Reports of the Department of Mathematical Information Technology
Series C. Software and Computational Engineering
No. C. 2/2009

Mining road traffic accidents

Sami Äyrämö Pasi Pirtala
Janne Kauttonen Kashif Naveed
Tommi Kärkkäinen

University of Jyväskylä
Department of Mathematical Information Technology
P.O. Box 35 (Agora)
FI–40014 University of Jyväskylä
FINLAND
fax +358 14 260 4980
http://www.mit.jyu.fi/

ISBN 978-951-39-3752-2
ISSN 1456-4378

Mining road traffic accidents*

Sami Äyrämö† Pasi Pirtala‡ Janne Kauttonen§
Kashif Naveed¶ Tommi Kärkkäinen‖

Abstract

This report presents the results from the research study on applying large-scale data mining methods into analysis of traffic accidents on the Finnish roads. The data sets collected from traffic accidents are huge, multidimensional, and heterogeneous. Moreover, they may contain incomplete and erroneous values, which make its exploration and understanding a very demanding task. The target data of this study was collected by the Finnish Road Administration between 2004 and 2008. The data set consists of more than 83000 accidents of which 1203 are fatal. The intention is to investigate the usability of robust clustering, association and frequent itemsets, and visualization methods to the road traffic accident analysis. While the results show that the selected data mining methods are able to produce understandable patterns from the data, finding more fertilized information could be enhanced with more detailed and comprehensive data sets.

1 Introduction

Killing more than 1,2 million and injuring between 20 and 50 million people every year, and thereby being the ninth most common cause of death in 2004, road traffic remains among the most central public health problems in the world [1]. A tragic fact is that among the young people aged between 15 and 29 years, a road traffic injury is the most common cause of death worldwide. While WHO reports that 90%

*This research was funded by the Finnish Road Administration, Keski-Suomi region.

†Department of Mathematical Information Technology, University of Jyväskylä, PO Box 35 (Agora), FI-40014 University of Jyväskylä, Finland, sami.ayramo@jyu.fi

‡Finnish Road Administration,Middle-Finland Region, PL 250, FI-40101 Jyväskylä, Finland,pasi.pirtala@ely-keskus.fi

§Department of Mathematical Information Technology, University of Jyväskylä, PO Box 35 (Agora), FI-40014 University of Jyväskylä, Finland, janne.kauttonen@jyu.fi

¶Department of Mathematical Information Technology, University of Jyväskylä, PO Box 35 (Agora), FI-40014 University of Jyväskylä, Finland, kashif.k.naveed@jyu.fi

‖Department of Mathematical Information Technology, University of Jyväskylä, PO Box 35 (Agora), FI-40014 University of Jyväskylä, Finland, tommi.karkkainen@jyu.fi

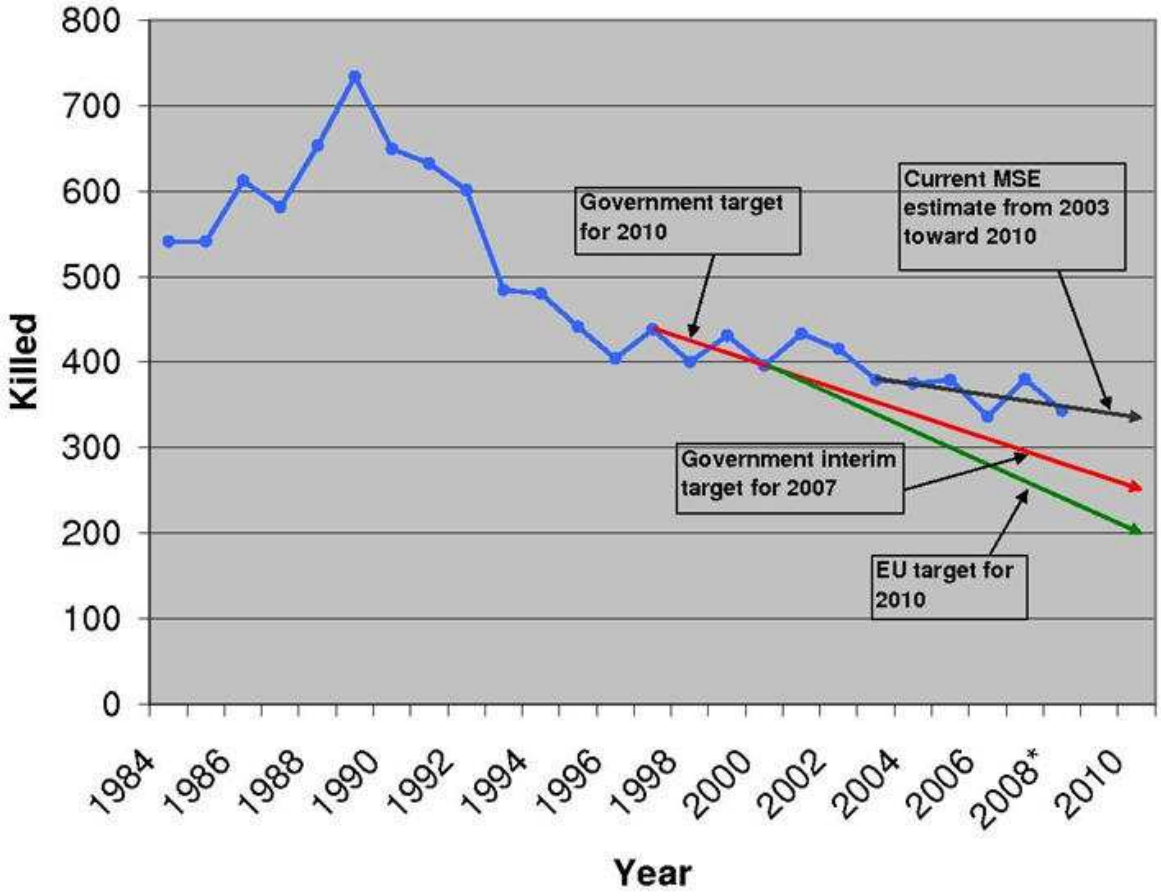

Figure 1: Trend curve, goals and recent estimates of road traffic accidents in Finland. The MSE estimate starts from 2003 due to the changes in the compilation of statistics in 2003. *The statistic not yet confirmed.

of the road traffic deaths occur in low-income or middle-income countries, about 39,000 people were killed in Europe and 1,400 in the five Nordic countries in 2008 [23]. In 2001, the EU set itself the goal of halving the yearly number of road deaths by 2010, but two years before the deadline it is already clear that the goal will not be reached [23].

In Finland, 344 people were killed in 2008, which makes 65 people per million inhabitants (http://www.stat.fi). With this number Finland holds the 10th position among the 27 EU countries in road deaths per million population [23]. Although Finland performs better than the EU average of 79 road deaths per million inhabitants, it is dropped down to middle level in EU in the pace of reducing road deaths. Figure 1 shows the progress in Finland over the past decades. While Finland has achieved remarkable improvements in the road safety over the past two decades, the good progress has decelerated during the last ten years so that the government target of 250 road deaths per year by the year 2010 has become unattainable for Finland.

When considered internationally, the yearly numbers of road deaths on the Finnish roads are at fairly satisfactory level, but the problem is that there has been only very slow progress during the last few years. Given the suffering experienced by the bereaved people, long-term consequences of the seriously injured victims and the increased load for the health care system, the road accident numbers can still be considered high in Finland. There are not many other individual causes that would kill or injure as many people as the road traffic does. Therefore, it is extremely important to keep on searching for new methods that will help us to reduce the number of road deaths. It seems that the recent actions have not been effective enough, because the decrease in the number of road deaths has almost stagnated during the last five years. Regenerating the good progress rate, and thereby catching up the top EU member states in the road safety development speed may require that completely new approaches must be found by the road and traffic safety administrators. Effective political or legislative resolutions and influential investments require thorough investigations and powerful analysis methods. Because traditional statistical analyzes are based on hypothesis testing on complete small scale samples, the findings are strongly driven by the analysts' prior assumptions. Data mining and knowledge discovery takes a different approach to the data analysis [29].

Data mining is an approach that focuses on searching for new and interesting hypotheses than confirming the present ones. Therefore, it can be utilized for finding yet unrecognized and unsuspected facts. In this study, feasibility and utility of data mining methods in the context of road traffic safety is studied. As data mining covers a large and versatile set of methods for large-scale data analysis, exploratory and descriptive methods are emphasized in this study. The intention is to find out whether robust clustering together with association and itemsets mining techniques is able to elicit reasonable, and hopefully novel, unsuspected and interesting facts from road traffic accident data.

The data used in the experiments consists of 83509 road traffic accidents in Finland between 2004 and 2008. Of these accidents, 17,649 injured and 1,203 caused death of at least one involved victim. Due to the small percentage of fatal accidents, it is important to include the whole data in the analysis of road network safety. Uniform distribution of the fatal road accidents over the 78,141 kilometers of highways of which Finnish Road Administration is in charge (www.tiehallinto.fi), yields one accident for every 65km during the five years period considered in this study. Even if the most of the accidents concentrate on the high volume main roads (class I and II), which account for 13,264 kilometers in total, it is still justified to expect that a significant amount of hidden information reside in those over 80,000 non-fatal accidents. Particularly fortuitous near-miss cases can be informative even if no one was killed or injured. The problem of how to recognize them from the data mass can be likely assisted with data mining techniques.

This report presents some results on real-world road traffic accident data. The results are neither conclusive nor exhaustive when it comes to determining or explaining the causes of traffic accidents. Instead of being conclusive, the results demon-

strate the capabilities of data mining and knowledge discovery in the field under this study.

1.1 Related work

Despite the wide variety of data mining applications, not so many research or development efforts in the context of road safety have been made. In this section a set of research efforts from this field are reviewed. Most of these studies have been accomplished in Europe, but some results from more exotic countries are available. The reports by WHO [28, 1], for example, show that differences in road traffic safety are huge between various countries around the world.

In Belgium, wide-range of research on mining road traffic data has been carried out by several researcher [16, 15, 11]. The researchers have applied, for example, model-based clustering methods, information criterion measures, and association analysis algorithms on traffic accident data.

Geurts et al. [16] used model-based clustering to cluster 19 central roads of the city of Hasselt for three consecutive three years time periods: 1992-1994, 1995-1997, and 1998-2000. Data consisted of 45 attributes that were very similar attributes used in this study, expect some more detailed variables, such as characteristics of the road user, fatigue, and rough physical geographic characteristics. They found out that that by clustering and generating frequent itemsets the circumstances that frequently occur together can be identified. Their analysis shows that different policies should be considered towards different accident clusters.

Geurts et al. [15] used data from the region of Walloon Brabant between 1997 and 1999. The data consisted of 1861 injurious accidents of which 81 were fatal. While they also used very similar attributes to this study, some more detailed variables were collected, such as characteristics of the road user, fatigue, and rough physical geographic characteristics. They found out novel explanations for traffic accidents "black" zones by frequent itemset mining and explained why accidents concentrate on certain road segments. They also found many interesting interaction patterns between accident factors which suggest that co-occurence of different factors depend on the accident zone.

The results by Depaire et al. [11] indicate that by clustering the roads into groups with equal accident frequencies enhance the understanding on traffic accidents. The data was limited to the Brussels capital region and accident with two involved road users. The final data set contained 29 variables and 4028 accidents over the period between 1997 and 1999. While the data contain very similar variables to the ones used in this research, some more detailed variables were collected, for example, hidden pedestrian, missing safety (e.g., wearing a helmet or safety belt), passenger positions in the vehicle, behavior (ignores red light, passes incorrectly, makes an evasive maneuver etc.) and accident dynamics (constant speed, acceleration, braking, not moving). Furthermore, detriment counts were calculated using a specific formula and the numbers of slightly and severely injured and deaths. Each cluster was characterized with cluster-specific attributes that should be affected for

enhancing road safety. Vehicle type, road type, and age were the main variables that contributed to the clustering result. They found out that 1) cluster models can reveal new variables influencing the injury outcome, 2) independent variables may have different influence on the injury outcome depending on the cluster, and 3) the effect of a single independent variable can differ in direction between different traffic accident types. A lot of this information may have remained hidden in the large heterogeneous full data set without clustering. While the results are promising, they could be further enhanced with more extensive traffic accident data.

Anderson [4] presented two-step methodology for profiling road accident hotspots. The first step identifies spatially high density accident zones by using Geographical Information Systems (GIS) and Kernel Density Estimation method. The second step recognizes the similar zones by adding environmental and land use data to the accident zones and classifies the hotspots by K-means clustering. The target data from the London area in the UK were collected by the Metropolitan Police between 1999-2003. The environmental attributes represented, for instance, road length, cycle lane length, pedestrian crossings, traffic lights, bus stops, schools, and speed cameras. The clustering process produced a meaningful hierarchical structure of five groups including 15 clusters altogether. The clusters described the spatial and environmental features of the accident hotspots.

Abugessaisa [2] from Linköping University (Sweden) developed a conceptual three-layer model that should advance the sharing of domain knowledge and communication between the information system developers and road safety organizations. Abugessaisa found out that visual and explorative data mining tools (e.g., dendrograms, K-means clustering, and self-organizing maps) may assist domain experts in observing hidden relationships and similarities in the road accident data sets and, thereby, formulate new and interesting hypotheses.

Sirviö and Hollmen [31] from Helsinki University of Technology (Finland) investigated the use of hybridized methods, including data clustering, principal component analysis, Markovian models, and neural networks in forecasting road conditions of Southern Finland. The results show that Markovian models are more straightforward and efficient than neural networks with the studied problem, but the authors suggested further research on clustering and neural network methods.

Chong et al. [9] also present interesting results for different machine learning techniques (neural networks, decision trees, support vector machines, and hybrid decision tree-neural network method) on the GES automobile accident data set that is collected from the United States. Their results show that hybrid decision tree-neural network approaches outperforms the single classifiers in traffic accident classifier learning.

In some studies the target data sets have been collected from countries with extremely high accident numbers like Korea and Ethiopia [32, 33, 36]. In [32], Sohn et al. compare neural network, decision tree, and logistic regression classifiers to build up classification models for accident severity prediction on traffic accident data set from Korea. Data set consists of 79 variables including many details, such as driver's education, violent driving, speed of car before the accident, rule violation, type and

status of driving licence, protective device, median barrier, injured body part, vehicle inspection status, loading condition, distance from driver's residence, curve radius, and the length of tunnel. The data contained 11564 accidents that were split into training data (60%) and validation data (40%) respectively. The class variable was the accident type (death, major injury, minor injury, injury report and property damage). After the variable selection 22 variable were used in classification. They found out that classifier accuracies were not significantly different on traffic accident data, but that decision trees produced the most understandable results. The protective device (e.g., seat belt) turned out to be the most significant factor to the accident severity.

In order to enhance the classification accuracy achieved in [32], Sohn and Lee [33] applied data fusion, ensemble and clustering algorithms to improve the accuracy of individual classifiers on road traffic accident data from Korea. As a result, the authors suggest that due to the large variations of observations in Korean road traffic accident data, the accidents should be clustered and then fit a classification model for each cluster accordingly.

Tesema et al. [36] applied adaptive regression trees to build a decision support system for classifying injuries into the predefined classes: fatal, serious, slight, and property damage. The data consisted of 5207 accidents described by 36 variables of which 13 were used in the classification. The data was obtained from Addis Ababa Traffic Office and collected between September 1995 to March 2005. The generated rules indicated that, for instance, accident cause, accident type, driver's age, road surface type, road condition, vehicle type, and light condition are important variables in the classification of accident severity. The authors concluded that their classification model is able to support the traffic officers at Addis Ababa Traffic Office when they are planning and making decisions in traffic control activities.

1.2 Preliminaries

Terms *data*, *information* and *knowledge* are used with the following meanings herein (see, [37, 34]):

- **Data** consist of not yet interpreted symbols, such as simple facts, measurements, or observations.
- **Information** consist of structured data with meaning.
- **Knowledge** emerges from information after interpretation and association with a context.

Concerning the formulae throughout the report, we denote by $(\mathbf{x})_i$ the ith component of a vector $\mathbf{x} \in \mathbb{R}^p$. Without parenthesis, $\mathbf{x}_i$ represents one element in the set of vectors $\{\mathbf{x}_i\}_{i=1}^n$. The l_q-norm of a vector $\mathbf{x}$ is given by

$$\|\mathbf{x}\|_q = \left(\sum_{i=1}^{n} |(\mathbf{x})_i|^q \right)^{1/q}, \quad q < \infty.$$

2 Data mining and knowledge discovery

Briefly, data mining (DM) and knowledge discovery in databases (KDD) refer to analysis of huge digital data sets. Hand et al. [19] define "*data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.*" The need for data mining arises from the huge digital data repositories. Data repositories are swelling both due to the increasing number of ways to measure different real-world phenomena and declined prices of digital storing facilities. In addition to the amount of data, quality of data (errors, missing data etc.) is another challenge. The digital data storages are often collected without any statistical sampling strategies [19].

While the traditional data analysis techniques have become inefficient to handle huge data sets, they are also based on the prior assumptions on data. In overall, data mining is more about the search than confirmation of hypotheses. Hence, data mining is not only concerned with algorithmic capabilities, but it also provide tools to accomplish analyzes without strong assumptions or knowledge on the data a priori. While the well-known data mining problem of "*the curse of dimensionality*" (e.g., [19]) pose requirements for the methods, at the same time it hinders analysts or decision makers from identifying previously unrecognized dependencies and similarities from the data.

Due to the explorative and descriptive nature, intelligible representation and visualization of the found patterns and models are essential for the successful mining process, particularly when the domain expert has limited knowledge of the data mining methodology.

While data mining is typically related to the algorithms, knowledge discovery from databases usually refers to the overall process [14]. Margaret Dunham [12] defines knowledge discovery in databases as "*...the process of finding useful information and patterns in data*". Data mining represents the step where the algorithms are applied to the target data.

This research on mining road traffic accidents is conducted according to the two-level knowledge mining (KM) process instead of the traditional KDD model [5]. The KM model provides well-defined interface for domain and method experts. The steps of the process are shown in the Figure 2.

2.1 Data mining tasks

Depending on the application, the target data, or the predefined goal of the analysis, various methods can be applied to the data mining task. Hand et al. [19] define the main types of DM tasks in the following way:

Exploratory Data Analysis (a.k.a. EDA) means explorative analysis of a data set, in which interesting and unexpected structures are visually observed from data. Graphical representation techniques, such as histograms, pie charts, scatter plots, and so on, are frequently utilized, but in case of high dimensional data

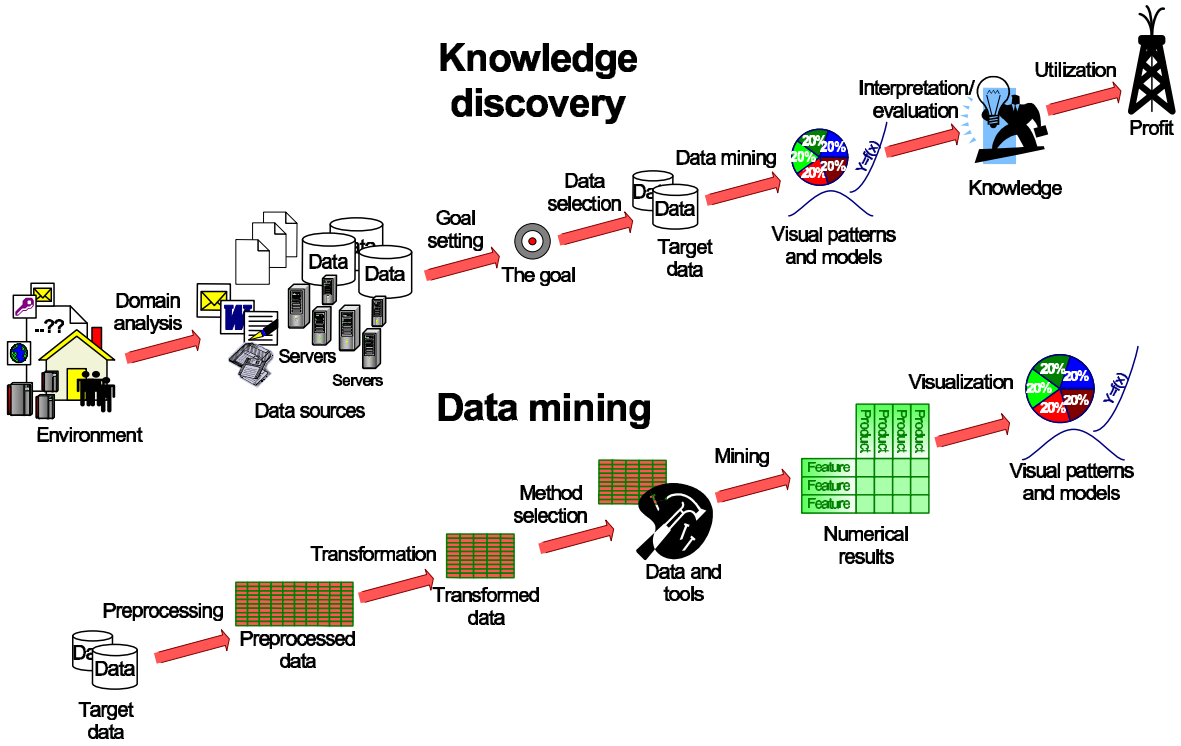

Figure 2: Knowledge mining process.

involving more than three dimensions, dimension reduction techniques, such as PCA and MDS, are needed to transform the data into a low-dimensional space. One should note that while a data miner is always performing search for unexpected novelties from data, she/he is rather exploring than confirming or discarding hypotheses. On this basis, the whole data mining and knowledge discovery methodology can be considered as an explorative data analysis approach.

Descriptive modelling is based on methods that describe a high-dimensional data set in a refined way without strong prior assumptions about the underlying classes and structures. Cluster analysis, segmentation, density estimation, and dependency modelling techniques are typically applied in this case.

Predictive modelling utilizes classification and regression techniques. A value of a particular variable is predicted from the values of the other known variables. In classification, the predicted variable is categorical (e.g., fatality of an accident), whereas in regression the variable is quantitative (e.g., traffic volume on a given road). Hence, predictive models are based on prior knowledge about the classes. Predictive data mining techniques are, for example, neural networks, nearest-neighbor classifiers, decision trees, and Bayes classifiers.

Discovery of patterns and rules searches for frequent itemsets, association rules and sequential patterns from data. Market basket analysis is the traditional exam-

ple. A large number of pattern and rule mining methods are based on the so-called Apriori principle [3].

Retrieval by content refers to finding interesting patterns from large data sets using, for example, a set of keywords. This approach is utilized in retrieval of documents or images from large databases. World-Wide-Web search engines are examples of the retrieval-by-content applications (for example, Google).

While the aforementioned classification is quite detailed, it is quite common to speak roughly about descriptive and predictive tasks (e.g., [35]).

The methods needed to accomplish the data mining tasks are typically made up of four elements [19]. **Model or pattern structure** determines an underlying structure or functional form of the data. **Score function** expresses how strictly a model fits the target data. In other words, it measures the error between a model and data. The best model produces the smallest error. **Optimization and search methods** are needed to minimize the error of model or pattern structure. The search methods are of two types: parameter search methods for a given model and model search from a model space. A parameter search problem is usually formulated through an optimization problem, for example, minimization of the least squares error. The pattern/model search problems are often solved by heuristic search techniques (e.g., the problem of best number of clusters). **Data management strategy** concerns the efficient data access during the model search or optimization phase. In data mining applications, the target data sets may exceed the capacity of primary data storages. Therefore, data management should not constitute a bottleneck for the advanced search and optimization algorithms.

Data mining methods and algorithms are introduced in many specific books (see, for example, [18, 12, 19]). Moreover, various related and useful methods that have been adapted to data mining requirements can be found in the literature in statistics, artificial intelligence, machine learning, pattern recognition, and database technology.

2.2 Frequent itemsets and association rule mining

A frequent itemset generation algorithm digs out frequently occurring itemsets, subsequences, or substructures from large data sets. A common example of frequent itemset applications is market basket analysis. Market basket analysis is a process that helps retailers to develop their marketing strategies by finding out associations between different items that customers place in their shopping baskets. Besides market basket data, frequent itemsets mining has been applied in, for example, bioinformatics and web mining.

Before illustrating the method with an example, a set of definitions are given. Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items (e.g., a set of products sold by a grocery store) and $T = \{t_1, t_2, \ldots, t_n\}$ be a set of database transactions (e.g., a purchase transaction in a grocery store) where each transaction t_i contains a subset of items chosen from I. A set of items is referred to as an itemset. An itemset that contains k items is

Accident	Gender	Age	Alcohol	Speed limit	Fatals
1	M	Young	Yes	≥ 100	Yes
2	M	Young	Yes	$70 - 90$	Yes
3	M	Middle	No	$70 - 90$	Yes
4	F	Young	No	≤ 60	Yes
5	M	Old	No	$70 - 90$	Yes

Table 1: Example traffic accident data.

an k-itemset. For example, the set {Bread, Milk, Beer} is a 3-itemset. The null (or empty) set is an itemset that does not contain any items. The occurrence frequency of an itemset (aka support count) is the number of transactions that contain the particular itemset. Transaction width is defined as the number of items included in the transaction.

In market basket analysis, the item-wise details, such as the quantity or the price of products sold, are usually ignored. Consequently, the items are represented as binary variables whose value is one when the item is present in a shopping basket and zero otherwise. If the presence of an item is considered more important than its absence (or vice versa), and item is considered as an asymmetric binary variable.

In the case of road traffic accident data, the concept of item must be treated in a slightly different way to the straightforward market basket application. While a typical market basket data deals only with the boolean associations (presence/absence), traffic accident analysis must usually be able to handle a heterogeneous set of different items and attribute types. Therefore, continuous attributes must be categorized into a smaller number of intervals within the range of attribute values using discretization or concept hierarchy formation techniques.

Mining of frequent itemsets using an artificial accident data given in Table 1 will be illustrated next. The data set contains five fictitious accidents that are described with four explanatory variables and one consequential variable. The frequent itemsets are generated according to apriori algorithm [3].

The relative minimum support threshold is fixed to 40%, which is equivalent with the minimum support count two. From this it follows that all such k-itemsets that appear in less than two accidents are discarded. Let us start the search of frequent itemsets by finding all the frequent 1-itemsets from the sample data. One can easily observe that there exist five items that satisfies the minimum support requirement. The attribute 'Gender' has value 'male' in four out of five accidents, which gives 80% support. Similarly, four other frequent attribute-value pairs are found from 'Age', 'Alcohol', and 'Speed limit' attributes. All the discovered frequent 1-itemsets are listed in the Table 2. During the next iteration 2-itemsets are searched. According to the Apriori principle all nonempty subsets of a frequent itemset must also be frequent. This means that all the 1-itemsets included in 2-itemsets must satisfy the minimum support threshold. The last iteration generates the 3-itemsets and finally we have the set of frequent itemsets listed in Table 1.

Size	Itemsets	support
1	{Gender = M}	0.8
	{Age = young}	0.6
	{Alcohol = no}	0.6
	{Alcohol = yes}	0.4
	{Speed limit = 70-90}	0.6
2	{Gender = M, Age = young}	0.4
	{Gender = M, Speed limit = 70-90}	0.6
	{Gender = M, Alcohol = yes}	0.4
	{Gender = M, Alcohol = no}	0.4
	{Age = young, Alcohol = yes}	0.4
	{Alcohol = no, Speed limit = 70-90}	0.4
3	{Gender = M, Age = young, Alcohol = yes}	0.4
	{Gender = M, Alcohol = no, Speed limit = 70-90}	0.4

Table 2: Frequent itemsets generated from the artificial accident data.

Association rule mining extract association rules from a given frequent itemset. For example, one may obtain rule $\{Gender = M\} \longrightarrow \{\text{Alcohol = no}\}$ from the fictitious data in Table 1. The left side of the rule $\{Gender = M\}$ is antecedent and the right side {Alcohol = no} is consequent of the rule. The association rules are usually assessed in terms of support and confidence. The support of the above rule is 0,4. Confidence of a rule is obtained by dividing the support count of the rule by the support count of the antecedent. Hence the confidence for the rule $\{Gender = M\} \longrightarrow \{\text{Alcohol = no}\}$ is $2/4 = 0.5$. Confidence measures the reliability of the inference derived from an association rule. One should note that the association does not necessarily mean causality between the items in the antecedent and consequent of a rule. Depending on the data, association rule mining algorithms may produce millions of rules, for which one may need to use also other interestingness measures besides support and confidence. In this study, we rank the rules according to lift measure, which computes the ratio between the rule's confidence and the support of the itemset appearing in the rule consequent.

A typical rule mining algorithm consists of two subtasks: frequent itemset generation and rule generation. The former finds the frequent itemsets that satisfy the minimum support requirement and the latter extract all the rules that satisfy the confidence requirement. Many techniques for rule mining are presented, for instance, by Tan et al. in [35].

2.3 Data clustering

Data clustering is a descriptive data analysis technique that is also related to unsupervised data classification [19, 35]. It is one of the core methods of data mining. As a result of data clustering the target data set is divided into groups (clusters) that are meaningful and/or useful. A cluster can be defined, for example, as "a set of enti-

ties which are alike, and entities from different clusters are not alike [22]." Cluster models provide valuable information about similarities, densities, and correlations of multivariate data objects and attributes. Depending on the application, the number of objects and dimensions may vary, but in data mining applications both are often huge. Unlike supervised classification, data clustering does not exploit any information about cluster memberships of data objects.

Clustering methods can be roughly grouped into two categories: partitioning and hierarchical methods. Other methods, such as density-based DBSCAN [13], fall somewhere in between the two major categories. Partitioning-based methods, such as K-means or K-spatialmedians [27, 5] are efficient methods that consume less memory than, for instance, hierarchical methods. This is a considerable advantage with large-scale data analysis tasks.

When considered purely from numerical perspective, the prototype-based data clustering problems, such as K-means or K-spatialmedians, are characterized as non-convex global optimization problems. However, the most fundamental problem in data clustering is to define such a score function that yields the most characteristic and informative clusters for the given data. Depending on the chosen validity measure, a local optimum may yield a more acceptable cluster structure than the global optimum. In fact, cluster models can not be generally evaluated regarding the numerical outcome of the method, since one analyst may see the value of the obtained clusters in completely different way than another. Hence, cluster validity is also a philosophical issue.

2.3.1 Clustering example

Figure 3 provides an artificial example of bivariate data clustering problem. The data consisting of three clusters is given in Table 3. We have 15 observations that are each described using two variables. An example of optimal K-means clustering is presented by marking each cluster by an individual symbol. The cluster prototypes are represented by the sample means (see the pentagram markers in Figure 3).

2.4 Robust prototype-based clustering methods

In this research, we are going to apply and evaluate feasibility of a robust prototype-based method on mining road traffic accident data. The principal idea of prototype-based data clustering is the following: for a set of n-dimensional data points (vectors) $\{\mathbf{z}_i\}_{i=1}^{N}$ ($\mathbf{z}_i \in \mathbb{R}^n$), a prototype-based clustering method finds a partition where intra-cluster distances are minimized and inter-cluster distances maximized. The number of clusters is denoted by K typically. Depending on the method each cluster is represented by, for instance, the sample mean, median or some other multivariate location estimator. K-means is definitely the best-known prototype-based clustering method [27].

In this study, we have chosen to use a robust and reliable prototype-based clustering method, namely K-spatialmedians, which is based on a statistically robust es-

id	v1	v2
1	-0.4	-1.2
2	-0.5	1.0
3	-1.1	-0.6
4	-2.6	0.6
5	-1.1	1.1
6	6.0	4.7
7	5.5	4.1
8	5.5	5.9
9	5.6	6.9
10	5.4	4.2
11	4.0	-2.1
12	4.3	-2.8
13	3.1	-1.8
14	4.2	-1.6
15	3.5	-2.4

Table 3: Artificial bivariate data.

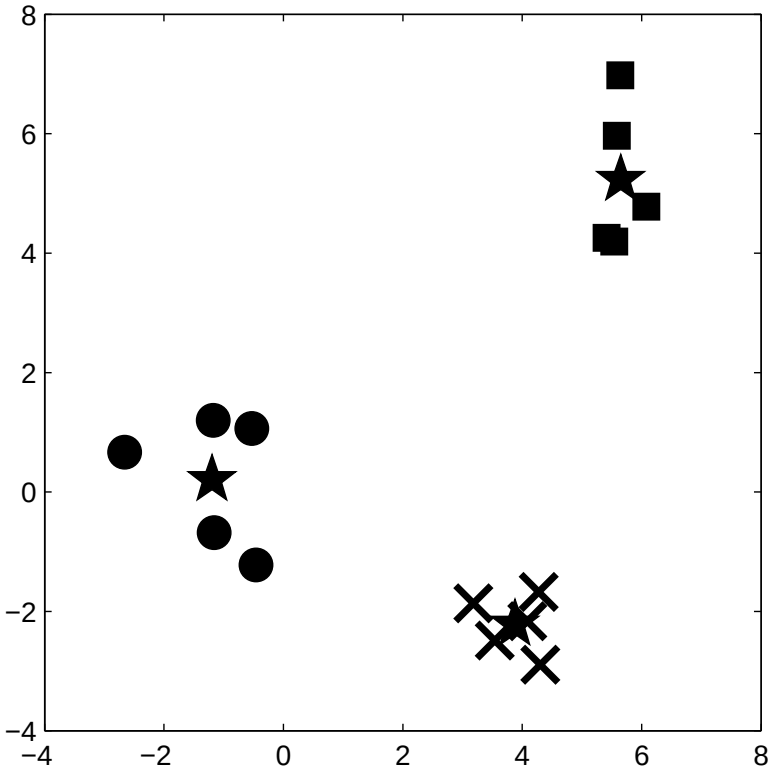

Figure 3: Artificial bivariate cluster data.

timation of prototypes and the K-means-wise expectation-maximization (EM) strategy [27, 10, 5]. According to a classical book by Huber [21] "*robustness signifies insensitivity to small deviations from the assumptions*". A small deviation from the assumptions may refer either to gross errors in a minor part of the data or small errors in a large part of the data. The primary goal of the robust procedures is to safeguard against those errors. A typical deviant in a data set is an outlier, which is an outlying observation with one or more data values that deviates significantly from the main bulk of the data [7]. An outlier can be caused, for example, by a failure in a data acquisition system or by a human mistake. On the other hand, an outlying value may also be a correct measurement of an object with deviating features. For instance, extremely high breath test value may be due to heavy drinking, measurement error, or misused encoding (e.g., 9,99). Another type of deviation from the complete and normal data sets are missing data that may exist due to various reasons [26]. The missing data mechanism may be fully unknown, which makes the estimation of the correct value difficult. With large data sets, manual missing data or outlier analysis and replacement is an insurmountable task, but if not taken into account, they prevent precise and correct inferences from erroneous and incomplete real-life data sets.

Robust estimation of prototypes can be realized by using the spatial median instead of the sample mean (in K-means) [24, 25]. When compared to the common traditional methods, K-spatialmedians provide prototypes that are more robust to extreme data values and gross errors. The breakdown point of the spatial median is 50%, which means that at least 50% of data points have to become disturbed in order to change the estimate infinitely. Note that in the univariate case the spatial median coincides with the coordinate-wise median. In addition to prototypes, robustness of the whole clustering method depends on the initialization approach, which determines the search space neighborhood. K-spatialmedians algorithm is reliable in the sense that it will not be failed by anomalous numerical conditions. This means that empty or singleton clusters and non-smoothness of the problems will not crash the method. Even if there exist no closed-form solution for the problem, the spatial median estimate be approximated precisely and in short time using the iterative Weiszfeld method that is accelerated with the successive over-relaxation step [25]. Although the use of robust estimates leads to computationally more intractable problems in general terms, the previous results indicate that the reduced number of clustering iterations obtained by a refinement initialization strategy compensates the cost of the estimator [6].

2.4.1 K-spatialmedians algorithm

At first the score function of the robust clustering method is defined. Based on the well-known K-means score function, a more general K-estimates clustering problem is defined in [5]:

$$\min_{\mathbf{c}\in\mathbb{N}^N,\mathbf{m}_k\in\mathbb{R}^n} \mathcal{J}(\mathbf{c},\{\mathbf{m}_k\}_{k=1}^K) = \sum_{i=1}^{N} \|\mathbf{P}_i(\mathbf{z}_i - \mathbf{m}_{(\mathbf{c})_i})\|_q^\alpha \tag{2.1}$$

$$\text{subject to } (\mathbf{c})_i \in \{1,\ldots,K\} \text{ for all } i = 1,\ldots,N,$$

where $\mathbf{c}$ is a code vector, which represents the assignments of the data points to the clusters and $\mathbf{m}_{(\mathbf{c})_i}$ is the prototype estimate (e.g., the sample mean in the K-means method) of the cluster, in which data point $\mathbf{z}_i$ is assigned to. $\mathbf{P}_i$ is the diagonal projector matrix where the j^{th} diagonal element equals to one given the j^{th} element exists in x_i, and otherwise j^{th} element of $\mathbf{P}_i$ equals to zero. On computer platforms (e.g., MATLAB) each diagonal matrix $\mathbf{P}_i$ of size p^2 are implemented as $p \times 1$ vector for minimizing the memory usage.

By choosing $q = \alpha = 2$ one obtains the aforementioned problem of K-means. However, by choosing $q = 2\alpha = 2$ one obtains the robust formulation, the problem of K-spatialmedians. This is the score function of the robust K-spatialmedians clustering method. By using P_i projections all available data values are exploited without need for manual or computerized missing data strategies.

The K-spatialmedians method finds K clusters from a given data set so that the sum of the cluster-wise Euclidean distances is minimized. This is obtained by a search method that iterates between the following two steps:

1. Reassign data points to their closest cluster prototypes according to Euclidean distance.

2. Update prototypes (computation of the sample spatial for each cluster).

If no more reassignments occur, then the algorithm terminates. As previously mentioned, the algorithm follows the well-known expectation-maximization (EM) strategy [10] and guarantees only a locally optimal solution. The prototypes are updated using the SOR-accelerated iterative Weiszfeld algorithm [25].

Computational complexity of the K-spatialmedians algorithm is $O(NnKt_{EM}t_{SOR})$, where t_{EM} is the number of clustering iterations and t_{SOR} is the number of SOR-iterations. Usually, $n, K, t_{EM}, t_{SOR} \ll N$, which means that the number of data points contributes the most to the computational cost. t_{SOR} depends on the required accuracy of the spatial median estimates. t_{EM} is usually very small, especially for the reasonably initialized robust clustering. Therefore, a clustering refinement method is used for the initialization step in this study. The clustering refinement is a density-estimation based method and the idea was initially presented by Bradley and Fayyad [8]. Their experiments indicate that the quality of the K-means clustering can be improved with this initialization method. The K-spatialmedians is shown to be more tolerable to noise and gross errors than the original K-means-based refinement approach and also comparable from the computational point-of-view [6].

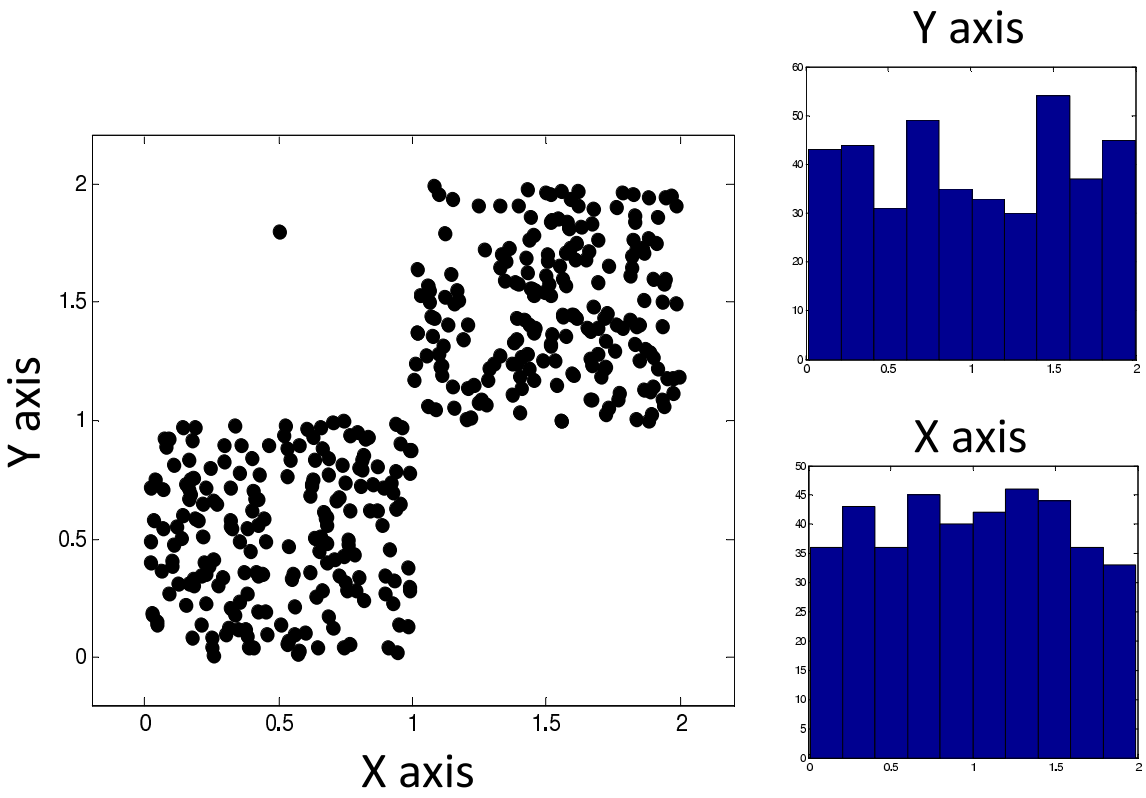

Figure 4: Two bivariate clusters with a single outlier. On the right-side the corresponding univariate distributions.

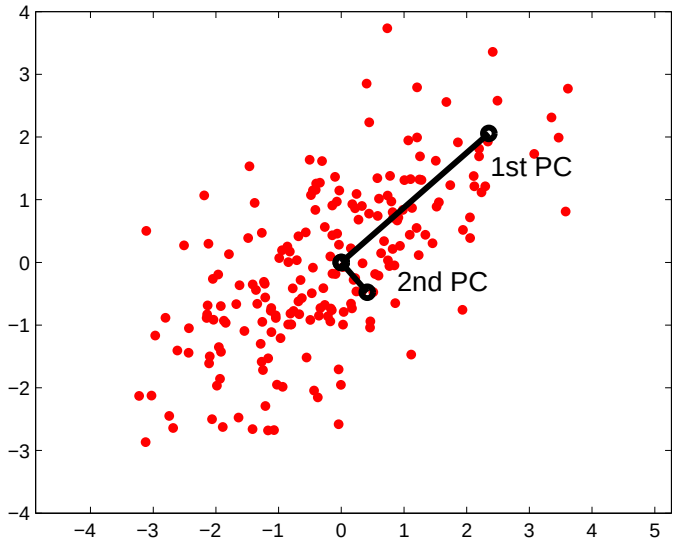

Figure 5: Example of principal components for bivariate data.

2.5 Dimension reduction

Dimension reduction and feature selection are also closely related to many data mining methods. From the perspective of the knowledge mining process, dimension reduction techniques can be used either for data preprocessing before the actual data mining step or they can be used as tools for exploring and visualizing the data. In data visualization, dimensionality of the data is reduced before the data exploration step. Common data visualization tools are, e.g., parallel coordinates and numerous plotting techniques (scatter, trellis, star, box plots etc.) [19, 35, 18]. The weakness of such methods is that high-dimensional data are not necessarily scattered or clustered in an interesting way in any direction of an individual coordinate or pair of coordinates (see Figure 4). Therefore, transformations that preserve the directions of maximal variances or class-discrimination are needed. Principal component analysis (PCA) is perhaps the best-known of such methods [20]. PCA finds the orthogonal directions that maximize the variability of the high-dimensional data (see Figure 5). The data points and cluster centers can be projected onto the most informative principal component axes. In data visualization two or three most informative axes are used in the data representation. A reduced dimension is obtained by projecting the original high-dimensional data from the original $\mathbb{R}^p$ space into the low dimensional $\mathbb{R}^q$ space ($p \gg q$) determined by the principal components.

Let us assume that $\mathbf{X}$ is a mean-centered and standardized p-dimensional data set. The q-dimensional projection $\mathbf{y}_i$ of any vector $\mathbf{x}_i \in \mathbf{X}$ is obtained by

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i,$$

where $\mathbf{A}$ is an orthogonal $q \times p$ transformation matrix. In the classical PCA method $\mathbf{A}$ is defined by the eigenvectors of the covariance matrix for $\mathbf{X}$. Let $\mathbf{\Sigma}$ be the covari-

ance matrix of $\mathbf{X}$. Eigenvectors $\mathbf{e}_i$ and the corresponding eigenvalues λ_i for $\mathbf{\Sigma}$ are obtained as a solution of the problem:

$$\mathbf{\Sigma}\mathbf{e}_i = \lambda_i \mathbf{e}_i, \quad \text{for} \quad i = 1, \ldots, p.$$

By substituting the largest eigenvectors to the rows of $\mathbf{A}$, one obtains a transformation matrix that can be used to map the points from the original space onto the low-dimensional orthogonal representation. In the case of data clustering, one can also determine the principal component directions using cluster prototypes. A lot of computational resources can be saved by using a small number of prototypes for computation of principal projections. Moreover, by using prototypes as an input data for PCA the low-dimensional data will likely preserve the most discriminative directions of the original data, because the prototypes usually try to maximize between-cluster distances.

While the principal component mapping produces a linear projection of data, multidimensional scaling (MDS) produces a non-linear transformation [20, 19]. MDS is based on the idea about preserving the pairwise distances between the data points. Let d_{ij} be the distance (e.g., Euclidean distance) between two p-dimensional observations $\mathbf{x}_i$ and $\mathbf{x}_j$. A standard MDS finds a set of q-dimensional ($q < p$) vectors $\{\mathbf{x}'_1, \ldots, \mathbf{x}'_n\}$ which minimizes a cost function given by

$$\mathcal{J}_{MDS}(\{\mathbf{x}'_1, \ldots, \mathbf{x}'_n\}) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} (d_{ij} - d'_{ij})^2,$$

where d'_{ij} is the distance between the unknown q-dimensional vectors $\mathbf{x}'_i$ and $\mathbf{x}'_j$. The cost function must be minimized by using an appropriate optimization algorithm. Perhaps the most popular variant of MDS is *the Sammon's mapping* given by

$$\mathcal{J}_{Sammon}(\{\mathbf{x}'_1, \ldots, \mathbf{x}'_n\}) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{(d_{ij} - d'_{ij})^2}{d_{ij}}.$$

By normalizing the pairwise distances in the original space, smaller distances are weighted. While MDS requires only pairwise distances and generalizes to any dissimilarity measure, it takes a lot of computational resources for large data sets.

3 The accident mining research process

Knowledge mining process steps (Figure 2) can be employed for enhancing communication between domain experts and data mining method developers. In this chapter we describe the realized analysis process by following the knowledge mining process.

KM1. Domain analysis

Although a road traffic accident is a familiar concept for anybody, continuous domain analysis must be carried on by the road administration. This enables timely

decisions and actions under the constantly changing circumstances. As a result of domain analysis interesting variables are measured and stored in databases for further analysis. Note that domain analysis and data gathering do not necessitate the further data mining efforts.

KM2. Goal setting

Based on the administrative responsibilities and existing data sources, the goal of the knowledge mining is specified together with the academic data mining researcher. In this study, the goal is to accomplish preliminary study on utilization of clustering and association rule mining techniques on road traffic accident data sets. A special effort is put on finding methods for detecting and understanding previously unknown risk factors behind fatal accidents and classifying the potential accident locations.

KM3. Data selection

After the initial experiments on the data set from Middle-Finland region were accomplished the study was extended to cover the whole Finland data. The data set consist of all the registered road traffic accidents on the Finnish road network during the years 2004–2008. It contains 83509 accidents including 1203 fatal and 17649 injurious accidents. Because the data were distributed into several tables and files, data integration was needed. Some duplicate values, such as a binary drunken driving variable and the permille count of alcohol, were pruned. Also some irrelevant attributes, for example, the dates related to the administrative processing, were removed. The following datasets were used in the Traffic accidents research project.

Accident The accident dataset contains the detailed information about accidents like accident severity (fatal, non-fatal etc), accident location (region, district, county, road number, road segment, distance from the start of road segments, address etc), temporal data like year, month, day, hour, and date, environment variables like weather, temperature, and lightness, road conditions like surface type, road width, walkways, junction types, traffic lights, speed limits, road works, heavy and light traffic volumes etc, the accident type like turn and hit, overtake, animal hit, the number of injuries, casualties, vehicle type and and so on.

Persons The Person dataset contains the information of the people involved in the accident like accused driver, non-accused driver, alcohol involved etc.

Participants The participants dataset contains the information about the other participants like vehicles, animals etc.

Population density Finnish Road Administration considers the population density as a possible and potential risk for the traffic accidents. The population density of the areas around the road network is defined in the population density

repository. The population information is recorded against the road number, starting segment, ending segment, starting distance from the starting segment, ending distance from the ending segment. The interesting fact is that the population density may remain the same over several road segments or may vary even within one segment after a small distance. As the population density is an independent repository, in order to retrieve the population density information for each traffic accident location, the population density and traffic accidents repositories were integrated together.

Altogether, 32 variables were qualified for further analysis as a result of data selection. Finally, the population densities at the accident locations were integrated to the accident data set from a separate database. The integration of the population density attribute is based on the mapping of the road number, road segment, and distance attributes between the road network and population density data sets. The complete set of target clustering attributes is presented in Table 4. The authors want to emphasize that none of the used variables risk the privacy of the involved victims.

KM4.1. Data preprocessing

Before employing any data mining method all the attributes were integrated into a single data matrix and inconsistencies were removed from data. Although missing value were not preprocessed (e.g., imputation) there are cases where a missing value has a meaningful explanation. For instance, a missing value in the *traffic lights* attribute indicates that there exist no traffic lights at the accident place. Thus, all the missing values in the *traffic lights* attribute were replaced by zeros. In the original data there exist four separate attributes for a pedestrian/bicycle way that were replaced by one binary valued *pedestrian/bicycle way* attribute which indicates whether there exist pedestrian/bicycle way or not. Weekday and hour information were categorized into ten class that were defined by the road administration representatives (see Table 5 in Appendix 1).

KM4.2. Data transformation

The target data set contains many different types of attributes on different scales. These must be transformed and scaled before the actual mining methods can be employed effectively [30]. The nominal attributes are transformed by creating binary-valued pseudo variables for each label value. After the binarization all the attributes are normalized. Typically all attributes values are transformed to the equal range, e.g., $[0, 1]$ by linear scaling transform, or another option is to normalize the standard deviations or median absolute deviations of the data distributions. We chose to transform the ratio and interval type of attributes to the range $[0, 1]$ and binary-valued attributes to the range $[0.25, 0.75]$. By using smaller range for the binary-valued attributes we prevent them having maximal weight in all distance computations. The rationale is that zero and one of a binary attribute represent lower and

ATTRIBUTE NAME	TYPE	VALUES
Accident-specific attributes		
accident type	nominal	0,...,99
accident category	nominal	1,...,13
accident scene	nominal	1,...,9
heavy traffic involved	nominal	yes/no
number of involved vehicles/animals	ratio	1,...
number of killed persons	ratio	0,...,23
number of injured persons	ratio	0,...,24
Driver-specific attributes		
gender	binary	male/female
drunken driver	binary	yes/no
age	ratio	5,...,98
Road-specific attributes		
population density	nominal	0,...,6
road pavement	nominal	0,...,6
traffic lights	nominal	0,...,4
speed limit type	nominal	1,...,6
motor/semi-motor highway	nominal	1,...,3
functional road class	nominal	1,...,4
maintenance class	nominal	1,...,8
pedestrian/bicycle way	binary	yes/no
arterial highway	binary	yes/no
speed limit	ratio	20,...,120
average daily traffic volume	ratio	9,...,88610
average daily heavy traffic volume	ratio	0,...,8359
number of roadways	ratio	0,...,4
roadway width	ratio	35,...,379
sight distance 150m	ratio	0,...,100
sight distance 300m	ratio	0,...,100
sight distance 460m	ratio	0,...,100
Circumstance-specific attributes		
time	nominal	1,...,10
road condition	nominal	1,...,6
lightness	nominal	1,...,4
weather	nominal	1,...,7
temperature	interval	-36,...,+35
Others attributes (only for interpretation)		
month	nominal	1,...,12
region	nominal	1,...,14

Table 4: Road accident attributes.

upper halves of the range, respectively. Instead of giving full-weight for the attribute, the values are scaled to the most representative points of the halves, that are 0.25 and 0.75. After the transformation, all attributes where some constant value exist in more than 95% of observations are removed, but taken into account again during the interpretation. 67 variables were used as input for the clustering method.

KM4.3. Method and parameter selection

In this pilot-study we chose to try the fast and robust k-spatialmedians algorithm in the clustering task. Our previous studies have shown that the method may produce more stable clusterings than, for example, the classical K-means method [27]. The SOR-based K-spatialmedians algorithm is also comparable with respect to computation cost due to the lesser number of clustering iterations. The K-spatialmedians algorithm projects all computational operations to the all existing values which means that no special missing value handling is needed.

KM4.4. Mining

All the computation were performed using MATLAB 7.5.0.338 (R2007b) software installed on HP ProLiant DL585 server with four AMD Opteron 885 (2,6GHz) dual core processors, 64GB of memory, and 64-bit x86_64 Red Hat Enterprise Linux Server release 5.3 OS. According to preliminary test runs we chose to use K=7 as a number of clusters. For the initialization so called clustering refinement principle was chosen (for more details, see [6]). Despite clustering refinement is based on subsamples, we chose not to apply subsampling strategy and used the whole data for each run. The number of refinement runs on the whole data was ten. Thereafter the obtained ten sets of seven prototypes were clustered by starting once from each set of prototypes. The prototypes with the smallest error were chosen as the initial clustering for the final clustering run. The maximum number of algorithm iterations was always 100.

KM4.5. Visualization

The clustering results are presented using PCA and MDS dimension reduction techniques, introducing the most characteristic attribute values for each cluster, and mining the frequent itemsets from the clusters using the Apriori method.

KM5. Interpretation/evaluation

The outcomes of the data mining process were evaluated through the numerical and visual representations in several meetings between the road administration and the data mining research group. The results were also discussed in a public multidisciplinary seminar in the presence of e.g., several road traffic safety, road adminstration and police representatives.

KM6. Utilization

The results published in this report will be used for developing knowledge among the road administrative and traffic safety related people for planning the further actions and projects in the field.

4 Data mining results

Based on the KM process described in the previous section we obtained seven disjoint groups (subsets) of the analyzed data. In this section these groups, i.e. clusters are interpreted separately. In short we obtained interpretable groups of accidents with varying risk levels. Due to the small relative number of fatal accidents (1203 fatal accidents in the whole data), the cluster-wise percentages of fatal accidents are naturally small. However, deviations between the clusters may indicate the need for a more thorough analysis of non-fatal accidents in the clusters of many fatalities. Therefore, it is important to define and discuss all the characteristic cluster features. Tables and figures of the summary statistics of the most significant attribute-value pairs are presented in Appendices 1-3. The table references will not be repeatedly used in the following discussions.

Tables in Appendix 1 show that different clusters possess varying proportions of fatal and injurious accidents. Table 6 reveals that cluster 7 have clearly higher percentages of fatal accidents than the other clusters.

Figure 6 presents the attribute ranking with respect to entropy-based clustering information gain. Entropy measures the impurity of a set of data. The range of each variable is divided into K (the number of clusters) bins of equal width. If all prototypes fall into at most one bin, entropy is the lowest. Other way round, entropy is highest when each prototype falls into a separate bin. If the information gain equals for a set of variables, then these variables are ranked with respect to their standard deviations.

Information Gain based attribute ranking indicates that the road characteristics dominates the clustering results. For instance, clusters 3 and 7 seem to be quite similar with respect to many attributes. The accidents have occurred on quite similar roads, but cluster 3 consist of almost completely animal accidents and based on lightness variables many accidents during the dark time of the day while cluster 7 contains more single vehicle accidents in daylight. Later we will see that despite the similar road conditions the severity of accident consequences are very different. The most significant InfoGain variable, that is the percentage of class I main road accidents, shows similarities with PCA-plot in Figure 9. By this way one obtained initial understanding about cluster similarities for more detailed cluster analysis.

4.1 Cluster 1: Motor- and semi-motorway accidents

Cluster 1 is the motorway cluster in which accidents happen mainly on multi-roadway main roads where the speed limit is mainly 80, 100, or 120 km/h. These roads are

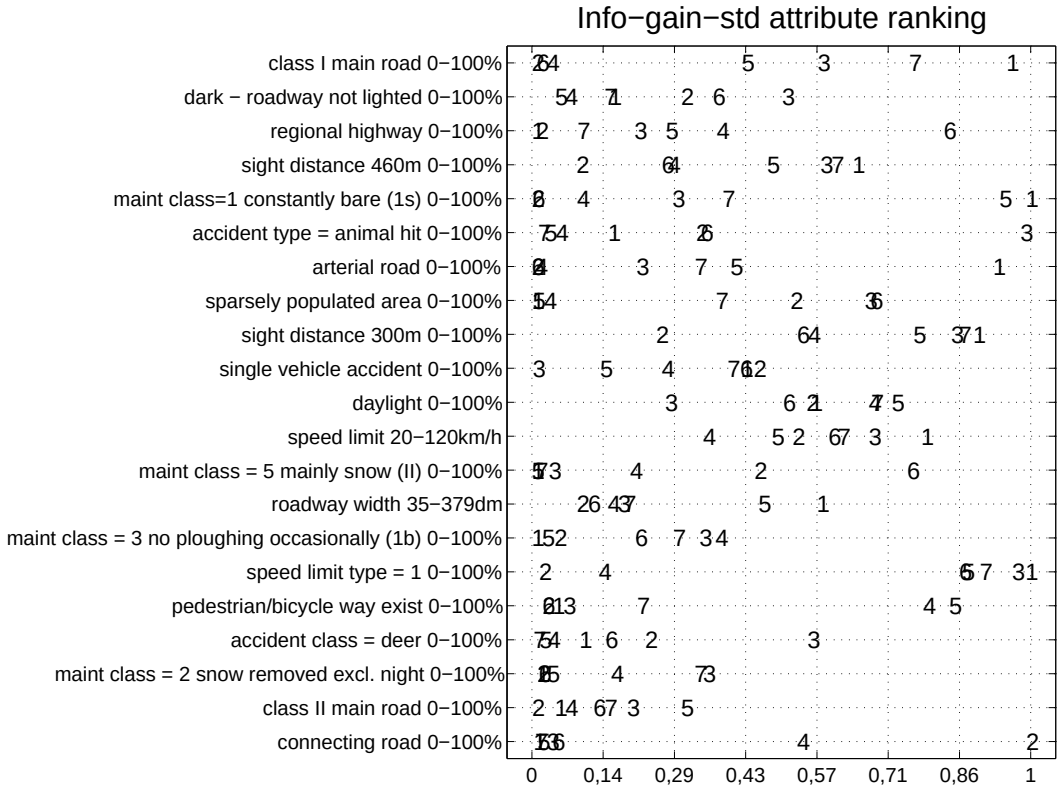

Figure 6: The most significant clustering attributes (from top to bottom) according to Information-Gain standard deviation measures.

characterized by the high traffic volumes and accidents occur during the morning and afternoon rush hours, especially Friday afternoon. The percentages of fatal and injurious accidents are very close to the average level of the whole data. The age distribution is slightly more skewed towards the 20-34 years old drivers than the full data.

Despite the high speed limits, the consequences are not more severe than the average of the data, because the separated opposing traffic lanes prevent the head-on collisions, and ramps, acceleration and deceleration lanes prevent the fatal intersection accidents. The most typical types of accidents are single vehicle run-offs, overtaking and head-to-tail collisions, and animal hits. While the overall number of accidents accumulated to December, January and June, the fatal accidents occurred most often in January and September. It seems that bad weather conditions have most impact on motor- and semimotorways. The percentages of bad weather (rain, snowfall, and sleeting) and road conditions (water in the ruts, snow, and slush) are among the highest in this cluster. This may indicate the need for more expeditious ploughing, gravelling and salting or variable speed limits. In overall, this cluster indicates that higher traffic volumes and speed limits do not increase the accident

risk as far as the head-on and intersection collisions are prevented with separated roadways and ramps.

4.2 Cluster 2: Alcohol-involved connecting road accidents

Cluster 2 consists mainly of animal, run-off, and head-on collision accidents on small connecting roads between the build-up areas with low maintenance priority. Despite the fact that this cluster represents the largest proportion (15,4%) of drink-drivers the severity of consequences do not differ much from the whole data averages. This may be due to the low traffic volumes which reduce the probability of head-on collision in single vehicle loss-of-control accidents. On the other hand, speed limits are low on the connecting roads which naturally reduce risk of head-on collisions. The role of alcohol can be seen in the fatal accidents of which 32,5% are caused by drink-drivers. Speeding by young drivers is probably another explanation to the fatal accidents, because in contrast to the overall accident distributions by month and age, the accidents happen mainly in summer and the proportion of young drivers is high among the fatal accidents. Moreover, a large part of the accidents occurred in the weekend nights. The large number of head-on collisions in curve may be also related with the narrow roadways and low maintenance priorities.

4.3 Cluster 3: animal hits

This cluster is a clear anomaly among the others by consisting almost entirely of animal hits. Risk of personal injury is clearly lowest of all. The driver in this cluster is typically a 25-64 years old man which is correspondingly seen in the high median age. The interesting detail is that the driver's median age among the fatal accidents is very low. Fifty percent of the fatal accidents were caused by the drivers under thirty years old, while the same age group have caused only 21.3% of all the cluster three accidents. Alcohol is involved only in one fatal case in this age group. On the other hand, even if the middle-aged drivers caused 3150 (20,6%) out of 15279 accidents, they were very seldom involved in the fatal cases, since only two (5%) out of total forty fatal accidents were caused by the 40-49 years old drivers. For a comparison, the group of 50-59 years old drivers caused seven (17,5%) out of forty fatal accidents which is very close to their proportion (19,9%) of cluster 3 accidents overall. The different consequences among the age groups may appear due to several reasons. Young drivers are usually more prone to speeding, but their cars are probably not as well equipped as the cars owned by the middle-aged drivers. On the other hand, the drivers over fifty years old may be more prone to fatal body injuries than the drivers under fifty years. Although alcohol is often related with the severe accidents by young drivers, drink-driving has almost no role in this cluster. In overall, the used data set does not provide enough detailed information about speeding, vehicle defects, protective devices (airbag), or type of fatal injuries, which makes it impossible to assess the underlying reasons. From the road network point-

of-view, the animal hits occur most often on one roadway main and regional roadways outside densely populated areas where the speed limits vary between 70 and 100km/h. According to common knowledge most of the animal hits are driven late in the evening or early in the morning in dark and wet autumn weather (see, Figures 16 and 10).

4.4 Cluster 4: Built-up area accidents

This cluster consists of accidents that are mainly happened inside built-up areas. The risk of fatalities is lower and the risk of personal injuries higher than in the whole data. This may be due to the low speed inside the build-up areas. The weekly distribution of accident times concentrates on the afternoon rush hours. The monthly distribution of all the accidents accumulates to December and January, while the fatal accidents concentrate on the period between June and October. Young drivers and alcohol have significant proportions in this cluster. On the other hand, in this cluster female drivers are involved more often than in the others, but their proportion is still only 28,4%. It is particularly noteworthy that the proportion of young drivers is larger among the fatal accidents than in overall among the clusters. While the overall distribution of accident types is diverse, the proportions of intersection, moped, bicycle, and pedestrian accidents are highest of all the clusters. Differently to other clusters, 9,1% of the accidents have occurred at crosswalks or on pedestrian/bicycle ways. This is not surprising since this is well-known characteristic feature to the built-up areas.

4.5 Cluster 5: Low-speed multi-lane roadway accidents

Cluster 5 consists of accidents that have happened on high traffic volume low-speed limit (50-80km/h) multilane roadways during the afternoon rush hours. A heavy vehicle has been involved in 24,4% of the accidents. The most representative accident types in this cluster are single vehicle run-offs, turning, overtaking, lane change, intersection, and head-to-tail accidents. 31% of the accidents are either a rear-end collision with a braking car or with a car standing due to a traffic obstruction. Because of the multiple roadways head-on collisions are not common for cluster 5. While the proportion of personal injuries equals to the whole data average, the proportion of fatal accidents is clearly smaller than the whole data average. While the age attribute follows very closely the distribution of the whole data, the highest 5% percentile of the drivers in the fatal accidents are at least 86 years old. This is, however, based on only thirty fatal accidents which means that the five percents is a result of two fatal cases by the oldest age group.

4.6 Cluster 6: Low traffic volume regional highways

Cluster 6 represents accidents that have mainly occurred on low traffic volume regional roads in sparsely populated areas. About 70% of the roads have oil gravel

pavement. The accidents have occurred for large part during the afternoon rush hours, but mainly in weekend evenings. The age distribution is almost equal with the whole data, but the median age in the fatal accidents (44 years) is five years over the whole data fatal accident median (39 years). The number of drink-driving cases is slightly higher than the average of whole the data. While the proportion of accidents resulting in one or more injured victims equals with the whole data, the proportion of fatal accidents exceeds the whole data count. While single vehicle accidents and animal hits dominate this cluster, the head-on collisions and relatively high speed limits increase the percentage of fatal accidents. 30,2% and 8,0% of fatal and all the accidents in cluster 6 are head-on collisions, respectively, which show the high risk of two lane roads without median barriers. The accidents concentrate to November and December, but there is a peak in the number of fatal accidents in August. Overall, in this cluster the median temperature is lowest. While the maintenance priority is also low, it is not surprising that the accidents concentrate to winter time in this cluster.

4.7 Cluster 7: Fatal main road accidents

Cluster 7 is the most severe one with having the highest percentages of fatal (3,0%) and injurious (28,0%) accidents. The cluster consists of accidents that have happened on one roadway class I main roads during morning and afternoon rush hours. The speed limit is 80 or 100 km/h in 83,3% of the accidents. A heavy vehicle is involved in 22,8% of the accidents. The accidents are distributed to both statistical built-up areas and sparsely populated areas. Driver attributes (gender, age, and drink-driving) are equally distributed with the whole data. The monthly accident distribution is bimodal with peaks in December-January and July. The distribution of the fatal accidents is slightly concentrated to the summer season, but there are also lot of accidents in January. About 30% of the accidents in this cluster have happened on icy, snowy, or slushy road surface and and 11% while snowing or sleeting.

4.8 Dimension reduction and visualization

Figures 7 and 9 depict the clusters prototypes through a two-dimensional MDS and PCA plots, respectively. The PCA plot is produced by using 0,5% samples from clusters and the K-means and K-spatialmedians prototypes. The transformation exploits all available data values. The MDS plot is created by the Sammon mapping stress function (MATLAB Statistics Toolbox Version 6.1 (R2007b)) that is initialized by PCA. The default MATLAB input parameter values were used. The data is sampled (0,5% of each cluster) before the transformation. The dissimilarity matrix is computed by following the principles of the Gower's general similarity measure [17]. Figure 8 presents the relationships of the inter-point distances between the original and scaled data space. The Shepard plot shows that the small distance are slightly under-estimated and large distances over-estimated. As the Sammon's mapping methods does not handle co-located points, all zero distances were re-

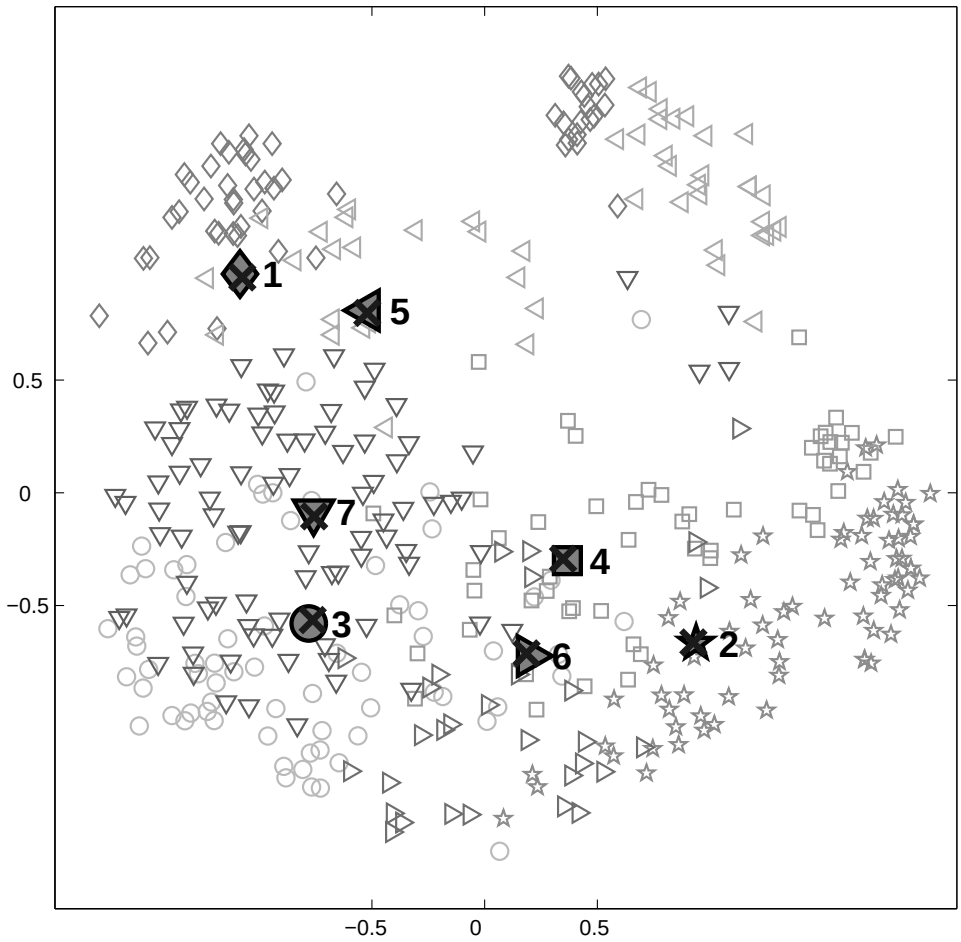

Figure 7: Sammon's mapping plot of the K-spatialmedians cluster prototypes. K-means prototypes are marked by 'X'. The cluster data points are represented by 0.5% samples of the whole clusters.

placed by the square root of machine epsilon. The transformation is performed using the combined data consisting of K-means and K-spatialmedians prototypes and the cluster data samples.

In both figures cluster centers obtained by the K-means method [27] are also presented for comparison. The results show that robust spatial median based method produce almost equal centers while being also robust against noise and gross errors. By comparing the figures we notice that the prototypes have very similar geometry for the both low dimensional projections. This supports the inter-cluster reliability of the clustering result. Figures 7 and 9 indicate that clusters 1 and 5 are somewhat similar. This is logical since they both consists of accidents on high traffic volume roads with multiple roadways. Clusters 3 and 7 are close to each other regarding the type of accident locations. Cluster 3 consist mainly of animal accidents and cluster 7 severe collisions on main roads. Clusters 4 and 6 consist of accidents in built-up

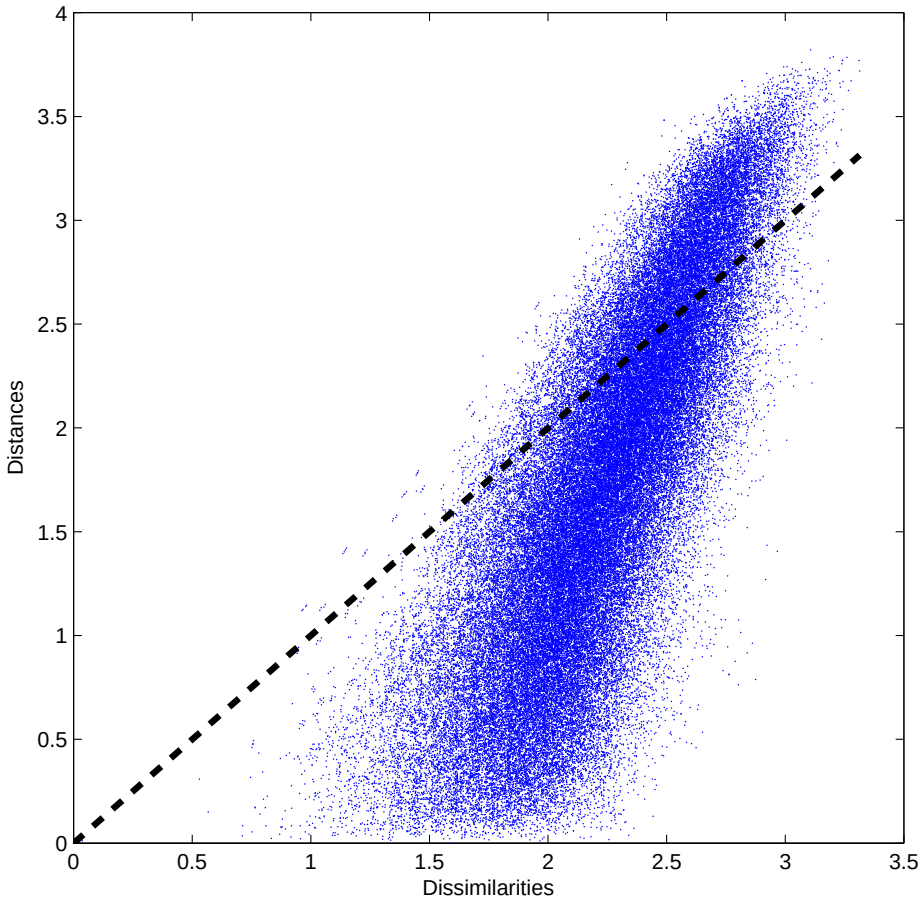

Figure 8: The Shepard plot of the inter-point distances with respect to the original Gower distances in Sammon's mapping.

areas and on small regional roads. Alcohol is quite often involved in these clusters. Cluster 2 is the small gravel road cluster with low maintenance priority where the number of alcohol cases is the highest. In overall, it seems that low dimensional plots provide a baseline for assessing similarities between the clusters.

4.9 Frequent itemsets and association analysis of cluster 7

In addition to robust cluster analysis, we also made an experiment using the frequent itemset and association rules mining methods as a tool in cluster presentation and interpretation. Table 11 shows the maximal frequent itemsets that are generated from cluster 7 with minimum support equal to 0,5. Because the number of rules is huge, we restricted the search to the itemsets of size eight and chose the twenty itemsets with the highest support. Table 11 shows that most of itemsets correspond to the resuls of cluster analysis. Some of the item are so frequent that they repeat

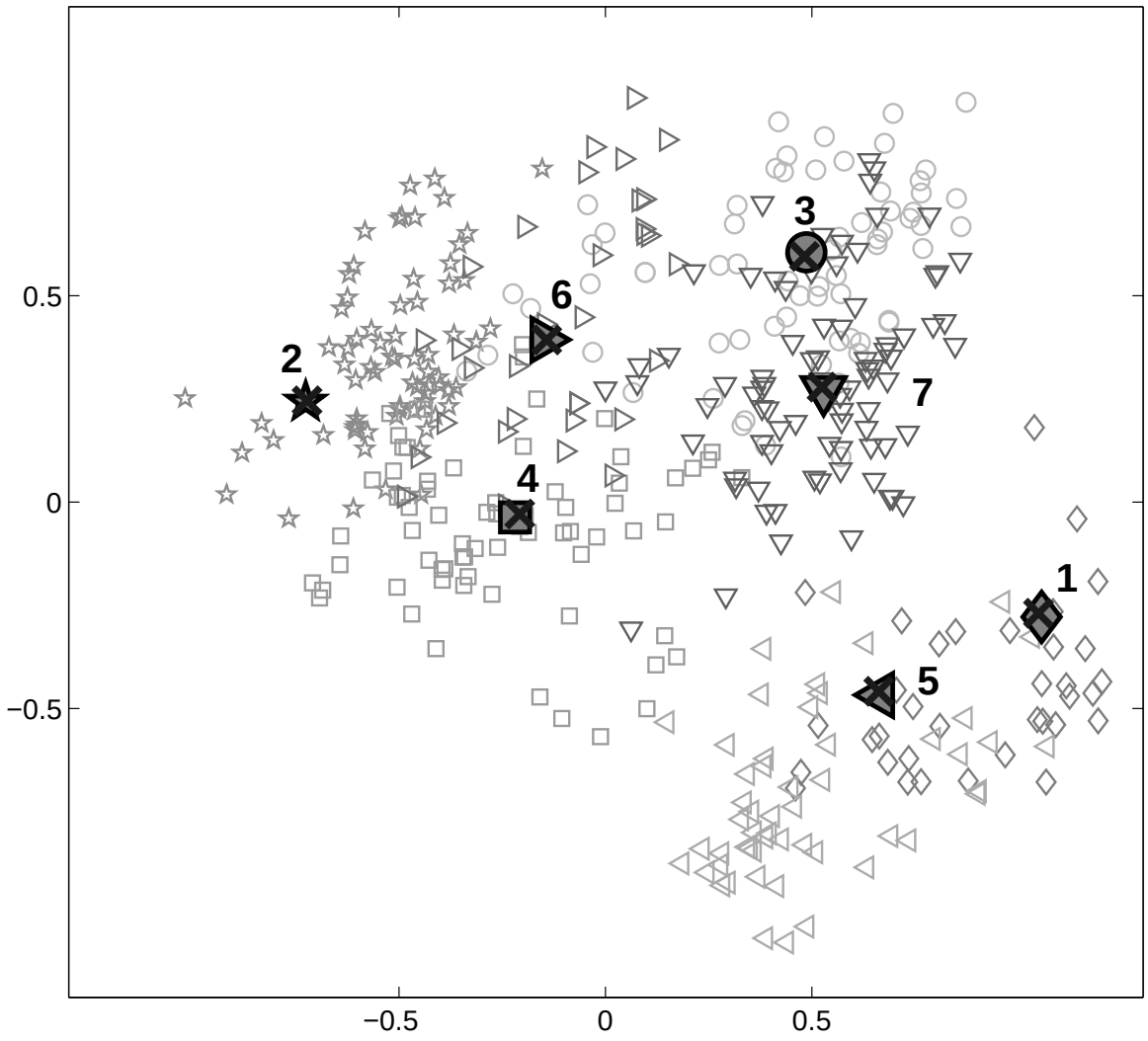

Figure 9: PCA plot of the K-spatialmedians cluster prototypes. K-means prototypes are marked by 'X'. The cluster data points are represented by 0,5% samples of the whole clusters.

in most itemsets while some, possibly interesting items do not show up due to the infrequent occurrences. The frequent itemsets support the interpretation made by cluster analysis. The accidents happen mainly on asphalt pavement single roadway roads. The accident scene is usually roadway and the driver's gender is male.

In order to avoid the most frequent items the generation of maximal itemsets were restricted to the fatal accidents of cluster 7. The results are shown in Table 12. The results show very small difference to the maximal frequent itemsets generated from the whole cluster. Cluster 7 clearly consists of accidents caused by a male driver on class I main roads in good weather conditions and daylight.

In addition to frequent itemsets, we also generated association rules from cluster 7. We ended up to use constraints on consequent part of the rule, because other-

wise it was very difficult to generate non-trivial rules. Without constraints most of the obtained rules are very obvious, for example {*no pedestrian/bicycle way, number of involved = 2, accident type=head-on collision on straight stretch* $\longrightarrow$ *accident class=head-collision*} or {*accident place=roadway, number of involved=2, accident type=head-on collision on straight stretch* $\longrightarrow$ *accident class=head-on collision*}. Since the overall proportion of fatal accidents is even in the most severe cluster only 3,0%, the support for any item sets restricted by the fatality requirement can not be greater than 3,0. Therefore, the support counts are very low. Nevertheless, we can assess different rules by using other measures such as lift metric. Table 13 shows the twenty constrained association rules ordered by their lift values. The rules may seem obvious, but they indeed describe cluster 7 quite well. The first rule is a very good example of accidents that lead to the highest fatality proportion of cluster 7. A single roadway road with relatively high speed limit (100km/h) and a male driver colliding with a heavy vehicle on straight stretch of a road is, without questions, a fatal combination. The second rule is almost equal with the first one. Overall, the twenty rules confirm that a head-on collision with a heavy truck on main roads where the driving speed are high are the most risk accidents. If we consider the interestingness or novelty of the found rules, this did not reveal any unexpected information about the data. While the rules may sound trivial findings, they show that combined clustering and rule mining can reveal the most dangerous conditions. With the used accident data it seems not possible to generate more detailed or unexpected rules from accident conditions, because the data does not contain very detailed information about accident locations or preceding moments that led to accidents.

4.10 Discussion

The obtained results show undoubtedly that using descriptive data mining methods, it is possible to create reasonable knowledge from the road traffic accident data. While the results seem quite obvious, it is also significant that they are very reasonable. It is interesting to observe that severity of accidents varies between the clusters. The considered clusters were still rather large. When the accidents in the Middle-Finland region were plotted on the road map, they were still difficult interpret due to the large number of accidents per cluster. One should perhaps select a small sample of accidents from a high risk cluster, for example, restrict to a certain region, and analyze those accident locations more thoroughly. The sample could consists, for example, of one hundred accidents that are most similar to the cluster prototype. Another option is to construct nested structure by clustering the current clusters hierarchically. Then one could inspect the most risky clusters and not only concentrate on the one with fatal accidents, but also on interesting subsets of non-fatal accidents that are similar to the fatal ones.

Different approaches of data mining clearly support each other. The attribute ranking graph, low-dimensional projections together with frequent itemsets and association rules gives straightforward information about cluster similarities and interesting and meaningful attributes. In this report we do not analyze all the clusters

thoroughly, but we have shown that the most significant features can be recognized without exploring the complete cluster data. After the explorative cluster analysis, more thorough investigation on the most important clusters can be done using the categorized cluster data tables.

5 Conclusion

Overall, most fatal accidents seem to happen conditions on single roadway main roads outside built-up areas where the speed limit varies typically between 80-100km/h. Multi-lane motor- or semi-motor highways are much safer, because the head-on collisions are prevented and side-collisions are not as critical as head-on impacts. Aged drivers have relatively large contribution to the high risk accidents in class I and II main roads. Despite animal accidents are common in Finland, the risk of death is not very high in such accidents. Young drivers are more often involved in the accidents that happen in built-up areas or small roads. Alcohol is often involved in the accidents caused by a young driver and the accidents are typically single vehicle run-off accidents. Figures 21-26 show the cluster-wise distributions by the age groups. The results show that young drivers have clearly the highest number of accidents, but the proportions of middle-aged drivers are relatively higher in fatal accidents than in the non-fatal accidents.

With the current data it is possible to recognize the risky road segments and the road user groups responsible for accidents in certain environments. However, it is not possible to find out very strict details for enhancing road construction plans from this data. More detailed location specific information from accident locations and situations are needed. The lack of detailed accident-specific data hinders the analysis from the road network engineering point of view, because it is currently difficult to analyze the local defects in a particular road segment that might cause further accidents. For example, the data contain no information about seasonal speed limits, "no passing" zones, roundabouts, priority, median barriers, uphill/downhill degrees, curve radius, gravelling, salting, speeding, traffic rule violations (use of seat belts or helmet, and aggressive/reckless/careless driving), type of vehicle (cross-country vehicle, trailer, etc.), vehicle defects, protective devices (airbag), status/type of driving licence, number of years with licence, apparent suicide cases, sleepiness, etc. The literature review shows that many of these attributes have been available in other international case studies. Without all this information it is difficult to evaluate the role of road building, deliberateness of accidents and so on. This means that there remains a lot of accidents that are not caused by the road conditions. On the other hand, there are accident that are perhaps caused by insufficient road traffic plans.

Although descriptive data mining methods are clearly able to uncover reasonable information from the selected traffic accident data set, the results remain at very general level so that they do not yet provide much previously unknown new knowledge for the traffic accident experts. Therefore, more detailed data is needed

for finding novel facts from data. Data mining seems to produce very understandable and useful results.

Acknowledgement

This work was supported by Finnish Road Administration (Keski-Suomi region) under the project Tie2Louhi. The authors are thankful to director Seppo Kosonen for his guidance during the project.

References

[1] *Global status report on road safety: time for action, WHO*, 2009.

[2] I. ABUGESSAISA, *Analytical tools and information-sharing methods supporting road safety organizations*, PhD thesis, Link

[3] R. AGRAWAL AND R. SRIKANT, *Fast algorithms for mining association rules in large databases*, in VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases, San Francisco, CA, USA, 1994, Morgan Kaufmann Publishers Inc., pp. 487–499.

[4] T. K. ANDERSON, *Kernel density estimation and k-means clustering to profile road accident hotspots*, Accident Analysis Prevention, 41 (2009), pp. 359 – 364.

[5] S. ÄYRÄMÖ, *Knowledge Mining using Robust Clustering*, PhD thesis, University of Jyväskylä, 2006.

[6] S. ÄYRÄMÖ, T. KÄRKKÄINEN, AND K. MAJAVA, *Robust refinement of initial prototypes for partitioning-based clustering algorithms*, in Recent Advances in Stochastic Modeling and Data Analysis, C. H. Skiadas, ed., World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2007, ch. 11, pp. 473–482.

[7] V. BARNETT AND T. LEWIS, *Outliers in statistical data*, John Wiley & Sons, 2 ed., 1984.

[8] P. S. BRADLEY AND U. M. FAYYAD, *Refining initial points for K-Means clustering*, in Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 1998, pp. 91–99.

[9] M. M. CHONG, A. ABRAHAM, AND M. PAPRZYCKI, *Traffic accident analysis using machine learning paradigms*, Informatica (Slovenia), 29 (2005), pp. 89–98.

[10] A. DEMPSTER, N. LAIRD, AND D. RUBIN, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society. Series B (Methodological), 39 (1977), pp. 1–38.

[11] B. DEPAIRE, G. WETS, AND K. VANHOOF, *Traffic accident segmentation by means of latent class clustering*, Accident Analysis & Prevention, 40 (2008), pp. 1257 – 1266.

[12] M. H. DUNHAM, *Data mining - introductory and advanced topics*, Pearson Education Inc, Upper Saddle River, New Jersey, USA, 2003.

[13] M. ESTER, H.-P. KRIEGEL, J. SANDER, AND X. XU, *A density-based algorithm for discovering clusters in large spatial databases with noise*, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), E. Simoudis, J. Han, and U. M. Fayyad, eds., AAAI Press, 1996, pp. 226–231.

[14] U. FAYYAD, G. PIATETSKY-SHAPIRO, AND P. SMYTH, *The KDD process for extracting useful knowledge from volumes of data*, Communications of the ACM, 39 (1996), pp. 27–34.

[15] K. GEURTS, I. THOMAS, AND G. WETS, *Understanding spatial concentrations of road accidents using frequent item sets*, Accident Analysis & Prevention, 37 (2005), pp. 787 – 799.

[16] K. GEURTS, G. WETS, T. BRIJS, AND K. VANHOOF, *Clustering and profiling traffic roads by means of accident data*, in Proceedings of the European Transport Conference 2003, Strasbourgh (France), October 8-10, 2003.

[17] J. GOWER, *A general coefficient of similarity and some of its properties*, Biometrics, 27 (1971), pp. 857–871.

[18] J. HAN AND M. KAMBER, *Data mining: concepts and techniques*, Morgan Kaufmann Publishers, Inc., 2006.

[19] D. HAND, H. MANNILA, AND P. SMYTH, *Principles of Data Mining*, MIT Press, 2001.

[20] T. HASTIE, R. TIBSHIRANI, AND J. FRIEDMAN, *The elements of statistical learning: Data mining, inference and prediction*, Springer-Verlag, 2001.

[21] P. HUBER, *Robust statistics*, John Wiley & Sons, 1981.

[22] A. K. JAIN AND R. C. DUBES, *Algorithms for clustering data*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[23] G. JOST, M. POPOLIZIO, R. ALLSOP, AND V. EKSLER, *3rd road safety pin report: 2010 on the horizon*, report, European Transport and Safety Council (ETSC), 2009.

[24] T. KÄRKKÄINEN AND S. ÄYRÄMÖ, *Robust clustering methods for incomplete and erroneous data*, in Proceedings of the Fifth Conference on Data Mining, 2004, pp. 101–112.

[25] ——, *On computation of spatial median for robust data mining*, in Proceedings of Sixth Conference on Evolutionary and Deterministic Methods for Design, Optimisation and Control with Applications to Industrial and Societal Problems (EUROGEN 2005), R. Schilling, W. Haase, J. Periaux, and H. Baier, eds., 2005.

[26] R. J. LITTLE AND D. B. RUBIN, *Statistical analysis with missing data*, John Wiley & Sons, 1987.

[27] J. MACQUEEN, *Some methods for classification and analysis of multivariate observations*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[28] M. PEDEN, R. SCURFIELD, D. SLEET, D. MOHAN, A. A. HYDER, E. JARAWAN, AND C. MATHERS, *World report on road traffic injury prevention*, 2004.

[29] G. PIATETSKY-SHAPIRO, *Knowledge discovery in real databases: A report on the IJCAI-89 workshop*, AI Magazine, 11 (1991), pp. 68–70.

[30] D. PYLE, *Data preparation for data mining*, Morgan Kaufmann Publishers, Inc., 2001.

[31] K. SIRVIÖ AND J. HOLLMÉN, *Spatio-temporal road condition forecasting with markov chains and artificial neural networks*, in HAIS '08: Proceedings of the 3rd international workshop on Hybrid Artificial Intelligence Systems, Berlin, Heidelberg, 2008, Springer-Verlag, pp. 204–211.

[32] S. SOHN AND H. SHIN, *Pattern recognition for road traffic accident severity in Korea*, Ergonomics, 44 (2001), pp. 107–117.

[33] S. Y. SOHN AND S. H. LEE, *Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea*, Safety Science, 41 (2003), pp. 1 – 14.

[34] D. STENMARK, *Information vs. knowledge: The role of intranets in knowledge management*, in Proceedings of the 35th Hawaii International Conference on System Sciences, IEEE, January 2002.

[35] P.-N. TAN, M. STEINBACH, AND V. KUMAR, *Introduction to data mining*, Addison-Wesley, 2005.

[36] T. TESEMA, A. ABRAHAM, AND C. GROSAN, *Rule mining and classification of road accidents using adaptive regression trees*, International Journal of Simulation Systems, Science & Technology, 6 (2005), pp. 80–94.

[37] I. TUOMI, *Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge management and organizational memory*, Journal of Management Information Systems, 16 (1999), pp. 107–121.

Appendix 1: Tables

Hour/Day	Mo	Tu	We	Th	Fr	Sa	Su
0	7	5	5	5	5	7	7
1	7	5	5	5	5	7	7
2	7	5	5	5	5	7	7
3	7	5	5	5	5	7	7
4	7	5	5	5	5	7	7
5	7	5	5	5	5	7	7
6	1	1	1	1	1	8	8
7	1	1	1	1	1	8	8
8	1	1	1	1	1	8	8
9	1	1	1	1	1	8	8
10	2	2	2	2	2	9	9
11	2	2	2	2	2	9	9
12	2	2	2	2	2	9	9
13	2	2	2	2	2	9	9
14	3	3	3	3	3	10	10
15	3	3	3	3	3	10	10
16	3	3	3	3	3	10	10
17	3	3	3	3	3	10	10
18	4	4	4	4	6	6	6
19	4	4	4	4	6	6	6
20	4	4	4	4	6	6	6
21	4	4	4	4	6	6	6
22	5	5	5	5	7	7	7
23	5	5	5	5	7	7	7

Table 5: Encoding of the accident time.

DRIVER INFORMATION								
	Cl1	**Cl2**	**Cl3**	**Cl4**	**Cl5**	**Cl6**	**Cl7**	**ALL**
CONSEQUENCES								
number of accidents	6052	15036	15279	13387	6427	8587	18741	83509
fatals %	1,5	1,3	0,3	1,0	0,5	1,9	3,0	1,4
injuries %	18,6	20,9	4,9	27,4	19,1	20,5	28,0	20,3
GENDER %								
male	70,8	71,7	74,3	67,5	71	70,9	71,1	71,2
female	25,7	23,9	21,5	28,5	25,6	25,4	25,9	25,1
missing	3,5	4,4	4,2	4	3,3	3,7	3	3,8
DRUNKEN DRIVING %								
Involved	7,8	15,4	0,3	10,2	5,1	11,6	9,3	8,7
AGE %								
missing	12,2	15,0	9,7	12,2	9,6	13,7	11,2	12,0
< 18	0,4	2,7	0,1	5,4	1,3	1,0	1,4	1,9
18-19	7,1	10,2	3,6	8,9	6,1	7,3	7,3	7,3
20-24	13,2	10,8	8,7	10,8	11,9	10,8	11,7	10,9
25-34	20,9	15,2	16,9	14,9	19,5	14,9	16,5	16,5
35-54	32,1	30,4	40,0	28,0	33,8	32,2	30,4	32,2
55-64	10,0	9,9	15,7	10,4	11,0	11,9	11,4	11,7
> 64	4,2	5,8	5,2	9,4	6,8	8,2	10,2	7,4
average	37,9	38,4	42,4	39,4	39,7	41,2	41,1	40,3
standard deviation	23,5	25,8	23,1	27,5	24,5	26,9	27,2	25,9
Percentiles all								
5 %	19	18	20	16	18	18	18	18
25 %	25	23	30	23	26	26	25	26
50% median	36	37	43	37	38	41	39	39
75 %	49	50	54	53	51	54	54	53
95 %	64	68	65	74	68	71	73	70
Percentiles fatals								
5 %	18	17	18	16	19	18,6	19	18
25 %	25	21,8	24	23	23	28,8	27	25
50% median	37,5	34	30	33,5	39,5	44	40	39
75 %	50	50	50,5	49	57	56	54	53
95 %	71	72	64,5	77,8	86	74	76,2	75,2

Table 6: Driver information.

ROAD INFORMATION							
	Cl1	Cl2	Cl3	Cl4	Cl5	Cl6	Cl7
FUNCTIONAL CLASS %							
1 Class I main road	93,9	0	55,6	3,2	41,6	1,1	74,1
2 Class II main road	5,8	0,1	19,6	7,2	30,2	13,3	15,5
3 Regional highway	0	1	21,4	38,1	26,8	81,1	9,7
4 Connecting road	0,4	98,9	3,3	51,5	1,3	4,5	0,7
MAINTENANCE CLASS %							
1 constantly bare (1s)	98,8	0	28,2	9,5	92,9	0,2	37,6
2 snow constantly removed excluding nighttime	1,2	1,3	33,7	16,3	3,5	1,6	32,5
3 snow not removed occasionally (1b)	0	5	33,9	37,2	2,3	22,5	28,9
4 class 3 but inside built-up areas (T1b)	0	0,5	0,1	16,2	1,2	0,1	0,1
5 mainly snowy (II)	0	44,8	3,6	19,7	0	73,1	0,8
6 snowy (graveling only in extreme conditions)	0	48,5	0,5	1	0	2,6	0,1
PAVEMENT							
1 Asphalt	99,9	17,2	98,2	96,7	99,8	29,6	98,5
2 Oil gravel	0,1	54,9	1,8	3,2	0,2	69,5	1,4
3 Gravel	0	27,8	0,1	0,1	0	0,8	0
MOTOR-/SEMIMOTORWAY %							
motorway	87,2	0	0	0	4,9	0	0
semi-motorway	9,3	0	0,2	0	0,6	0	0,1
NUMBER OF ROADWAYS %							
1	9,8	100	99,1	97,2	8,2	99,9	98,5
2	90,2	0	0,9	2,8	91,8	0,1	1,5
PEDESTRIAN/BICYCLE WAY %							
yes	4,5	2,3	7,1	77,4	82,7	2,4	21,8
ARTERIAL ROAD %							
yes	90,7	0	21,2	0,8	40	0	32,4
SPEED LIMIT %							
≤40 km/h	0,5	2,3	0	22,5	1,3	0,3	0,4
50 km/h	4,2	9,1	0,2	29,5	17,3	1,3	2,5
60 km/h	3,4	21,7	1,9	36,2	24,9	7,7	12,8
70-80 km/h	17,3	66,8	59,4	11,7	50,5	83,8	61,3
100 km/h	51,5	0	38,5	0	5,9	7	22,9
120 km/h	23	0	0	0	0	0	0
ROAD WIDTH (percentiles)							
5 %	126	51	76	66	111	65	81
25 %	221	61	81	76	170	71	83
median 50%	235	66	91	81	186	71	96
75 %	247	71	101	96	215	76	106
95 %	288	76	126	126	258	91	133
SIGHT DISTANCE (average of percents)							
150m	99,6	69,4	99,2	91,7	97,9	91,7	99,5
300m	88,3	24,9	83,7	55,5	76,4	53,4	85,5
460m	64,1	9,1	57,5	27,5	47,0	26,3	59,9

Table 7: Road information.

TRAFFIC AND ENVIRONMENT INFORMATION							
	Cl1	**Cl2**	**Cl3**	**Cl4**	**Cl5**	**Cl6**	**Cl7**
DAILY TRAFFIC VOLUME (percentiles)							
5 %	8896	66	1090	814	9526	258	1422
25 %	15418	159	2382	2002	16879	573	2953
median 50%	22059	376	3653	3607	23860	880	4994
75 %	30852	706	5621	6491	44868	1347	7453
95 %	48666	1605	10365	12995	63395	2722	12407
DAILY HEAVY TRAFFIC VOLUME (percentiles)							
5 %	917	3	74	24	570	15	117
25 %	1470	7	198	73	1204	35	283
median 50%	1962	17	426	153	1815	56	545
75 %	2605	33	688	330	3303	89	840
95 %	3818	71	1158	908	6442	186	1379
POPULATION DENSITY %							
missing	0	0,3	0,1	1,7	0,2	0,1	0,1
undefined	99,9	0	0,8	1,7	96,1	0,1	1,2
1 built-up area traffic sign	0	2,5	0,1	54,4	1	0,4	1,5
2 statistical built-up area	0	12	8	28,2	1,5	5,3	32,9
3 built-up area A > 60/km^2	0	5,3	4,2	4,3	0,6	3,3	6,2
4 built-up area B 30-60/km^2	0	13,1	8,8	4,3	0,2	9,5	10,7
5 built-up area B 15-30/km^2	0	16,7	13,5	2,7	0,2	15,5	11,3
6 sparsely populated area < 15/km2	0	50,2	64,5	2,6	0,2	65,8	36,1

Table 8: Road information.

ACCIDENT INFORMATION							
	Cl1	Cl2	Cl3	Cl4	Cl5	Cl6	Cl7
ACCIDENT CLASSES %							
1 single vehicle collision	41,6	44,2	0,2	27,1	14,6	41,8	39,7
2 turning	1,6	2	0,1	11,6	8,5	3,9	14,2
3 overtaking	14,8	0,8	0,4	2,2	17,1	2	6
4 intersection	2,3	4,1	0,1	23,3	12,4	3,3	13,1
5 head-on collision	1,7	8,6	0,6	3,8	0,9	8	7,2
6 head-to-tail collision	12,2	0,5	0,2	6,3	34,6	1,4	8,9
7 moped	0,1	2,1	0,1	7,5	1,3	0,9	1,8
8 cycle	0,2	0,9	0	6,7	1,8	1	1,6
9 pedestrian	0,5	0,8	0,1	2,6	0,7	0,8	0,8
10 moose	4,9	8,8	39	0,9	0,9	17	0,3
11 deer	10,2	23,2	55	3,5	1,7	15,1	0,3
12 other animal	1,4	1,6	3,8	0,8	0,2	2,1	0,6
13 other	8,5	2,4	0,3	3,7	5,4	2,8	5,4
ACCIDENT SCENE %							
roadway	84,3	99,3	99,8	89,3	88,5	99	96,6
crosswalk	0,1	0,1	0	3,9	1	0	0,3
pedestrian/bicycle way	0	0,1	0	5,2	1,6	0	1
grade-separated junction ramp	13,5	0	0	0,2	7,3	0	0,8
TRAFFIC LIGHTS %							
In function	2,4	0	0	6,3	35,7	0,1	2,5
NUMBER OF INVOLVED %							
1	41,6	44,9	0,2	28,1	14,7	42	39,9
2	49,8	53,9	96,5	67,4	69,6	55,4	52,1
3	5,9	1,1	2,8	4,1	11,3	2,3	6,8
≥4	2,8	0	0,4	0,5	4,4	0,3	1,2
HEAVY VEHICLE %							
Involved	17	8,8	13,8	10,3	24,4	13	22,8
TYPE OF ACCIDENT %							
90 animal accident	16,5	33,6	97,8	5,3	2,8	34,1	1,3
08 rear-end collision with a vehicle stopped due to an obstruction	2,9	0,1	0	3,3	17,3	0,4	3,2
06 head-to-tail collision with a braking vehicle	5,9	0,3	0,1	2,6	13,7	0,8	4,2
80 run-off-road to right on straight stretch	7,7	7,4	0	4,9	2,8	8,6	11,9
40 hitting from intersecting directions	0,6	1,2	0	11,8	4	1	5,8
84 run off the right side in left curve	5,8	11,1	0	4,4	1,2	9	4,9
80 run-off the left side on straight stretch	5,5	5	0,1	3	2,3	6,8	7,8
84 run-off the left side in right curve	7,4	7	0	2,3	1,9	4,8	2,8
03 left lane change	4,3	0	0	0,4	6,6	0,1	0,3
02 right lane change	2,1	0	0	0,4	6,5	0	0,2
53 left turn to the front or side of the intersecting vehicle	0,6	1,7	0	6,2	4,3	1,2	3,9
82 run-off the right side in right curve	5,9	4,2	0	1,8	1,4	4	2,8
12 head-to-tail collision in left turn	0,1	0,4	0	3,6	1,1	1	5,1
21 head-on collision in curve	0,9	5,1	0,3	2,3	0,5	4	2,8
00 overtaking	5,1	0,5	0,3	1,1	1,4	1,3	3,6

Table 9: Accident information.

WEATHER INFORMATION							
	Cl1	**Cl2**	**Cl3**	**Cl4**	**Cl5**	**Cl6**	**Cl7**
TEMPERATURE							
missing	5,9	4,9	5,9	4,9	8,9	4,6	3,6
< -7°C	5,5	6,6	4,1	7,1	5,4	9,5	7,4
-7...+2°C	36,4	30,8	28,9	31,5	25,1	40,7	36
> +2°C	52,1	57,6	61,1	56,5	60,6	45,2	53
TEMPERATURE (percentiles)							
5 %	-8	-10	-7	-10	-8	-10	-10
25 %	-1	-1	0	-1	0	-2	-2
median 50%	5	6	6	5	8	2	4
75 %	13	14	12	15	15	11	15
5 %	21	22	20	22	23	20	22
WEATHER							
1 clear	33,6	38,9	31,8	40,6	41,3	33,5	38,2
2 fair clouds	38,8	44,4	48,7	40,8	37,7	46,7	39,8
3 fog	1,5	1,8	3	1,2	1,1	2	1,4
4 rain	11,1	7,8	10,8	9,3	11,9	7,4	9
5 snowfall	10,9	4,4	3,4	5,5	4,7	6,8	8,3
6 sleeting	2,9	1,6	1,2	1,9	1,9	2,5	2,7
ROAD CONDITION							
1 bare, dry	45,9	49,1	55,3	48,7	52	38,8	47
2 bare, wet	21,1	16,4	26,2	19,2	27,2	13,8	18,2
3 ruts filled with water	1,6	0,3	0,6	0,5	1	0,4	0,7
4 snowy	6,4	11,5	4,1	8,7	4,4	10,8	6,6
5 slushy	5,3	3,1	1,4	3,1	2,7	4,7	4,6
6 icy	16,6	17,2	8,6	16,4	9	28,3	19
7 bare ruts	2,2	1,1	2,7	2,6	2,3	2,2	3,4
LIGHTNESS							
1 daylight	52,4	52,7	26,3	64,6	68,6	48,5	65,3
2 dusk/dawn	11,3	12,7	17,6	8	7	12,1	7,7
3 dark - not lighted	16,4	30,9	49,3	7,3	5,2	36,9	15,5
4 dark - lighted roadway	19,6	3,1	6,4	19,7	18,6	1,9	11,3

Table 10: Accident types.

Appendix 2: Graphs

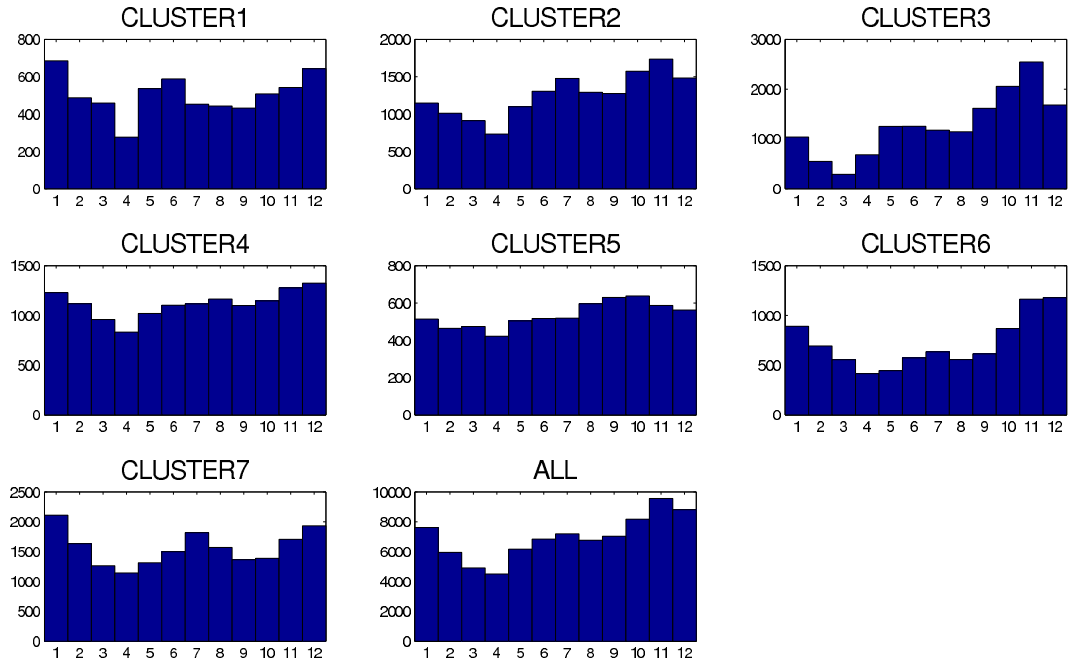

Figure 10: Accident time distributions by months (all accidents).

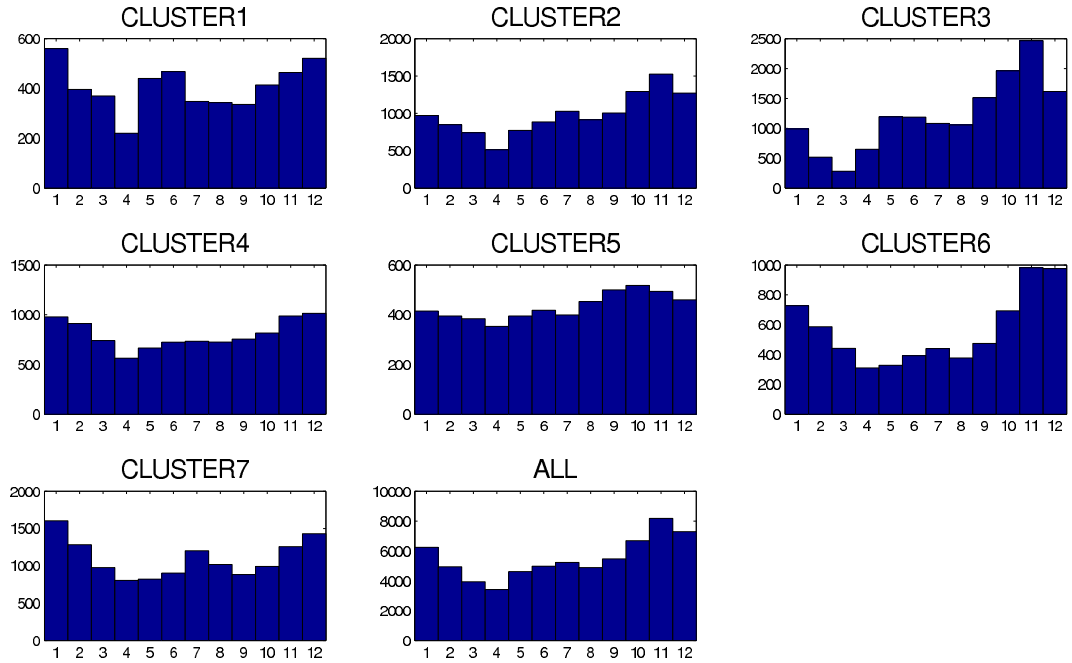

Figure 11: Accident time distributions by months (accidents without personal injures).

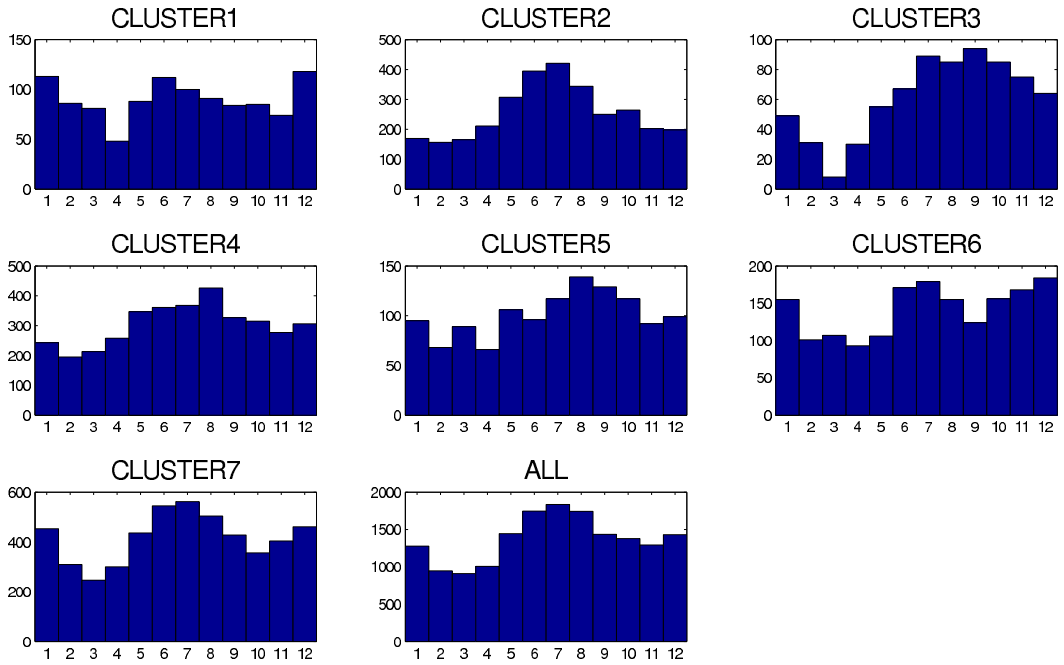

Figure 12: Accident time distributions by months (non-fatal injurious accidents).

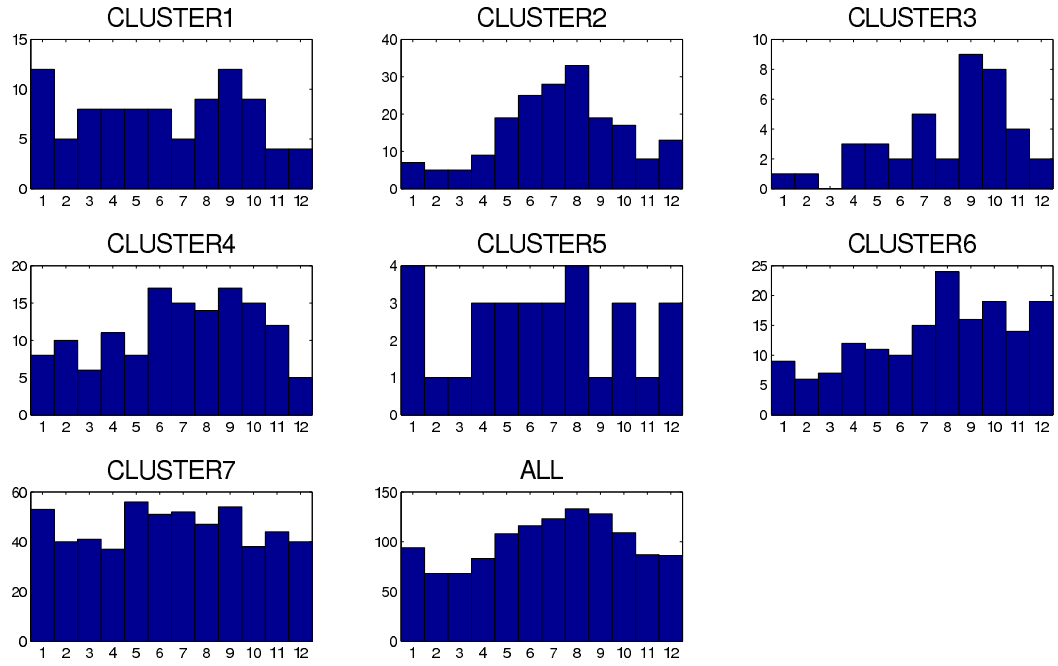

Figure 13: Accident time distributions by months (fatal accidents).

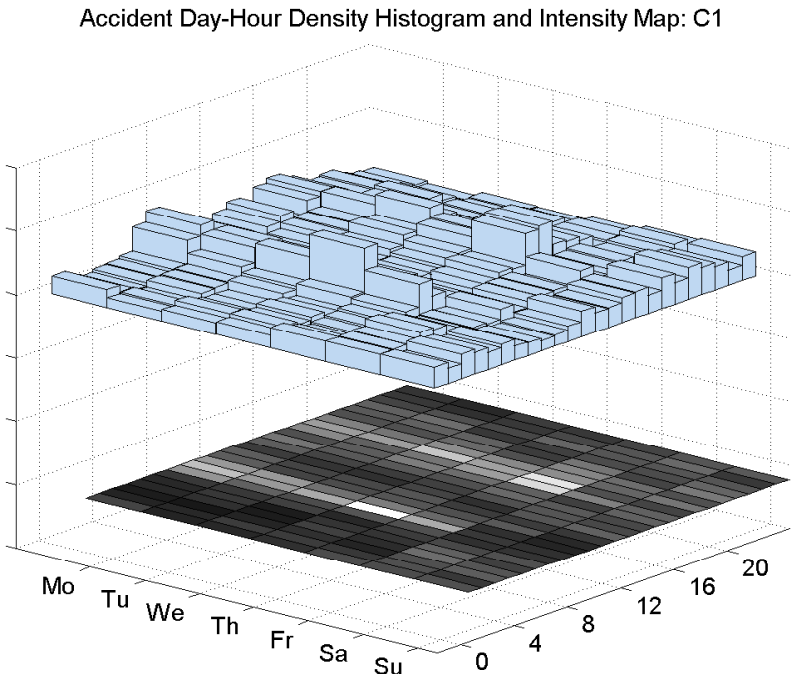

Figure 14: Time distribution of accident occurrence times by weekdays and hours in cluster 1.

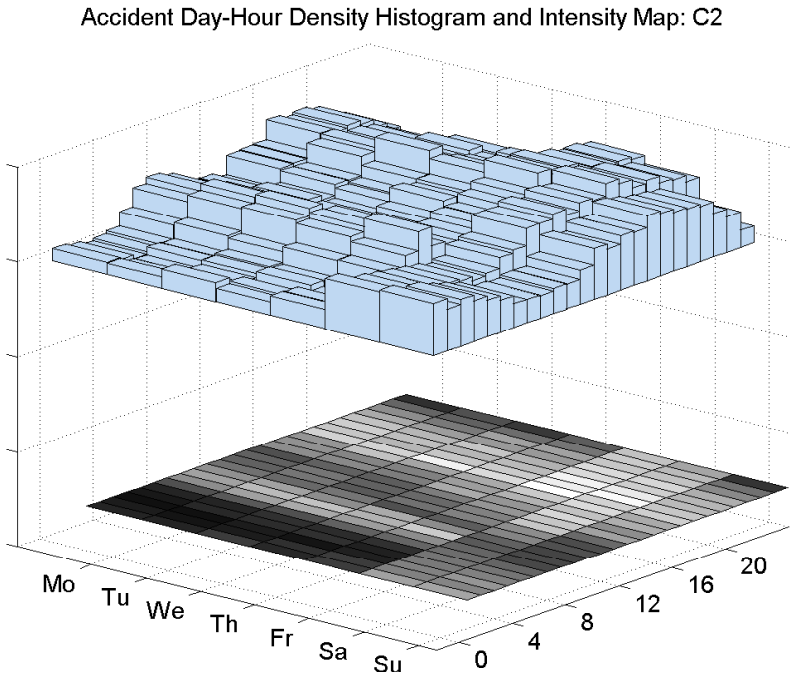

Figure 15: Time distribution of accident occurrence times by weekdays and hours in cluster 2.

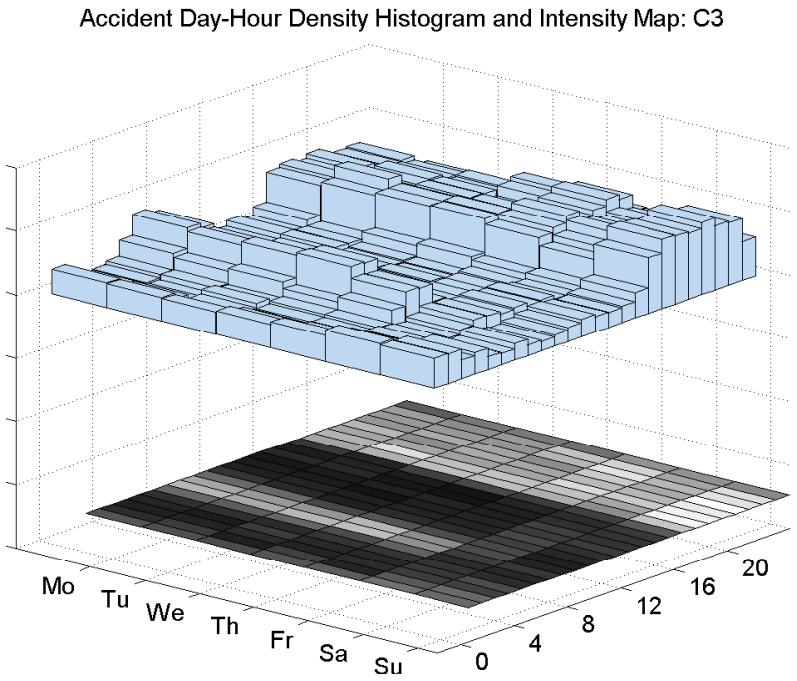

Figure 16: Time distribution of accident occurrence times by weekdays and hours in cluster 3.

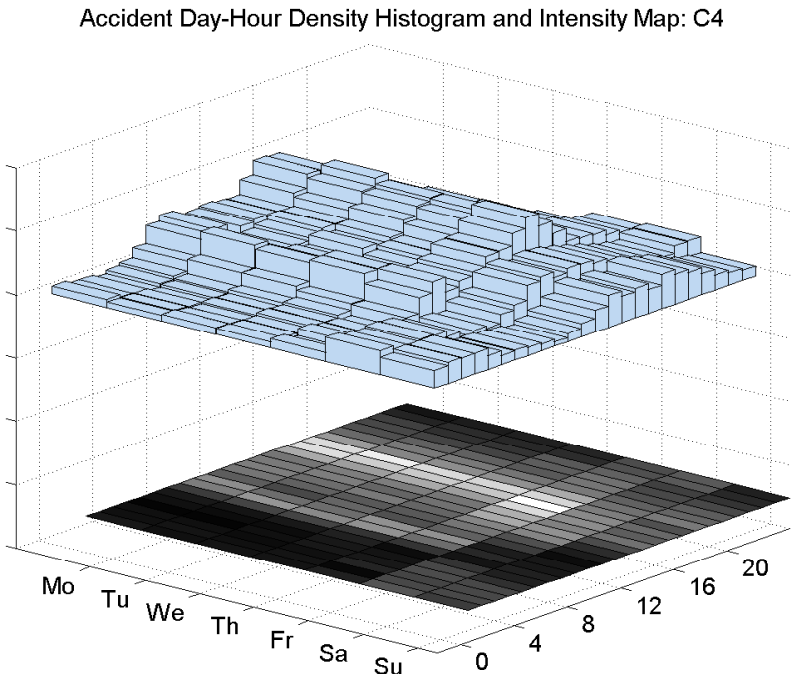

Figure 17: Time distribution of accident occurrence times by weekdays and hours in cluster 4.

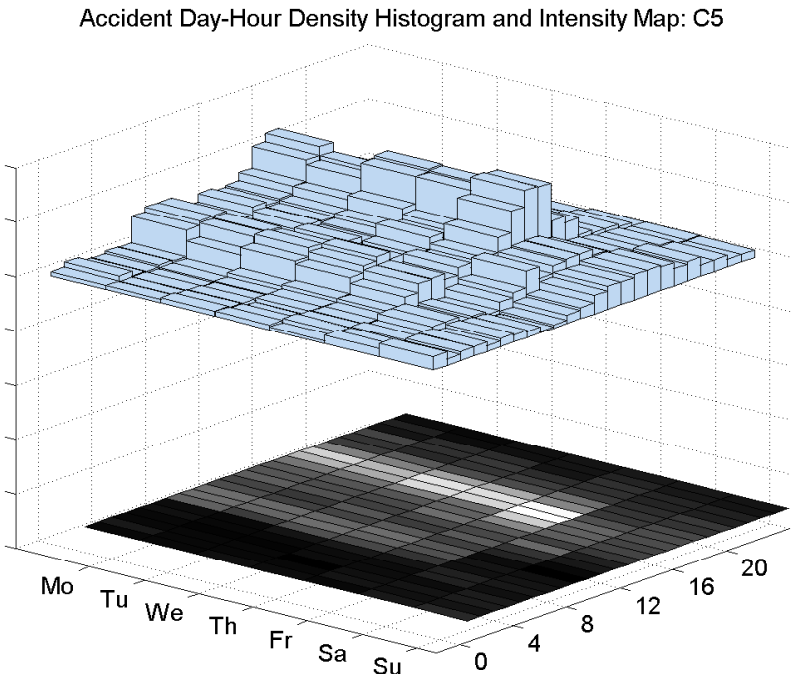

Figure 18: Time distribution of accident occurrence times by weekdays and hours in cluster 5.

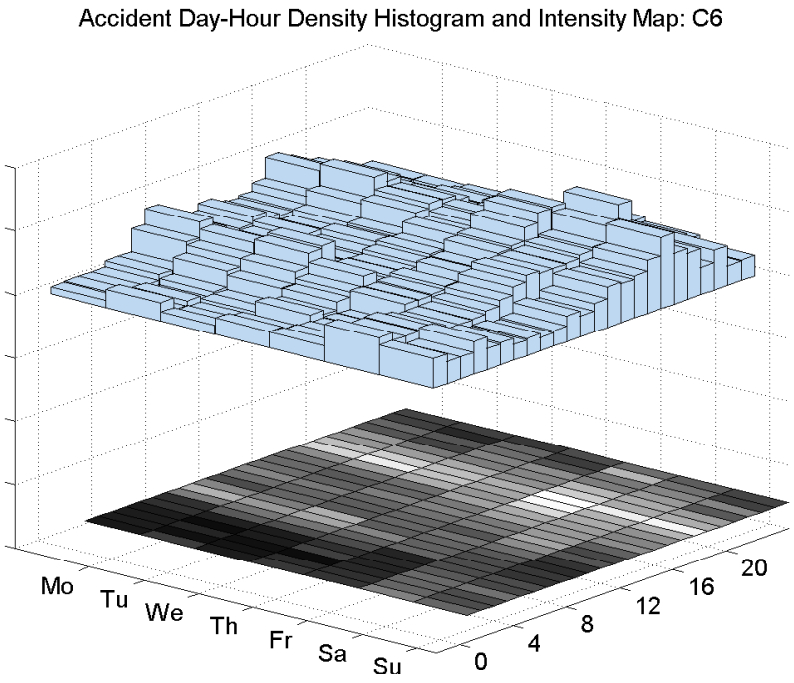

Figure 19: Time distribution of accident occurrence times by weekdays and hours in cluster 6.

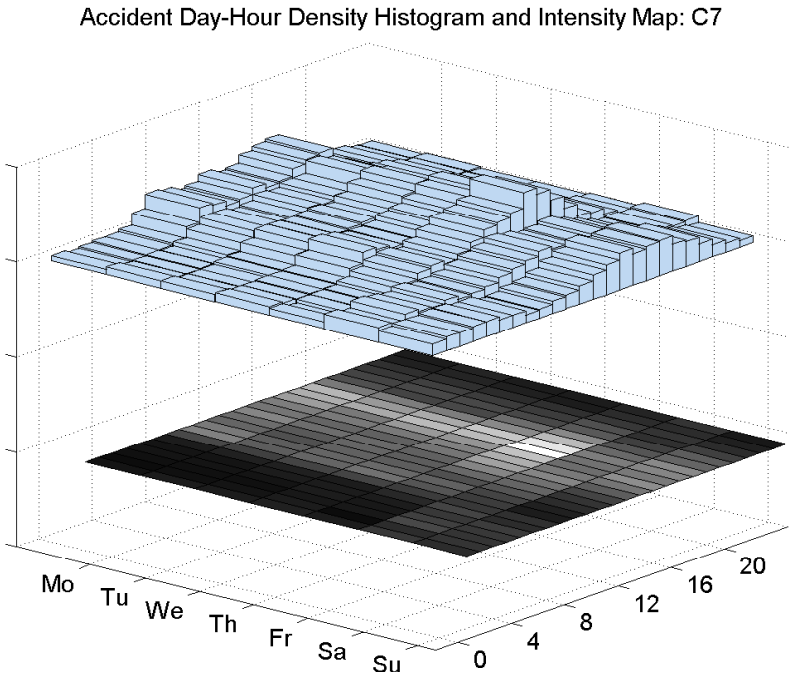

Figure 20: Time distribution of accident occurrence times by weekdays and hours in cluster 7.

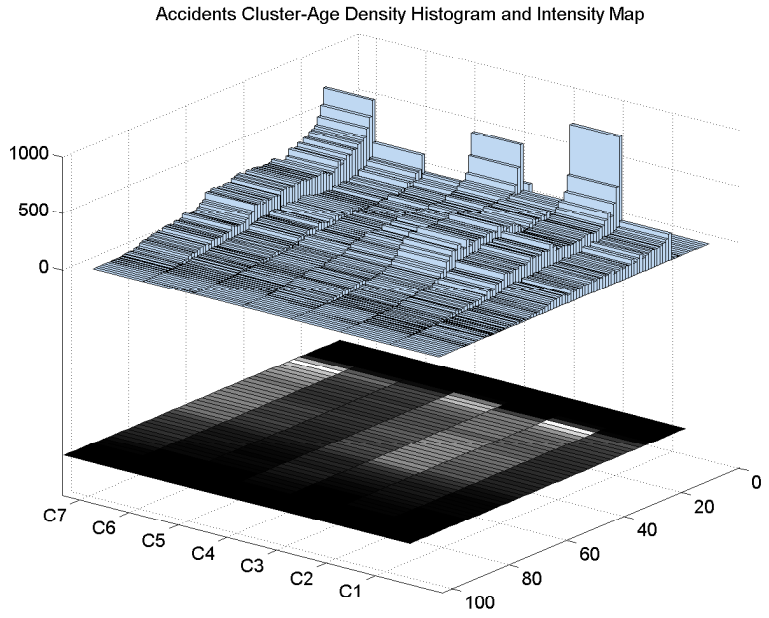

Figure 21: Driver's age distributions in the accidents.

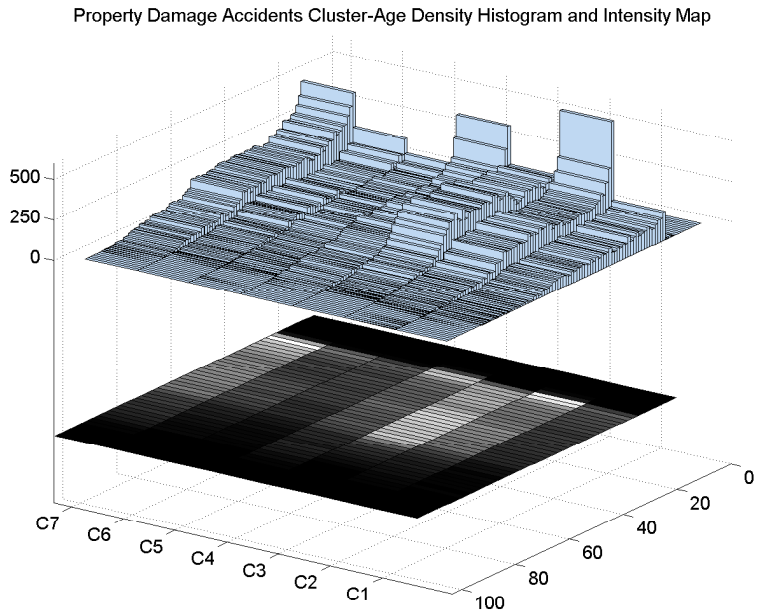

Figure 22: Driver's age distributions in the accidents not causing personal injuries.

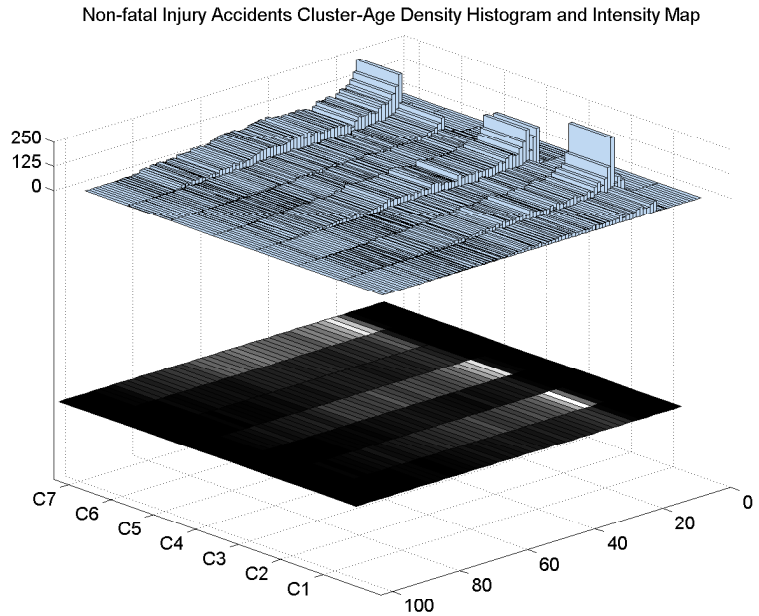

Figure 23: Driver's age distributions in non-fatal injurious accidents.

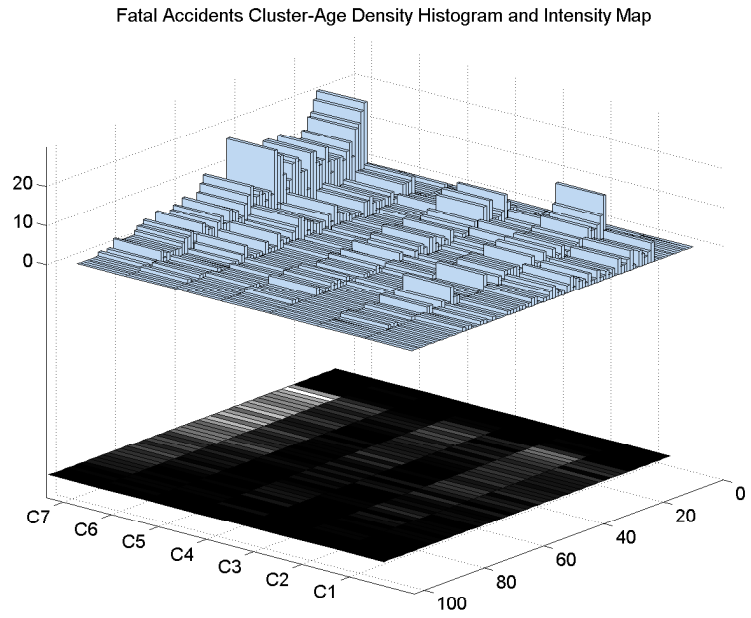

Figure 24: Driver's age distributions in fatal accidents.

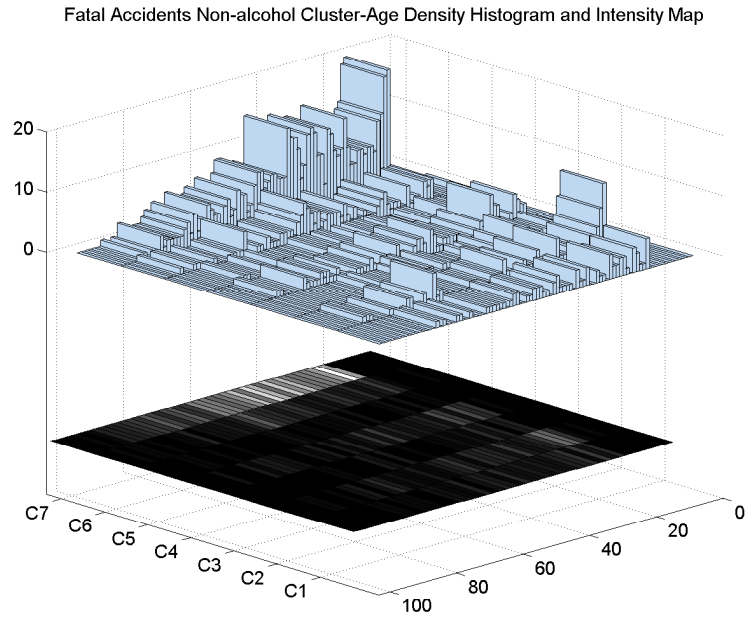

Figure 25: Driver's age distributions in fatal accidents excluding drink drivers.

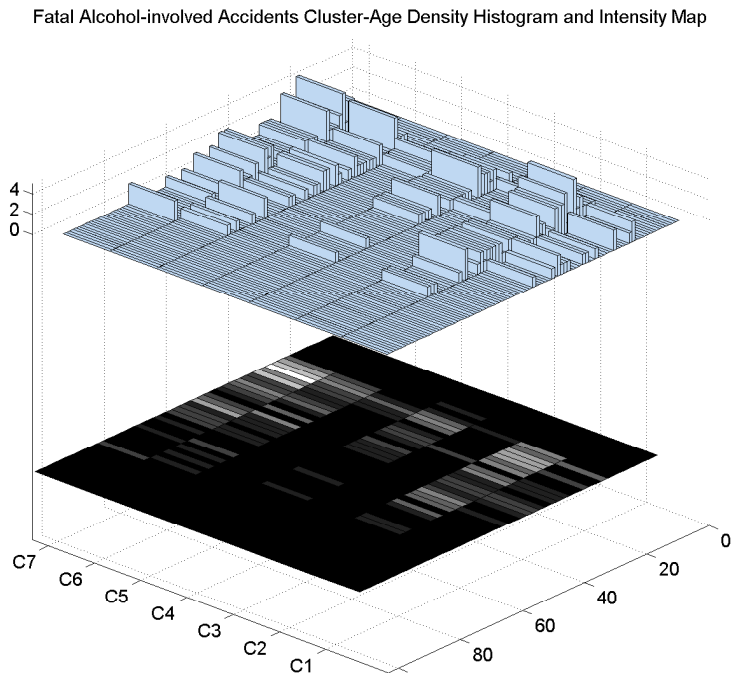

Figure 26: Age distributions in fatal accidents by drink drivers.

Appendix 3: Itemsets and association rules

Itemset	Sup	Items								
1	0.628	no ped/bic way	speed_lim_type=1	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
2	0.621	sd 300m=75-100%	speed_lim_type=1	acc_place=roadway	no traffic lights	#roadways=1	pavement=asphalt	no motorway	sd 150m=75-100%	
3	0.606	heavy vehicle=no	speed_lim_type=1	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
4	0.585	class I main road	speed_lim_type=1	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
5	0.563	sd 300m=75-100%	no ped/bic way	acc_place=roadway	no traffic lights	#roadways=1	pavement=asphalt	no motorway	sd 150m=75-100%	
6	0.562	alcohol=no	speed_lim_type=1	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
7	0.559	male	speed_lim_type=1	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
8	0.542	sd 300m=75-100%	heavy vehicle=no	acc_place=roadway	no traffic lights	#roadways=1	pavement=asphalt	no motorway	sd 150m=75-100%	
9	0.541	heavy vehicle=no	no ped/bic way	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
10	0.531	arterial road=no	speed_lim_type=1	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
11	0.524	class I main road	sd 300m=75-100%	acc_place=roadway	no traffic lights	#roadways=1	pavement=asphalt	no motorway	sd 150m=75-100%	
12	0.523	class I main road	no ped/bic way	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
13	0.514	alcohol=no	no ped/bic way	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
14	0.508	sd 300m=75-100%	no ped/bic way	speed_lim_type=1	acc_place=roadway	no traffic lights	#roadways=1	no motorway	sd 150m=75-100%	
15	0.507	male	no ped/bic way	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
16	0.506	daylight	speed_lim_type=1	acc_place=roadway	sd 150m=75-100%	no traffic lights	#roadways=1	pavement=asphalt	no motorway	
17	0.506	sd 300m=75-100%	no ped/bic way	speed_lim_type=1	no traffic lights	#roadways=1	pavement=asphalt	no motorway	sd 150m=75-100%	
18	0.503	sd 300m=75-100%	no ped/bic way	speed_lim_type=1	acc_place=roadway	no traffic lights	pavement=asphalt	no motorway	sd 150m=75-100%	
19	0.503	sd 300m=75-100%	no ped/bic way	speed_lim_type=1	acc_place=roadway	#roadways=1	pavement=asphalt	no motorway	sd 150m=75-100%	
20	0.501	male	sd 300m=75-100%	acc_place=roadway	no traffic lights	#roadways=1	pavement=asphalt	no motorway	sd 150m=75-100%	

Table 11: Maximal frequent itemsets generated from all the accidents in cluster 7. Relative minimum support 0.5. Totally 2296 itemsets satisfied the minimum support requirement 0.5 of which 100 were maximal. The twenty largest maximal itemsets are shown.

Itemset	Sup	Items									
1	0.584	male	no ped/bic way	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway	
2	0.566	sd 300m=75-100%	no ped/bic way	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway	
3	0.553	class I main road	no ped/bic way	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway	
4	0.548	class I main road	male	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway	
5	0.548	sd 300m=75-100%	male	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway	
6	0.519	class I main road	sd 300m=75-100%	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway	
7	0.550	daylight	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway		
8	0.544	#involved=2	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway		
9	0.515	alcohol=0	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway		
10	0.510	speed limit = 80	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway		
11	0.505	arterial road=no	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway		
12	0.503	temperature ≥ +3	speed_lim_type=1	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway		
13	0.512	surface=bare,dry	sd 150m=75-100%	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway			
14	0.510	daylight	no ped/bic way	acc_place=roadway	pavement=asphalt	no traffic lights	#roadways=1	no motorway			
15	0.508	sd 300m=75-100%	male	no ped/bic way	sd 150m=75-100%	no traffic lights	#roadways=1	no motorway			
16	0.506	daylight	no ped/bic way	sd 150m=75-100%	acc_place=roadway	no traffic lights	#roadways=1	no motorway			
17	0.505	daylight	no ped/bic way	sd 150m=75-100%	pavement=asphalt	no traffic lights	#roadways=1	no motorway			
18	0.503	sd 300m=75-100%	male	no ped/bic way	sd 150m=75-100%	acc_place=roadway	no traffic lights	no motorway			
19	0.501	#involved=2	sd 300m=75-100%	sd 150m=75-100%	acc_place=roadway	no traffic lights	#roadways=1	no motorway			
20	0.501	sd 300m=75-100%	male	no ped/bic way	sd 150m=75-100%	acc_place=roadway	no traffic lights	#roadways=1			

Table 12: Maximal frequent itemsets generated from the fatal accidents in cluster 7. Relative minimum support 0.5. Totally 2548 itemsets satisfied the minimum support requirement 0.5 of which 70 were maximal. The twenty largest maximal itemsets are shown.

Rule	Sup	Conf	Lift	Rule body					Rule head
1	0.0076	0.437	14.80	#roadways=1	male	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
2	0.0076	0.434	14.69	speed_lim_type=1	male	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
3	0.0076	0.434	14.69	no traffic lights	male	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
4	0.0076	0.434	14.69	pavement=asphalt	male	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
5	0.0076	0.430	14.56	no motorway	male	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
6	0.0091	0.424	14.35	speed_lim_type=1	male	surface=bare,dry	heavy vehicle=yes	acc_class=head-on collision (5)	fatal
7	0.0077	0.417	14.12	class I main road	male	surface=bare,dry	heavy vehicle=yes	acc_class=head-on collision (5)	fatal
8	0.0094	0.415	14.06	pavement=asphalt	male	surface=bare,dry	heavy vehicle=yes	acc_class=head-on collision (5)	fatal
9	0.0088	0.415	14.05	speed_lim_type=1	male	arterial_road=yes	heavy vehicle=yes	acc_class=head-on collision (5)	fatal
10	0.0088	0.415	14.05	#roadways=1	speed_lim_type=1	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
11	0.0088	0.415	14.05	#roadways=1	no traffic lights	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
12	0.0088	0.415	14.05	pavement=asphalt	#roadways=1	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
13	0.0106	0.414	14.04	speed_lim_type=1	male	surface=bare,dry	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
14	0.0108	0.414	14.02	speed_lim_type=1	male	temperature ≥ +3	heavy vehicle=yes	acc_class=head-on collision (5)	fatal
15	0.0077	0.414	14.02	#roadways=1	class I main road	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
16	0.0088	0.412	13.97	no traffic lights	speed_lim_type=1	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
17	0.0088	0.412	13.97	pavement=asphalt	speed_lim_type=1	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
18	0.0088	0.412	13.97	pavement=asphalt	no traffic lights	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
19	0.0084	0.411	13.94	#roadways=1	sd 150m=75-100%	speed limit=100	heavy vehicle=yes	acc_type= head-on collision on straight (20)	fatal
20	0.0087	0.411	13.93	speed_lim_type=1	class I main road	surface=bare,dry	heavy vehicle=yes	acc_class=head-on collision (5)	fatal

Table 13: Constraint association rules generated from the fatal accidents in cluster 7. Only rules with fatal accident as a consequence are accepted. Relative minimum support 0.75% (that is 141 out of 18471 accidents). Minimum condfidence 0.3. Size of the rule set 2-6. Totally 2755 rules were found. Twenty best according to Lift-value are shown.