EXPLORING THE POTENTIAL OF DATA MINING TECHNIQUES FOR THE ANALYSIS OF ACCIDENT PATTERNS

Carlo G. Prato, Technical University of Denmark

Shlomo Bekhor, Technion – Israel Institute of Technology

Ayelet Galtzur, Technion – Israel Institute of Technology

David Mahalel, Technion – Israel Institute of Technology

Joseph N. Prashker, Technion – Israel Institute of Technology

ABSTRACT

Research in road safety faces major challenges: individuation of the most significant determinants of traffic accidents, recognition of the most recurrent accident patterns, and allocation of resources necessary to address the most relevant issues. This paper intends to comprehend which data mining techniques appear more suitable for the objective of providing a broad picture of the road safety situation and individuating specific problems that the allocation of resources should address first. Descriptive (i.e., K-means and Kohonen clustering) and predictive (i.e., decision trees, neural networks and association rules) data mining techniques are implemented for the analysis of traffic accidents occurred in Israel between 2001 and 2004. Results show that descriptive techniques are useful to classify the large amount of analyzed accidents, even though introduce problems with respect to the clear-cut definition of the clusters and the triviality of the description of the main accident characteristics. Results also show that prediction techniques present problems with respect to the large number of rules produced by decision trees, the interpretation of neural network results in terms of relative importance of input and intermediate neurons, and the relative importance of hundreds of association rules. Further research should investigate whether limiting the analysis to fatal accidents would simplify the task of data mining techniques in recognizing accident patterns without the "noise" probably created by considering also severe and light injury accidents.

Keywords: traffic accidents, data mining, clustering analysis, decision trees, neural networks, association rules.

INTRODUCTION

Research in road safety faces major challenges: individuation of the most significant determinants of traffic accidents, recognition of the most recurrent accident patterns, and allocation of resources necessary to address the most relevant issues. Typically, traffic accidents have been regarded as random events and statistical models have been extensively employed to investigate the determinants of fatal and injury accidents. Logistic regression, Poisson regression, Negative Binomial regression and ordered choice models have been commonly applied. However, most of these models rely on specific assumptions and pre-defined underlying relationships between dependent and independent variables. Alas, whenever these assumptions are violated, these models could lead to erroneous estimation of the likelihood of accident occurrence under the hypothesized conditions.

Data mining techniques constitute an alternative approach that has received increasing attention from researchers in recent years (e.g., Abdel-Aty and Keller, 2005; Chang, 2005; Chang and Chen, 2005; Chang and Wang, 2006; Geurts et al., 2005). Clustering analysis, decision trees and neural networks have been generally utilized to replace existing algorithms for accident classification (e.g., Abdel-Aty and Keller, 2005, Chang and Wang, 2006; Geurts et al., 2005) and crash frequency estimation (e.g., Chang, 2005; Chang and Chen, 2005). The implementation of data mining techniques avoids the limitation represented by specific assumptions and potentially uncovers not previously hypothesized underlying relationships. However, the analysis of findings from data mining applications in the transportation field generates doubts about the ability of data mining techniques to provide non-trivial insights in road safety research (e.g., Chang and Cheng, 2005; Geurts et al., 2005; Tseng et al., 2005).

With this premise, this paper investigates the implementation of descriptive (i.e., K-means and Kohonen clustering) and predictive (i.e., decision trees, neural networks and association rules) data mining techniques for the analysis of traffic accidents. The purpose of this study is to comprehend which data mining techniques appear more suitable for the objective of providing a broad picture of the road safety situation and individuating specific problems that the allocation of resources from the government should address first.

The remainder of this paper is organized as follows. A review of existing literature in data mining applications to road safety research is followed by a description of the applied methodology and the data used for the analysis. Results from the implementation of the described data mining techniques are then presented before a summary and discussion of the major findings of this research.

LITERATURE REVIEW

Data mining techniques have been commonly employed in business administration, industry, and engineering (see Witten and Frank, 2005). In recent years there has been a growing interest in applying data mining techniques in the transportation field as well, for example in

the areas of traffic engineering and travel behaviour. Smith et al. (2001) show that advanced data mining techniques hold high potential to provide automated tools that assist traffic engineers in signal control system design and operations. Yang et al. (1993), Dougherty (1995) and Yamamoto et al. (2002) apply decision trees and production rules to investigate drivers' route choice behaviour under advanced information systems, even though in current practice relatively little information has been successfully extracted from the wealth of data collected by intelligent transportation systems.

In the late 90's and the beginning of this century there have been several attempts to use data mining techniques in the area of traffic safety. In particular, frequent patterns in accident data have been searched by implementing spatial data mining (Zeitouni and Chelghoum, 2001), clustering techniques (e.g., Ljubic et al., 2002; Geurts et al., 2003; Bayam et al., 2005), rule induction (e.g., Geurts et al., 2003; Geurts et al., 2005; Kavsek et al., 2006), decision trees (e.g., Strnad et al., 1998; Clarke et al., 1998; Bayam et al., 2005) and neural networks (e.g., Mussone et al., 1999; Bayam et al., 2005). Some applications have combined data mining techniques with technological enhancements, for example Ng et al. (2002) present a combination of cluster analysis, regression analysis and Geographical Information System (GIS) platforms to group homogeneous accident data, to estimate the number of accidents and to assess the crash risk.

Following these attempts, in recent years there has been a growing body of research exploring whether data mining techniques are potentially more suitable than classical econometric models to uncover relations between the variables that affect accidents, such as road characteristics, driver characteristics and attitudes, vehicle features and seasonal factors. For example, Cameron (1997) indicate that clustering methods seem an important tool when analyzing traffic accidents as these methods are able to identify groups of road users, vehicles and road segments which would be suitable targets for countermeasures. Lee et al. (2002) present a review and discusses limitations of classical econometric models that have been widely used to analyze road crashes. Chen and Jovanis (2002) show that certain problems may arise when using classic statistical analysis on datasets with large dimensions, namely the exponential increase in the number of parameters as the number of variables increases and the invalidity of statistical tests as a consequence of sparse data in large contingency tables. Chang and Chen (2005) compare prediction performances of decision trees and negative binomial regressions to determine that decision trees are a better method for analyzing freeway accident frequencies. Chong et al. (2005) evaluate the performance of four machine learning paradigms applied to modelling the severity of injury that occurred during traffic accidents: neural networks, support vector machines, decision trees and a hybrid model involving decision trees and neural networks.

Even though data mining techniques have been shown to be potentially a powerful tool for dealing with accident classification and prediction problems, a general problem seems to be the triviality of the results. For example, Geurts et al. (2003) classify accident data obtained from the National Institute of Statistics for the region of Flanders (Belgium) with Kohonen network clustering and Geurts et al. (2005) identify accident circumstances that frequently occurred at high frequency accident locations with association rules. Only 16 association

rules out of thousands are described to conclude, for example, that accidents are influenced by rainy weather, wet road surface and slippery conditions. Chang and Chen (2005) investigate freeway accident frequency by means of the Classification and Regression Tree (CART) method and conclude that conflicts between vehicles and exposure to potential crash risk are expected to increase with increasing number of vehicles, rainy conditions, road steepness and the effect of speed differentials. Tseng et al. (2005) examine traffic accident patterns in relation with driver inattention with Kohonen networks and conclude that when inattention and physical-mental conditions take place at the same time, driver have a higher tendency of being involved in a crash where they collide with static objects. The common trait of these results is that these trivial statements do not seem to actually require the implementation of complex techniques to be either defined or predicted.

METHODOLOGY

This section illustrates the data mining techniques applied in this study, by distinguishing methods for descriptive and predictive analysis. Descriptive analysis is used to uncover groups or clusters of data objects based on similarities among these objects occurring as a result of interactions among independent variables. Predictive analysis is used to forecast future events or behaviours based on mapping a set of input values to an output value.

Descriptive analysis

K-means and Kohonen networks perform the task of segmenting a heterogeneous population into homogeneous subgroups, thus they perform descriptive analysis.

K-means clustering

The most popular non-hierarchical clustering method is the K-means technique. “K” refers to the number of clusters chosen to reduce the dimensionality of the problem, while “means” refers to the cluster being represented by the mean of observations on selected variables.

The implementation of K-means clustering in this research utilizes the “maximin” method to first select cluster centres. Initially, the algorithm positions the first cluster centre as the first record of the data file, and the remaining centres are created by searching for positions in an n-dimensional space that are as far as possible from any other cluster centre already generated. Then, the algorithm calculates the Euclidean distance between each record and every cluster centre and assigns each record to the cluster with the smallest squared Euclidean distance. After the assignment of all the cases is completed, the location of each cluster centre is recomputed as the average of all the cases within the cluster and this iterative process is repeated until one stopping criteria is reached, namely a change in means below a defined threshold or a maximum number of iterations.

The typology of the input variables for K-means clustering is not an issue, as long as the calculation of the Euclidean distance follows the standardization of the data. Fields of range

type are transformed into a scale that varies between 0 and 1, flag fields are coded such as that the false value equals 0 and the true value equals 1, and categorical fields are recoded as flag variables for each category. The application of K-means clustering to this research proposes exploring results from various numbers of groups and fixing interruption criteria for the algorithm in 1×10^{-6} for the variation of mean change or in 20 iterations maximum.

Kohonen networks

Kohonen networks are a type of neural network based upon the idea of self-organized learning. Since the algorithm does not attempt to predict values of target variables, these networks are suitable for clustering. The clusters are formed from patterns that share similar features. The network consists of a one or two-dimensional grid of neurons. Each neuron is connected to each of the inputs, and weights are considered for each connection, as illustrated in figure 1.

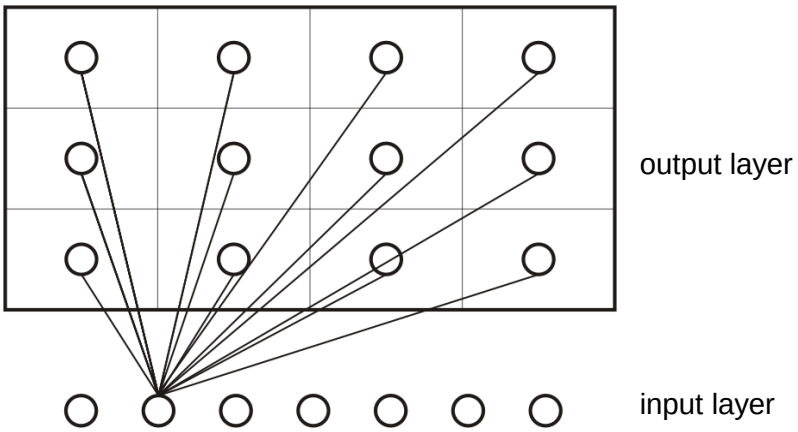

Figure 1 - Illustration of a Kohonen network

The network trains by presenting cases to the grid and comparing characteristics of each record with those of all neurons in the grid, after giving random weights initially. The neuron in the grid with the most similar pattern “gets” the examined record, and its weight is adjusted to be more similar to that of the record just acquired to enhance the likelihood of similar records to be captured by the same node. The network adjusts the weights of the surrounding neurons in the grid as well, and after the data pass through the network a number of times, the result consists of a map containing clusters of records corresponding to different types of patterns in the data. Similar patterns (i.e., similar neurons in the grid) should be closer in the map than dissimilar ones.

The algorithm works in two phases, namely a first stage with large-scale changes and a second stage with smaller changes in the weights in order to perform a fine-tuning of the map. Also, the algorithm requires the number of neighbours around the acquired node whose weight is modified at each phase and defines a learning parameter eta at each phase. As the training process involves long iterations with weight adjustments, the Kohonen network takes longer to train than K-means clustering. The implementation of Kohonen networks in this study considers exploring clusters from different dimensions of the map and defining the

parameters for both phases: the first phase has learning parameter equal to 0.3, neighbours equal to 2 and 20 cycles, the second phase has learning parameter equal to 0.1, neighbours equal to 1 and 100 cycles, and convergence is reached when the learning parameter arrives to a null value.

Predictive analysis

Decision trees, association rules and neural networks examine the data and estimate the outcome values of a dependent variable, thus they perform predictive analysis.

Decision trees

Decision trees are mining methods able to forecast or classify future observations according to decision rules. Information is divided in classes and is utilized to generate rules able to classify previous and new cases with absolute precision. The interpretation of trees is simpler than other techniques since the configuration of the branches exemplifies the structure. Further, decision trees are easily converted into "if-then" rules to enhance the comprehension of the model and of the relationships between variables.

The algorithm C5.0 divides the records according to the field that yields the maximum information gain, defined as the difference between the average information needed to identify the class of a record within the entire data and the expected information required once the data has been partitioned into each outcome of the field being tested. Each subgroup defined at the first division is further examined until additional division of the subgroups is not feasible, and the lower level subdivisions are examined to remove or cut those not giving significant contribution to the model. This process is known as pruning of the tree, which is used to decide whether a branch should be "simplified" back toward the parent node on the basis of the comparison between the predicted errors for the unpruned branches and those for the pruned node.

The CHAID algorithm, acronym of Chi-squared Automatic Interaction Detection, utilizes chi-squared statistics to identify optimal subdivisions of the dataset. Initially, the CHAID algorithm analyzes the contingency tables of each independent variable and verifies their significance by means of a chi-squared independency test. Then, the algorithm selects the most significant predictor and merges categories of the variable that are yielding similar results, while proceeding with the division in subgroups of the data according to the new categories created. Categories are merged when their differences are equal to the difference obtained with the independency test. While the C5.0 algorithm is stable in presence of missing data and large number of input fields, does not require long training time and the interpretation of the tree is easier, the CHAID algorithm is efficient in presence of missing data and generates trees for categorical predictors with more than one branch for each subgroup.

Note that when the algorithms produce a decision tree, each leaf describes a certain subgroup of the training data and each record enters only one single leaf, but when the

algorithm produces induction rules, these present a simplified version of the information contained in the decision tree and each record applies to more than one rule as well as any rule at all. The implementation of the C5.0 algorithm in this study uses a 75% pruning and a minimum of 20 records per child branch to reduce the effect of the noise in the data. Further, the algorithm is instructed to cut the tree in two phases: the first executes a local cut that examines the sub-trees and compresses the branches that increase model precision, while the second explores the whole tree and compresses weak sub-trees. The implementation of the CHAID algorithm in this study introduces the level of significance for merging the categories, equal to 0.05, and the convergence criterion, with a maximum number of 100 iterations if the optimal value of the chi-squared test is inferior to 0.001. For predicting purposes, both algorithms use part of the data for training and part of the data for testing. The comparison between actual and predicted values provides a measure of goodness-of-fit of the estimated models.

Neural networks

A neural network consists of a number of neurons that are arranged in layers and are linked to every neuron in the previous layer by connections with different weights. The learning adapts the weights iteratively and provides the system of a method to learn by example.

The Multi-Layer Perceptron (MLP) is a simplified model of the human mind elaboration process, and works by simulating an elevated number of simple elaboration units that resemble abstract versions of neurons. As illustrated in figure 2, an input layer represents the input fields, an output corresponds to the output fields and one or more hidden layers represent the propagation from each neuron to each other neuron in the following layer.

Initially, the network assigns random weights and the initial answers appear without sense in the beginning of the learning phase. In the following runs, the network encounters examples with known output, the provided answers are compared to these output, and the weights are updated in order to have the closest possible level of similarity between predicted and observed values. The replication of the results increases during the learning phase and the network can be applied to future cases with unavailable results. This process reiterates and the network improves its forecasts until one or more interruption criteria are satisfied.

Several parameters determine the development of the learning phase. The alpha parameter refers to the momentum used in updating the weights when trying to locate the global solution and tends to move the weight changes in a constant direction to reduce the training time. The eta parameter refers to the learning rate and determines how much adjustment is feasible at each update and decreases according to a predetermined number of decay cycles. The persistence parameter defines the number of cycles for which the network trains without improvement to reach the stopping point. The implementation of MLP in this study uses an Exhaustive Prune training method where alpha and eta are respectively equal to 0.9 and 0.3, the number of decay cycles is equal to 100, the rate of eliminated neurons is equal to 0.02 and the persistence is equal to 100 cycles. Overtraining is avoided by considering randomly 50% of the dataset for training and the other 50% for test before the validation. As

for decision trees, the comparison between actual and predicted values provides a measure of goodness-of-fit of the constructed neural networks.

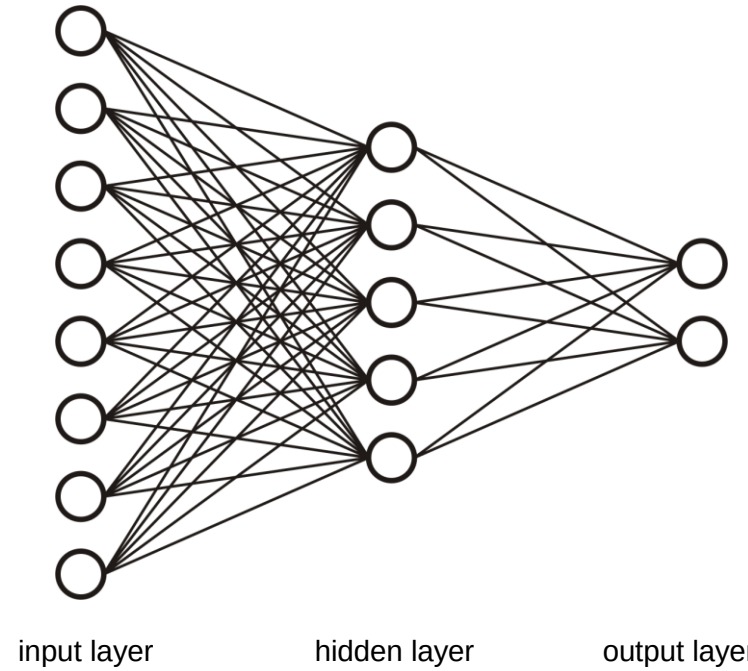

Figure 2 - Illustration of a MLP neural network

Association rules

Association rule discovery, generalized rule induction, affinity analysis and market analysis are terms that describe a type of pattern algorithm that differentiates itself from decision trees. These methods generate rules that are independent of other rules and are not restricted to a single output or dependent field, while revealing which values of two or more fields occur together. Unfortunately, the search space for independent rules grows exponentially with the number of attributes, thus association rule algorithms are very expensive from the computational perspective.

Given that an association rule consists of some conditions, also named antecedents, that are followed by some conclusions, the evaluation of the rules necessitates of two criteria: the support is the percentage of records in the dataset for which the conditions hold, and the confidence is the proportion of records meeting the conditions that also meet the conclusion. The support indicates the generality of the rules, while the confidence points out how likely is the conclusion, given that the conditions are met.

The a-priori rule discovery algorithm works only with symbolic data that are coded as flag fields and indexes, and passes through the complete dataset to generate the association rules. The support is generally under 10% to generate more potential rules, and the process is usually reiterated by using initially only a portion of data in order to evaluate support and confidence optimal for each case study. The confidence is set to 80% or 90% to avoid the generation of too many rules in the final runs of the model. The Generalized Rule Induction

(GRI) algorithm applies to a broader range of data and applies a different measure to determine the interest in a particular rule. The method generates associations based on the information content of a rule, which is assessed with a J measure that trades off support and confidence. GRI accounts for 0% support and minimum confidence equal to 50%. Association rules are complex to interpret, as exemplified in the literature and illustrated in the implementation of these algorithms in the present research.

DATA

The analyzed dataset, provided by Israel Central Bureau of Statistics (CBS), consists of traffic accidents that occurred between 2001 and 2004 and resulted in the injury of at least one person involved in the crash. For each year, databases consist of lists of records corresponding to each single crash and three different files register information about every accident with injury: the accident file, the vehicle and driver file, and the injured person file. Overall, detailed information covers every aspect of the accident including the accident location, infrastructure characteristics, vehicle features, information about the involved people, weather conditions, and traffic light situation.

Information in the accident file includes time and date of the accident, and its whereabouts in terms of police district and location related characteristics (i.e., urban or rural road, intersection or road section). The accident is classified according to three levels of severity (i.e., fatal, severe and light), type, modality and cause. Further information details the characteristics of the infrastructure in terms of allowed speed, presence and condition of median barrier, traffic control and signal system, condition of the surface related also to the weather at the moment the accident occurred. Last, information involving pedestrian and collided objects is provided for specific typologies of accidents. The vehicle and driver file includes records of each vehicle and driver involved in the accident. Each record corresponds to one vehicle and its driver, and lists generic vehicle information (e.g., type, age, motor, weight and direction of travel of the vehicle), as well as drivers' socio-economic characteristics (e.g., gender, age, year of licensure, and past offences of the driver). The injured person file comprises records of each person injured including pedestrians. Each record registers generic information about an injured person (e.g., gender, age, nationality, place of residence and type of injury sustained).

While assessing the data quality, several problems were encountered. First, the CBS coding system changed over the years. The consistency of the coding system is important for analytical purposes, since the lack of common definitions for variables such as accident type or median barrier identification could lead to a bias in the application of data mining methods. With this problem in mind, and with the contemporary objective of analyzing tens of thousands of records, the analysis focuses on the period between 2001 and 2004 for which the coding system is fairly consistent and allows avoiding potential problems related to large heterogeneity in the definition of the variables. Second, a problem consists in merging the three data files. Note that the accident file contains one accident for each record, but the other two files contain multiple records corresponding to the same accident, as long as many drivers, vehicles and injured persons were involved. A unique file is composed by

considering as a base file the accident file, and defining as vehicle 1, vehicle 2 and so on the vehicles. The same applies to the drivers and to the injured persons. At the end, a unique database lists as many records as accidents, each with its number of vehicles, drivers and injured persons. Note that the limitations to the number of columns in the SPSS program, used according to the format provided by the Central Bureau of Statistics, advises to eliminate records with more than eight vehicles and drivers involved. Last, the files contain records with missing values, which are not analyzed to assure the quality of the data. The data source for accident analysis consists of 72,056 records.

The variables considered for descriptive and predictive analysis are illustrated in table 2. Among these variables, two fields are tested as valuable categorical dependent variables: accident location in order to individuate whether specific conditions are identified for specific road type, and accident severity in order to comprehend whether outcomes are dependent on specific characteristics of the road, the vehicle or the driver. For predictive purposes the dataset is divided into a training set, containing the accidents that occurred between 2001 and 2003, and a test set, containing the crashes that happened in 2004. As neural networks randomly divide the input dataset into training and test sets, the database accounting for the collisions taken place in 2004 is considered as validation set.

Table 1 – Categorical variables for descriptive and predictive analysis

Variables	Categories
accident severity	fatal accident – severe injury – light injury
type of accident	pedestrian – front/side crash – front/rear crash – side/side crash – front/front crash – collision with stopped car – collision with parked car – collision with object – rolling/slipping – fire – other crashes
accident modality	entrance of intersection – exit of intersection – parking or gas station – slope – curve – bridge or tunnel – railway crossing – straight road or junction – other
cause of the accident	offense of the driver – pedestrian action – passenger behavior – cyclist behavior – car malfunctioning – other
location of the accident	urban intersection – urban section – interurban intersection – interurban section
allowed speed	50 km/h – 60 km/h – 70 km/h – 80 km/h – 90 km/h – 100 km/h
day / night	day – night
day of the week	Sunday – Monday – Tuesday – Wednesday – Thursday – Friday – Saturday
season of the accident	spring – summer – autumn – winter
weather conditions	clear – rainy – hot – foggy – other
number of ways on the road	one way – two ways with separation line – two ways without separation line – other

median on the road	painted line – safety rail – no safety rail – non built separation – other
shoulders of the road	good condition – bad condition – rough road – bad condition and rough road
width of the road	up to 5 m. – 5 to 7 m. – 7 to 10 m. – 10 to 14 m. – over 14 m.
regulation of intersection	no control – functioning traffic light – malfunctioning traffic light – blinking yellow – stop sign – right of way sign – other
illumination on the road	normal daylight – limited visibility because of the weather – night with lighting – night without lighting – malfunctioning lighting – unknown night conditions
surface conditions of the road	dry – wet from water – wet from slippery material – covered with mud – covered with sand – other
location of crossing pedestrians	crossing on crosswalks with traffic light – crossing on crosswalks without traffic light – crossing out of crosswalks next to an intersection – crossing out of crosswalks far from an intersection – not specified crossing position
location of standing pedestrians	pedestrian standing on the road – pedestrian standing on the median – pedestrian standing on the sidewalk or shoulders – pedestrian playing on the road – pedestrian in the traffic direction – pedestrian against the traffic direction
type of collision with objects	with street signal – with safety rail – with building – with bridge – with light or phone pole – with tree – with other object
distance of colliding objects	up to 1 m. – up to 3 m. – object on the road – object on the median – unknown position of the object
vehicles involved	one vehicle – two vehicles – three vehicles – four or more vehicles
speed offences	at least one driver with previous speed violations – no driver with previous speed violations
alcohol or drugs offences	at least one driver with previous alcohol or drug violations – no driver with previous alcohol or drug violations
private vehicles	no private vehicle involved – one private vehicle involved – two private vehicles involved – three private vehicles involved – four or more private vehicles involved
public vehicles	no public vehicle involved – one public vehicle involved – two public vehicles involved – three or more public vehicles involved
light commercial vehicles	no light commercial vehicle involved – one light commercial vehicle involved – two light commercial vehicles involved – three or more light commercial vehicles involved
heavy commercial vehicles	no heavy commercial vehicle involved – one heavy commercial vehicle involved – two heavy commercial vehicles involved – three or more heavy commercial vehicles involved
motorcycles	no motorcycle involved – one motorcycle involved – two motorcycles involved – three or more motorcycles involved
bicycles	no bicycle involved – one bicycle involved – two bicycles involved – three or more bicycles involved

ANALYSIS OF RESULTS

Descriptive analysis

The definition of the considered number of clusters constitutes a compromise between a small number, which would give problems in terms of excessive dimension of the clusters, and a large number, which would cause difficulties in terms of their semantic interpretation. Further, different data mining techniques are applied with the same number of clusters in order to evaluate the dependency of the implementation of different methods on this number.

K-means clustering is applied by testing solutions with 5, 6 and 7 clusters. Kohonen networks are constructed by experimenting linear maps with 5 and 6 clusters, as well as bi-dimensional maps that unfortunately do not converge. The semantic interpretation of the clusters constitutes a difficult task, and relates to the frequency of each category of the input variables in the records belonging to each cluster.

The first solution with 5 clusters obtained with the K-means algorithm classifies:

1. front to side accidents that occurred in urban intersections, where the median was not constituted by a safety rail and the allowed speed was 50 km/h (23,510 cases);

2. accidents that happened in autumn or winter in rainy conditions and consequently wet road surface, mainly in road sections rather than in road intersections (6,806 cases);

3. accident that took place outside urban areas, in large two way roads where the allowed speed was over 80 km/h and there was a significant percentage of collisions with the safety rail (15,441 cases);

4. accidents that involved pedestrians, with the majority of them crossing on crosswalks without traffic lights or not crossing close to intersections, in two way roads without a safety rail as a median (8,515 cases);

5. accidents that occurred in urban areas, in road sections without a white line to separate the two ways, where the majority of the collisions were front to side or front to back crashes (17,784 cases).

The solution with 6 clusters divides the first cluster into accidents that happened during either day or night. The solution with 7 clusters maintains the groups obtained with 6 clusters and splits the third cluster again according to crashes occurring either during day or night.

The first solution from the Kohonen network with 1×5 map defines different clusters:

1. front to side accidents that happened at night, mostly in urban areas and during autumn and winter with rainy conditions (19,650 cases);

2. accidents that took place in road sections at night, on either two-way roads where the allowed speed was 50 km/h or two-way roads where the allowed speed was over 80 km/h (6,364 cases);

3. accidents that occurred in road sections at day, mostly on two-way roads without white line to separate them, and with most of the accidents involving pedestrians (17,573 cases);

4. front to side accidents that took place at day, mostly in urban areas and during spring and summer (9,946 cases);

5. front to side accidents that happened at day, mostly in intersection and preferably inside urban areas, where there was not safety rail as median (18,523 cases).

The division of larger clusters into smaller clusters, observed for K-means algorithm when increasing the number of groups, is not verified with the Kohonen networks. The Kohonen network with 1×6 map actually defines different clusters with respect to the 1×5 map:

1. front to side accidents that happened during day in urban intersections, in either one-way or two-way roads without white line, and where the median was not a safety rail (15,822 cases);

2. accidents that occurred during day in urban areas, when the traffic lights were not working or there was a right of way or stop sign, and mostly with pedestrians involved (6,884 cases);

3. accidents that took place during day in urban areas and not in intersections, mostly involving pedestrians that crossed large roads without safety rail (11,873 cases);

4. accidents that happened during day outside urban areas, mostly in roads where the allowed speed was at least 90 km/h and the median was a safety rail (12,249 cases);

5. accidents that occurred at night outside urban areas, mostly during autumn and winter on large roads where the allowed speed reached 80 km/h and the median was a safety rail (7,243 cases);

6. front to side accidents that took place in urban areas during the night, mostly in autumn and winter on narrow roads without safety rail and often even without a white line to separate the two ways (17,985 cases).

The solution from the Kohonen network with a 3×3 map is difficult to explain, especially since at least four clusters present similar characteristics without a clear cut distinction among one another. The remaining five clusters are more similar to the solution from the linear 1x5 map rather than to the solution from the linear 1x6 map.

Notably, the results from the Kohonen clusters provide evidence of the importance of the vicinity property, as both solutions consider night and day accidents in adjacent groups. The

importance of this property is also proven by the fact that non-linear maps do not converge, and by the fact that the solutions of this clustering technique appear to be not strictly interrelated, especially in the 3×3 map where the least significant clusters are positioned in the middle of the map and separate the remaining resulting groups. In addition, given the differences in the nature of the algorithms, it is not surprising that the two clustering techniques work in two different directions: K-means creates clusters based on the location and the typology of the accident, while Kohonen networks generate clusters based on the day or night attribute and the accident location.

Predictive analysis

Decision trees

Decision trees produce rules that help classifying the accidents according to the chosen dependent variables, and the most interesting results are described in the following sections.

Considering the location of the accident as the dependent variable, the decision trees generated with the C5.0 and CHAID algorithms are similar. For the C5.0 algorithm, the most significant variables consist of the allowed speed followed by the typology of the accident, the existence of previous speed violations for at least one of the involved drivers and the number of vehicles implicated in the crash. For the CHAID algorithm, the most relevant fields consist of the regulation of the intersections, followed by the allowed speed, the number and the type of vehicles involved and the physical condition of the median.

The confusion matrices in tables 2 and 3 provide insight into the predictive ability of the two algorithms, which is 67.1% for the C5.0 and 87.8% for the CHAID algorithm. The rows represent the observations and the columns the predicted values, thus the elements in the diagonal indicate the correct predictions for accidents occurred in 2004 on the basis of trees generated from data about accidents occurred between 2001 and 2003. The C5.0 algorithm predicts well crashes in rural areas, but confuses collisions inside urban areas as the majority of incorrect forecasts concerns urban accidents that occurred in road junctions and are instead predicted to take place in a section, and vice versa. The CHAID algorithm does not present the same problem, although it predicts an excess of accidents in urban areas.

Table 2 - Confusion matrix for accident location with C5.0 algorithm

	Interurban intersection	Interurban section	Urban intersection	Urban section
Interurban intersection	1000	458	460	192
Interurban section	371	1758	143	341
Urban intersection	44	29	4966	1805
Urban section	47	139	1812	4172

Table 3 - Confusion matrix for accident location with CHAID algorithm

	Interurban intersection	Interurban section	Urban intersection	Urban section
Interurban intersection	1270	133	638	69
Interurban section	0	2037	0	576
Urban intersection	33	4	6242	565
Urban section	0	147	0	6023

Decision rules with confidence level over 75% are presented in tables 4 and 5. These rules suggest that accidents in interurban intersections occur mainly in conditions of limited visibility, for example at night with only the natural night light available, and when the median is not physical but is only a white line drawn on the road. These rules leads to think about problems related to excessive speed, a concept that is confirmed by the rules regarding crashes in interurban sections, for example with single-vehicle accidents where a car digress from the road. In urban areas, pedestrians, cyclists and motorcyclists are often involved, and sometimes supposedly their behaviour causes crashes that have severe consequences.

Table 4 - Rules for accident location with C5.0 algorithm

IF the speed allowed is 80 km/h, AND the accident occurs with more than one vehicle involved, AND the accident is a front to side collision, AND the only light in the location of the accident is the natural night light THEN the likelihood of the accident occurring in an interurban intersection is 75.3%
IF the speed allowed is 90 km/h, AND the accident occurs with more than one vehicle involved, AND the accident is a front to side collision, AND the only light in the location of the accident is the natural night light THEN the likelihood of the accident occurring in an interurban intersection is 75.0%
IF the speed allowed is 90 km/h, AND the accident occurs with more than one vehicle involved, AND the accident is a front to back collision, AND the accident results in fatalities THEN the likelihood of the accident occurring in an interurban section is 92.2%
IF the speed allowed is 80 km/h, AND the accident occurs with only one vehicle involved, AND the accident is the getting off the road of a car, THEN the likelihood of the accident occurring in an interurban section is 89.1%
IF the speed allowed is 90 km/h, AND the accident occurs with more than one vehicle involved, AND the accident is a front to back collision, AND the accident results in severe injuries THEN the likelihood of the accident occurring in an interurban section is 85.2%
IF the speed allowed is 80 km/h, AND the accident occurs with only one vehicle involved, AND the accident is a front to side collision,

THEN the likelihood of the accident occurring in an interurban section is 85.1%
IF the speed allowed is 50 km/h, AND the accident is a front to side collision, AND the road signs in the location of the accident are in poor conditions THEN the likelihood of the accident occurring in an urban intersection is 75.6%
IF the speed allowed is 50 km/h, AND the accident occurs with at least one pedestrian involved, AND the accident occurs with a parking vehicle THEN the likelihood of the accident occurring in an urban section is 92.0%
IF the speed allowed is 50 km/h, AND the accident occurs with at least one pedestrian involved THEN the likelihood of the accident occurring in an urban section is 83.7%

Table 5 - Rules for accident location with CHAID algorithm

IF the regulation of the intersection is a not working traffic light, AND the allowed speed is over 90 km/h THEN the likelihood of the accident occurring in an interurban intersection is 96.9%
IF the regulation of the intersection is a blinking yellow traffic light, a “stop” or a “right of way” sign, AND the allowed speed is between 80 and 90 km/h THEN the likelihood of the accident occurring in an interurban intersection is 96.2%
IF the regulation of the intersection is a not working traffic light, AND the allowed speed is 90 km/h, AND the median is constituted by a safety rail THEN the likelihood of the accident occurring in an interurban section is 95.6%
IF the regulation of the intersection is a not working traffic light, AND the allowed speed is over 80 km/h, AND accident is either a collision with an object or the rolling of a car THEN the likelihood of the accident occurring in an interurban section is 85.7%
IF the regulation of the intersection is a blinking yellow traffic light, a “stop” or a “right of way” sign, AND the allowed speed is 50 km/h, AND the accident occurs with at least one motorcycle involved THEN the likelihood of the accident occurring in an urban intersection is 99.3%
IF the regulation of the intersection is a not working traffic light, AND the allowed speed is 50 km/h, AND the accident type is a front to side collision THEN the likelihood of the accident occurring in an urban intersection is 96.1%
IF the regulation of the intersection is a not working traffic light, AND the allowed speed is 50 km/h, AND the accident occurs with at least a pedestrian involved, AND the two-way road does not present a white line in the location of the accident, AND the accident results in fatalities or severe injuries THEN the likelihood of the accident occurring in an urban section is 95.6%
IF the regulation of the intersection is a not working traffic light, AND the allowed speed is 50 km/h, AND the accident is a front to side collision, AND the two-way road either presents or not a white line in the location of the accident, AND the accident occurs with at least one motorcycle involved THEN the likelihood of the accident occurring in an urban section is 88.6%

Considering the severity of the accident as dependent variable, the decision trees resulting from the implementation of the C5.0 and CHAID algorithms are extremely different. From the application of the C5.0 algorithm, the nodes at the top of the tree are the number of vehicles implicated, the presence of a bicycle or a commercial vehicle in the collision, the existence of previous speed violations for at least one of the drivers and the involvement of a pedestrian. From the application of the CHAID algorithm, the nodes at the top of the tree are the type of accident, the location of the crash and the presence of private vehicles.

The predictive abilities of the two algorithms are comparable, with 87.2% and 86.6% correctly predicted for the C5.0 and CHAID respectively, but only the C5.0 algorithm is able to predict fatal accidents. Accordingly, the C5.0 algorithm performs significantly better than the CHAID algorithm. The rules for fatal and severe accidents are summarized in tables 6 and 7 for both algorithms, and show clearly the better performance of the C5.0 algorithm, as no rule for fatal accidents and only a single rule for crashes resulting in severe injuries is produced by the CHAID method with low confidence level.

Table 6 - Rules for accident severity with C5.0 algorithm

IF the accident occurs with only one vehicle involved, AND the allowed speed is 90 km/h, AND the accident occurs with at least one pedestrian involved THEN the likelihood of the accident being fatal is 77.5%
IF the accident occurs with only one vehicle involved, AND the allowed speed is 100 km/h, AND the accident occurs with at least one pedestrian involved and crossing the road THEN the likelihood of the accident being fatal is 75.0%
IF the accident occurs with only one vehicle involved, AND the allowed speed is 70 km/h, AND the width of the road where the accident occurs is more than 10.5 m., AND the accident occurs with at least one pedestrian involved and crossing the road, THEN the likelihood of the accident resulting in a severe injury is 70.8%

Table 7 - Rules for accident severity with CHAID algorithm

IF the accident happens in the spring, AND the allowed speed is over 60 km/h AND the accident occurs with at least one pedestrian involved and crossing the road THEN the likelihood of the accident resulting in a severe injury is 41.2%

Fatal accidents mainly involve pedestrians that cross wide roads where the allowed speed is high, most likely highways or major arterials. This phenomenon is frequently seen in Israel, especially when driving on major arterials that cross villages or cities. These roads form a barrier dividing the community into separate parts, although such division is inconsistent with the land use pattern. Further, the involvement of bicycles and commercial vehicles increases the likelihood of accidents to result in fatalities, especially during night and in seasons like autumn and winter. Last, rules for fatal and severe accidents point out that single-vehicle crashes produce the most severe outcomes in terms of injuries.

Neural networks

Neural networks generate connections between input variables and output predictors by processing the same database used for the construction of the decision trees. Neural networks might produce results similar to the decision trees in terms of predictive ability, but their cost in terms of computational time is higher, as the convergence of a network takes around two hours while the convergence of a decision tree takes less than two minutes. Further, the interpretation of a neural network is far more complicated than the interpretation of a decision tree, as figures 3 and 4 exemplify.

Figure 3 illustrates the circular representation of a neural network, in which input and output variables are collocated in a circle and connections from the input to the output variables are represented with various degree of thickness directly proportional to their strength. Figure 4 presents the reticular representation of a neural network, in which input variables are collocated in the design space according to their distance from the output variables and connections are thicker when relationships are stronger. This problem of the interpretation of the results in neural network implementation, as opposed to the relative easiness of interpretation of the rules generated with decision trees, gives the edge to the latter technique for prediction purposes.

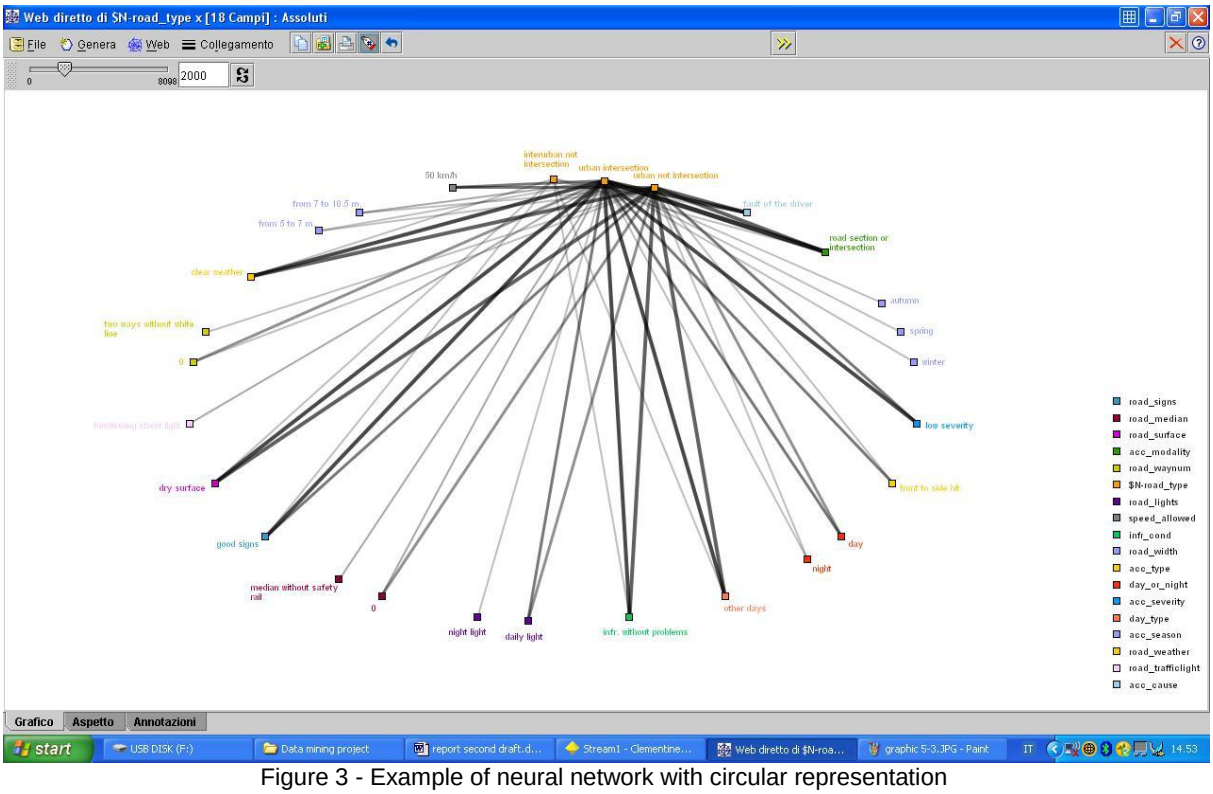

Figure 3 - Example of neural network with circular representation

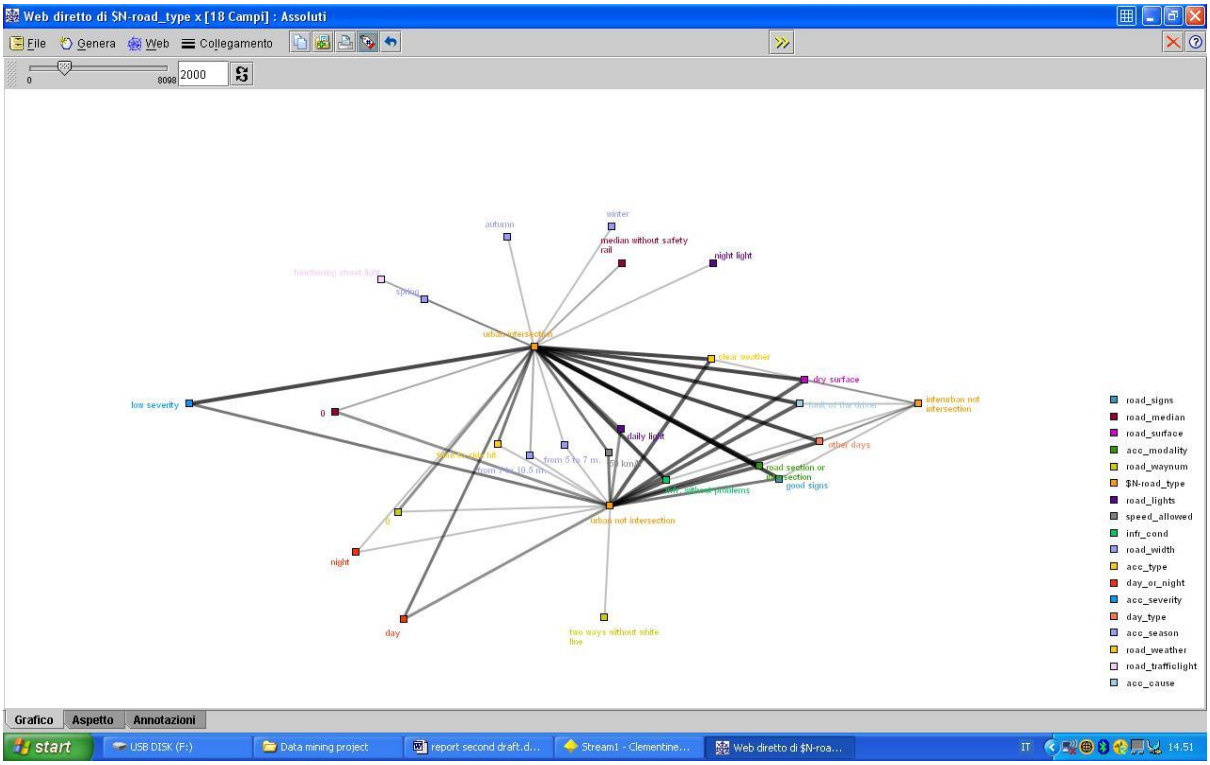

Figure 4 - Example of neural network with reticular representation

When analyzing the location of the accident, two input variables are by far more significant than others to explain where the accident occurred: the type of control of the intersections and the allowed speed. Table 8 details the most relevant input variables in predicting the location of the accidents.

Table 8 - Relevant input variables for accident location with MLP network

Variables	Relative importance
regulation of intersection	0.3889
allowed speed	0.3458
number of ways on the road	0.0404
type of accident	0.0287
other variables	< 0.0200

Results are similar to the ones obtained with the CHAID decision tree algorithm. The estimated precision in training is equal to 81.1%, and the prediction accuracy reaches 81.4% when the forecasted results are compared to the observed locations for crashes taking place in the year 2004. Interestingly, the confusion matrix in table 9 suggests that the high predictive ability does not reflect the goodness of the model, as any accident is predicted to be located in an intersection outside an urban area. The first column emphasizes that the model does not predict any accident to happen in an intersection outside an urban area, rather all the crashes are forecasted to take place in urban intersections. The remaining part of the matrix shows greater predictive accuracy.

Table 9 – Confusion matrix for accident location with MLP network

	Interurban intersection	Interurban section	Urban intersection	Urban section
Interurban intersection	0	135	1909	66
Interurban section	0	2182	0	431
Urban intersection	0	6	6280	558
Urban section	0	194	1	5975

When analyzing the severity of the accident, the MLP network estimates that the outcome of the accident is explained mainly by the type of accident and the involvement of bicycles, motorcycles, commercial vehicles. The network constructs two hidden layers and reaches a precision of 86.6% during the training and test phase, exactly identical to the prediction accuracy when comparing the predicted values with the actual outcomes of crashes occurred in the year 2004. Table 10 details the most relevant variables to classify whether accidents resulted in fatalities or severe or light injuries.

Table 10 - Relevant input variables for accident severity with MLP network

Variables	Relative importance
accident type	0.1478
bicycles	0.1340
motorcycles	0.1122
private vehicles	0.0731
heavy commercial vehicles	0.0589
allowed speed	0.0537
location of the accident	0.0474
regulation of intersection	0.0436
condition of infrastructure	0.0433
speed offences	0.0407
other variables	< 0.0400

The high predictive ability does not actually indicate that the model is good, since the model predicts almost all accidents as resulting in light injuries. This problem, encountered also with the CHAID algorithm but to a lesser extent, demonstrates that the evaluation of the goodness of the model is not only related to the overall prediction accuracy, but also to the actual interpretability of the model. Again, the significant higher computational cost does not

produce any advantage in terms of model goodness of fit, providing more insight into the evaluation of the decision trees as better methods than neural networks. Notably, the cause of inaccuracy is somewhat different with respect to the previous case. The prediction accuracy of accident severity is influenced by the presence of a dominant category, namely crashes that resulted in light injuries. The same does not apply to the accident location, where none of the categories is dominant, and the problem is related to the model rather than to the data.

Association rules

Association rules are applied to analyze characteristics of the accidents occurred in black spots, which are defined as specific locations in the road network where the frequency of fatalities results higher than the expected average. The list of the black spots defined for the national road network by the Israeli Road Directorate is matched to the accident data in order to create an additional binary variable, where 1 indicates that the accident occurred in a black spot and 0 otherwise. This variable is then assumed to be the categorical predictor for the analysis, but given the definition of the black spot data and the strict correlation with the severity of the accidents, not surprisingly the results are trivial as fatal accidents result to happen in black spots.

Given the trivial results obtained, the database is divided into two parts: accidents occurred in black spots and accidents taken place in other network sections. The severity of the accidents is considered as categorical dependent variable and the aforementioned association rule algorithm with a confidence rule equal to 90% and a support equal to 5% is tested. The a-priori algorithm produces around 5000 rules for both parts of the database, but all these rules have lift very close to one and thus the selection of rules according to this criterion is not possible. This means that none of the rules is clearly noticeable among the thousands generated, and that any attempt to synthesize the results would be based purely on subjective rather than objective criteria. Considering that the literature exhibits the very same problem of rule selection (see Geurts et al., 2005), and that existing studies do not offer a remedy for rule selection, association rules are not further investigated.

SUMMARY AND CONCLUSIONS

This study focuses on the search for the most promising data mining techniques for accident analysis, and contemporarily provides some classification and predictive rules to understand accident occurrence. Accordingly, the presentation of the conclusions and the proposition of further research are developed from both perspectives of data mining techniques and safety recommendations.

This study applies several data mining techniques and focuses on the understanding of both descriptive and predictive methods. The initial selection of eligible techniques for this study is based on literature survey findings, which proposes investigating implementation of some among the large number of alternative techniques available. Namely, the descriptive

techniques that are suitable for categorical data are K-means clustering and Kohonen networks, while the predictive methods that are suitable for the same type of data are CHAID and C5.0 algorithms for the construction of decision trees, the MLP neural networks, and the Association Rules algorithms. Results show that descriptive techniques are useful to classify the large amount of analyzed accidents, even though both K-means clustering and Kohonen networks present issues with the clear-cut definition of the clusters and with the triviality of the description of the main characteristics of the accidents, thus suggesting that predictive techniques might be more suitable. Results show that also predictive techniques pose challenges. Decision trees produce large numbers of rules, but are better than neural networks in terms of their definition of the most relevant variables for the outcome of an accident. Neural networks generate results whose interpretation is difficult to grasp. Association rules produce thousands of rules, whose relative importance is impossible to define.

From the analysis of the clusters and the rules, it appears evident that there are safety issues with respect to accidents involving pedestrians, cyclists and motorcyclists. For pedestrians in particular, two problems emerge: the first concerns pedestrians crossing illegally major arterials that divide communities, perhaps due to the lack of safe passages. The second problem concerns drivers' limited visibility in bad weather conditions.

Accident severity seems to be influenced mainly by the typology of vehicles involved. Obviously, the involvement of vulnerable road users and heavy commercial vehicles causes the outcome results to be more severe than when only private vehicles are engaged. Also relevant for the severity of the accident are the road infrastructure conditions, the illumination on the infrastructure, and the regulation of the traffic. In particular, the number of front to side accidents can be reduced by authorities assuring regular and emergency traffic signals maintenance, as non-working traffic lights appear related to high likelihood of accident occurrence.

Accident location appears to be affected by the existence of previous speed violations by at least one of the implicated drivers, the number of vehicles involved and the typology of the accident. Typically, accidents in interurban intersections take place mainly in conditions of limited visibility and when the median is not constructed, that leads to think about problems of speed violations. This concept is confirmed by the rules regarding crashes in interurban sections, for example with single-vehicle accidents where cars get off the road. In urban areas, vulnerable road users are often involved and supposedly sometimes their behaviour causes crashes with severe consequences.

From the data mining perspective, further research should investigate whether limiting the analysis to fatal accidents would simplify the task of data mining techniques in recognizing accident patterns without the "noise" probably created by considering also severe and light injury accidents and without the rather trivial accident patterns likely related to this "noise". From the safety recommendation perspective, applying the same techniques while considering different partition of the same dataset (e.g., by accident location, by accident

type) would allow removing the trivial results and obtaining stable accident patterns that provide a trustworthy representation of the safety situation in the country.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of the Ran Naor Foundation for the Advancement of Safety Research, and the helpful comments of Shalom Hakkert, Tsippy Lotan and Yechiel Millman. This study was conducted when the first author was affiliated with the Transportation Research Institute at the Technion – Israel Institute of Technology.

REFERENCES

Abdel-Aty, M. and J. Keller (2005). Exploring the overall and specific crash severity levels at signalized intersections. Accident Analysis and Prevention, 37 (3), 417-425.

Bayam, E., J. Liebowitz and W. Agresti (2005). Older drivers and accidents: A meta analysis and data mining application on traffic accident data. Expert Systems with Applications, 29, 598-629.

Cameron, M. (1997). Accident data analysis to develop target groups for countermeasures. Monash University Accident Research Centre, Reports 46 & 47.

Chang, L.Y., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Safety Science, 43 (8), 541-557.

Chang, L. and W. Chen (2005). Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research, 36, 365-375.

Chang, L.Y. and H. W. Wang (2006). Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention, 38 (5), 1019-1027.

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

Chong, M., A. Abraham and M. Paprzycki (2005). Traffic accident analysis using machine learning paradigms. Informatica, 29, 89-98.

Clarke, R., R. Forsyth and R. Wright (1998). Machine learning in road accident research: decision trees describing road-accidents during cross-flow turns. Ergonomics, 41 (7), 1060-1079.

Dougherty, M., 1995. A review of neural networks applied to transport. Transportation Research Part C, 3 (4), 247-260.

Geurts, K., G. Wets, T. Brijs and K. Vanhoof (2003). Profiling high frequency accident locations using association rules. Proceedings of the 82th Annual Meeting of the Transportation Research Board, Washington, D.C.

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

Kavsek, B., N. Lavrac and V. Jovanoski (2006). Apriori-sd: adapting association rule learning to subgroup discovery. Applied Artificial Intelligence, 20, 543-583.

Lee, C., F. Saccomanno and B. Hellinga (2002). Analysis of crash precursors on instrumented freeways. Transportation Research Record, 1784, 1-8.

Ljubic, P., L. Todorovski, N. Lavrac and J.C. Bullas (2002). Time-series analysis of UK traffic accident data. Proceedings of the Fifth International Multi-conference Information Society, Ljubljana, Svlovenia, pp. 131–134.

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

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

Smith, B. L., W. T. Scherer and T. A. Hauser (2001). Data mining tools for the support of traffic signal timing plan development. Transportation Research Record, 1768, 141-147.

Strnad, M., F. Jovic, A. Vorko, L. Kovacic and D. Toth (1998). Young children injury analysis by the classification entropy method. Accident Analysis and Prevention, 30, 689-695.

Tseng, W. S., H. Nguyen, J. Liebowitz and W. Agresti (2005). Distractions and motor vehicle accidents: data mining application on fatality analysis reporting system (FARS) data files. Industrial Management and Data Systems, 105 (9), 1188-1205.

Witten, I. H. and E. Frank (2005). Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, Amsterdam.

Yamamoto, T., R. Kitamura and J. Fujii (2002). Drivers' route choice behavior: analysis by Data mining algorithms. Transportation Research Record, 1807, 59-66.

Yang, H., R. Kitamura, P. P. Jovanis, K. M. Vaughn and M. Abdel-Aty (1993). Exploration of route choice behavior with advanced traveler information using neural network concepts. Transportation, 20 (2), 199-223.

Zeitouni, K. and N. Chelghoum (2001). Spatial decision tree- application to traffic risk analysis. Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon.