Exploratory research by means of data mining for investigating the relationship between the infrastructure, the drivers and the characteristics of offences

Final report: draft version

Presented to the Ran Naor Foundation

for the Advancement of Road Safety Research

Researchers: Dr. Shlomo Bekhor, Prof. David Mahalel, Prof. Joseph Prashker, Dr. Carlo Giacomo Prato, Dr. Ayelet Galtzur, Roni Factor.

August 2007

Table of Contents

List of figures and tables

1 Introduction

1.1 Motivation and background

Research on road safety has been conducted for several years, yet many issues still remain undisclosed and unsolved. Specifically, the relationships between drivers' characteristics and road accidents are not fully understood. It is not possible to accurately predict the odds that a driver may be involved in an accident, it is not known any particular insight on how personal and socio-economic characteristics might be related to the cause of accidents, and so on.

The lack of knowledge in this area causes problems in allocating efforts to decrease the number of accidents in an efficient manner. It is difficult, for example, to elaborate focused training programs that are well suited to drivers with different backgrounds. It is not available at hand a risk profile of a driver based on personal or socio-economic group characteristics. As a result, the training programs might not produce the expected results.

Conventional statistical methods, such as Poisson or Negative Binomial regression models, have been employed to analyze vehicle accident frequency for many years. However, these models have their own assumptions and pre-defined underlying relationship between dependent and independent variables. If these assumptions are violated, the model could lead to erroneous estimation of accident likelihood.

Data mining techniques have been commonly employed in business administration, industry, and engineering. These techniques do not require any pre-defined underlying relationship between target (dependent) variable and predictors (independent variables) and have been shown to be a powerful tool, particularly for dealing with classification and prediction problems. Therefore, the motivation of this research is the evaluation of the potential of data mining in accident analysis and the investigation of relationships between accidents, drivers, and road conditions using these techniques with available databases.

1.2 Research objectives

The proposed research will investigate several data mining techniques in an attempt to find the most suitable method for accident analysis and to understand relationships between driver characteristics and accident data. We will explore the underlying variables that can lead to car accidents. The expected outcome of the research is a better understanding of the suitability of

data mining methods to the safety research field and of the relationships that could help in the planning of efforts to reduce car accidents.

The approach proposed in this research is purposely oriented to explore the accumulated knowledge available in existing databases. The significance of this research is in the development of new insights related to road accidents. These new insights will provide valuable help in developing new methods to increase road safety, particularly in the phase of choosing the appropriate means and budget allocation of resources.

The basic hypothesis of the research is that accidents are not randomly scattered along the road network, and that drivers are not involved in accidents at random. The hypothesis is that there are complex circumstantial relationships between the several characteristics (driver, road, car and so on) and the accident occurrence.

Similar to Kononov and Janson (2002), we also believe that it is not possible to develop efficient means to improve safety without developing the ability to relate frequency and severity of accidents to the several variables that might affect them.

2 Literature review

2.1 Data mining

Knowledge discovery databases (KDD) and data mining aim at extracting useful knowledge from large collections of data, meaning to find interesting patterns and/or models that exist in databases but are hidden among the large volumes of data. KDD is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data. Data mining is a step in the knowledge discovery process consisting of specific algorithms that, under some acceptable computational efficiency limitations, find patterns or models in data.

Data mining can be applied through the implementation of one among several methods. The choice of the method to be implemented stems from persistence between the characteristics of the data mining method and the nature of the problem addressed. Laube (2001) summarizes several data mining techniques for spatial dynamic data. Techniques for classification of datasets are diverse and include k-means, k-medoids, multiple linear regression, discriminant analysis, decision trees, k-nearest neighbour, neural networks, MARS and kernel methods (Loess smoothing). Following is a short description of three data mining methods which are more suited to the present research.

Clustering is a technique used for combining observed objects into groups or clusters such that each group or cluster is homogeneous or compact with respect to certain characteristics. That is, objects in each group are similar to each other. In addition, each group should be different from other groups with respect to the same characteristics. That is, objects of one group should be different from the objects of other groups

The nature of the clusters found enables to simplify the complexity of the entire problem and understand better the meaningful differences among the objects. Once clusters have been detected, other methods must be applied in order to figure out what the clusters mean. The most common method of clustering is the K-means method.

There are several advantages related to this method. The chief strength of automatic cluster detection is that it is undirected. Hence, it can be applied even when there is no prior knowledge of the internal structure of a database. In addition, by choosing different distance measures, automatic clustering can be applied to almost any kind of data. Most cluster

detection techniques require very little manipulation of the input data and there is no need to identify particular fields as inputs and others as outputs.

Along with the advantages mentioned above, the performance of automatic cluster detection algorithms is highly dependent on the choice of a distance metric or other similarity measure. For example, in the K-means method, the original choice of a value for K determines the number of clusters that will be found. If this number does not match the natural structure of the data, the technique will not obtain good results. In addition, the aforementioned strength of automatic cluster detection as an unsupervised knowledge discovery technique might result in clusters that have no practical value.

Decision trees are powerful and popular tools for classification and prediction as they represent rules. Rules can readily be expressed so that we humans can understand them and implement them in a database access language. The ability of decision trees to generate rules, which can be translated into comprehensible English or SQL, is the greatest strength of this technique. Even when a complex domain causes the decision tree to be large and multifaceted, it is generally fairly easy to follow any one path through the tree. So the explanation for any particular classification or prediction is relatively straightforward.

The algorithms used to produce decision trees generally yield trees with a low branching factor and simple tests at each node. Typical tests include numeric comparisons, set membership, and simple conjunctions. When implemented on a computer, these tests translate into simple Boolean and integer operations that are fast and inexpensive. Decision-tree methods are equally adept at handling continuous and categorical variables. Categorical variables, which pose problems for neural networks and statistical techniques, come ready-made with their own splitting criteria. Continuous variables are equally easy to split by picking a threshold in their range of values .

Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous variable such as the sufficient training period for drivers. The method is also problematic for time-series data unless a lot of effort is put into presenting the data in such a way that trends and sequential patterns are made visible. Some decision-tree algorithms can only deal with binary-valued target classes. Others are able to assign records to an arbitrary number of classes, but are error-prone when the number of training examples per class gets

small. This can happen rather quickly in a tree with many levels and/or many branches per node .

Association rule analysis is a method based on transaction analysis and produces rules underlying dependencies within the data. The association rules in the form of “if-then” rules makes the results easy to understand and facilitates turning the results into action. The concepts behind association rules and suggested algorithms for finding such rules were first introduced by Agrawal et al. (1993). Generating association rules involves looking for frequent item sets in the data. By looking for frequent item sets, it is possible to determine the support of each rule.

This technique can be used to efficiently search for interesting information in large amounts of data. Informally, the support of an association rule indicates how frequently that rule occurs in the data. The higher the support of the rule, the more prevalent the rule is. Confidence is a measure of the reliability of an association rule. The higher the confidence of the rule, the more confident we are that the rule really uncovers the underlying relationships in the data. It is obvious that we are especially interested in association rules that have a high support and a high confidence.

Association rule analysis is an appropriate technique, when it can be applied, to analyze a large set of data for which the starting point is unknown. The technique can handle variable-length data without the need for summarization, as they can handle transactions without any loss of information. In addition, the computations needed to apply association rule analysis are rather simple, although the number of computations grows very quickly with the number of transactions and the number of different items in the analysis .

Probably the most difficult problem related to this technique is determining the right set of items to use in the analysis in a way that the frequencies of the items used in the analysis are about the same. As this method works best when all items have approximately the same frequency in the data, items that rarely occur are in very few transactions and will be pruned.

After the data mining techniques are implemented, it is necessary to interpret (post-process) discovered knowledge, especially the interpretation in terms of description and prediction - the two primary goals of discovery systems in practice.

Feelders et al. (2000) described the different stages in the data mining process and discussed some pitfalls and guidelines to circumvent them. In their paper, they exemplified the correct procedure to treat data from several sources, using accident data as an example. Despite the predominant attention on analysis, data selection and pre-processing are the most time-consuming activities, and have a substantial influence on ultimate success. Successful data mining projects require the involvement of expertise in data mining, company data, and the subject area concerned. Despite the attractive suggestion of "fully automatic" data analysis, knowledge of the processes behind the data remains indispensable in avoiding the many pitfalls of data mining.

2.2 Data mining implementations in the transportation and traffic safety areas

There has been an increasing interest in applying data mining techniques in the transportation literature in recent years. Smith et al. (2001) investigated the application of statistical clustering and classification techniques to aid in the development of traffic signal timing plans. The authors used the k-Means Hierarchical Cluster Analysis to identify temporal interval break points that support the design of a signal control system. The results of their research indicated that advanced data mining techniques held high potential to provide automated tools that assist traffic engineers in signal control system design and operations.

Yamamoto et al. (2002) applied decision trees and production rules, which are among the methods used in knowledge discovery and data mining, to investigate drivers' route choice behavior. However, in current practice, relatively little information has been successfully extracted from the wealth of data collected by intelligent transportation systems (ITS).

In recent years, exploratory research has been conducted using data mining techniques with the purpose to discover relations between the several characteristics that affect accidents such as the road, driver, car, day of week, season, and so on. For example, Cameron (1997) indicated that clustering methods are an important tool when analyzing traffic accidents as these methods are able to identify groups of road users, vehicles and road segments which would be suitable targets for countermeasures.

Lee et al. (2002) performed a review of classical statistical 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 such large dimensions such as an exponential

increase in the number of parameters as the number of variables increases and the invalidity of statistical tests as a consequence of sparse data in large contingency tables.

Mussone et al. (1999) used neural networks to analyze vehicle accident that occurred at intersections in Milan, Italy. They chose feed-forward Multi Layer Perceptron neural networks using BP learning. The model had 10 input nodes for eight variables (day or night, traffic flows circulating in the intersection, number of virtual and real conflict points, intersection type, accident type, road surface condition, and weather conditions). The output node was called an accident index and was calculated as the ratio between the number of accidents for a given intersection and the number of accidents at the most dangerous intersection. Results showed that the highest accident index for running over of pedestrian occurs at non-signalized intersections at nighttime.

In the late 90's and beginning of this century there have been several attempts to use data mining techniques in road accidents analyses. For example, clustering techniques (Ljubic et al., 2002) and spatial data mining (Zeitouni and Chelghoum, 2001) were used to discover frequent patterns in accident data. Additionally, the data mining technique of rule induction were used to identify rule sets representing interesting subgroups in accident data (see e.g. Kavsek et al., 2002). It seems that decision trees (see e.g. Strnad et al., 1998; Clarke et al., 1998) and neural networks (see e.g. Mussone et al., 1999) were the predominant methods used to model and analyze road accidents.

Ng, Hung and Wong (2002) used a combination of cluster analysis, regression analysis and Geographical Information System (GIS) techniques to group homogeneous accident data together, to estimate the number of traffic accidents and to assess the risk of traffic accidents in a study area.

Bayam et al. (2005) illustrate how data mining techniques could be used to analyze relationships between senior driver characteristics and accidents. The most frequent policy recommendation to improve senior drivers' safety is to increase light-controlled intersections with protected left-turn signals.

Chang and Chen (2005) analyzed freeway accident frequency using 2001–2002 accident data of National Freeway 1 in Taiwan. The authors developed both decision tree models and a negative binomial regression model to establish the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors.

The decision tree findings indicated that the average daily traffic volume and precipitation variables were the key determinants for freeway accident frequencies. By comparing the prediction performance between the decision tree and the negative binomial regression models, the authors found that decision tree is a better alternative method for analyzing freeway accident frequencies.

Chong et al. (2005) evaluated the performance of four machine learning paradigms applied to modeling the severity of injury that occurred during traffic accidents. They considered neural networks trained using hybrid learning approaches, support vector machines, decision trees and a concurrent hybrid model involving decision trees and neural networks. Experiment results reveal that among the machine learning paradigms considered the hybrid decision tree-neural network approach outperformed the individual approaches.

An interesting application of data mining methods to investigate accident data can be found in Geurts et al. (2003). In the initial part, the researchers found two main clusters in their accident data, obtained from the National Institute of Statistics for the region of Flanders (Belgium). The first cluster included 35 accidents in 13 traffic roads, whereas 107 traffic accidents in 6 traffic roads were included in cluster 2. After obtaining the clusters, the researchers profiled them in terms of accident related data and the degree in which these characteristics can discriminate between the clusters. In their research, association rules were used to identify accident circumstances that frequently occurred together at high frequency accident locations. A comparative analysis between high frequency and low frequency accident locations was conducted to determine the discriminating character of the accident characteristics of black spots and black zones. For example, they found that sideway collisions involving female road users are typical accident pattern for traffic roads with high accident risk. This type of accident occurs when the maximum speed limit was 50 Km/h, when non priority is given and the age of at least one road user was between 18 and 29.

Table 1 summarizes the literature on data mining techniques applied to road safety projects that were presented in this section.

AUTHOR	YEAR	SUBJECT	DATA MINING TECHNIQUE
Cameron	1997	Developing target groups	Cluster Analysis
Strand et al.	1998	Young children injury analysis	Decision Trees
Clarke et al.	1998	Cross-flow turn accidents	Decision Trees
Mussone et al.	1999	Accidents at intersections	Neural Networks
Zeitouni and Chelghoum	2001	Traffic risk analysis	Decision Trees
Ng et al.	2002	Risk of traffic accidents	Cluster Analysis
Geurts et al.	2003	High frequency accidents (black spots)	Association Rules
Brijs et al.	2003	Ranking hazardous sites	Bayesian Model
Kavsek et al.	2002	Subgroups in accident data	Rule Induction
Bayam et al.	2005	Senior drivers' characteristics	Decision Trees
Chong et al.	2005	Injury severity in traffic crashes	Decision Trees and Neural Networks
Chang and Chen	2005	Freeway accident frequency	Decision Trees

TABLE 1. Summary of the main data mining applications in the traffic safety area

3 Methodological framework

This section provides information regarding the KDD process and the data mining techniques applied in this study. Further, some details regarding the options implemented in the same techniques are presented.

3.1 Description of the KDD process

Knowledge Discovery and Data Mining (KDD) is a multi-stage process. The overall methodology will follow the main stages of KDD process, as depicted in figure 1 and subsequently described.

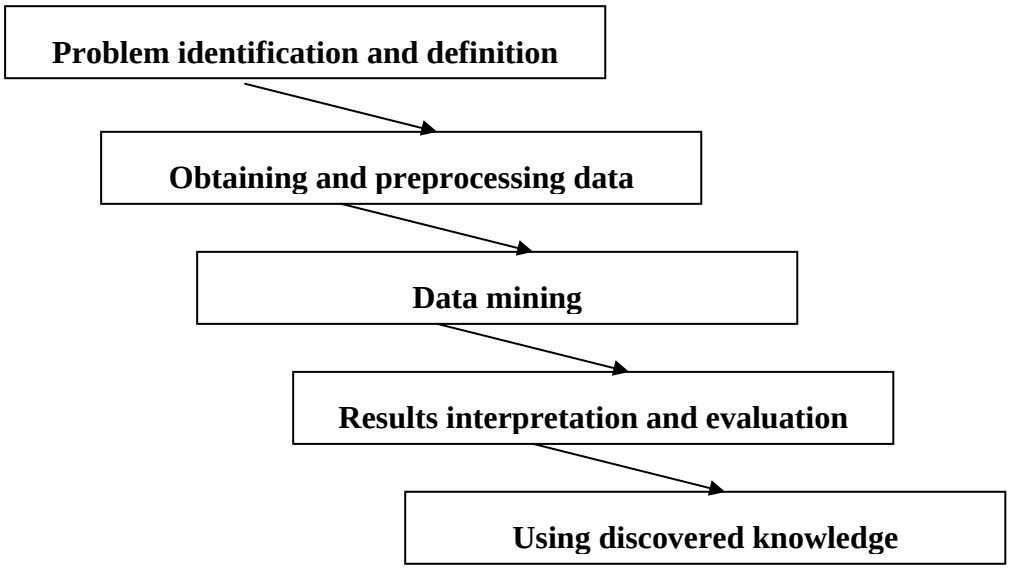

FIGURE 1. KDD process

Each one of the above steps has to be performed efficiently and thoroughly, in order to ensure the quality of the next step and of the final outcome of the entire process.

3.1.1 Problem identification and definition

The first step in the process is the identification and the definition of the problem itself. It requires deep understanding of the domain to which the KDD is applied. More specifically, it requires the definition of the type of knowledge that the KDD process is aimed to find and the data that serves as the basis for the searching process.

3.1.2 Obtaining and preprocessing data

This step is composed of several sub-steps and requires understanding of both the domain to which the KDD is applied and of statistical methods for improving data quality. Hence, the cooperation of experts from various disciplines is most crucial.

3.1.2.1 Selecting data sources

The selection of the data sources reflects the decisions made in the Problem identification and definition stage regarding the type of data required for obtaining new knowledge. Ensuring that the chosen data sources contain all the relevant data is not sufficient. Considerations regarding the completeness of the data and its reliability must be taken into account. These aspects are especially important in cases where some of the information is contained in more than one data source.

3.1.2.2 Treating missing and incorrect data

In many cases the raw data contains missing information and/or incorrect data. Improving the data quality requires the elimination of missing and incorrect data. Alternatively, missing and incorrect data can be replaced by valid values.

Identifying missing data is rather straightforward, while tracking invalid values is more complicated. In order to seek incorrect information it is required to define rules for the validity of the range of values for each attribute within the data set.

Once the task of marking the missing and incorrect data is complete, it is required to either eliminate it or replace it by reliable values. Naturally, an elimination process is a rather simple one, but sometime results in a set of examples that is too small for the data mining process. Hence, filling the missing data or replacing invalid values can be done by an inference process that stems from the knowledge of the domain for which the data mining is applied.

3.1.2.3 Data transformation

Data transformation aims to manipulate the data so that its content and its format are most suitable for the data mining process. The transformation process effects the distribution of the various features and the structure in which their values are stored. Various data mining

techniques present different requirements regarding these characteristics of the data. The requirements of each technique should be taken into account prior to its application.

Handling the distribution of certain features is essential for the one whose values have undesired characteristics such as skewed distribution. The desired distribution is achieved by applying mathematical transformation function over the data, such as normalization methods or smoothing techniques. As emphasized by Witten et al. (2005), this process should be done carefully in order to avoid the creation of artifacts data structures, loosing fundamental relationships between various features or diminishing extreme values that reflect a rare but meaningful phenomenon.

Handling the format of the data involves the transformation of values into specific structures and boundaries. These process should consider structures such as binary representation (true or false) or discrete representation (dividing analog values into a finite number of categories etc.).

3.1.2.4 Reducing the data dimension

Reducing the dimension of the data is required when the size of the database damages the efficiency of the data mining process. Harming the efficiency might be reflected either in a very long process time or in patterns found by the data mining process that are misleading.

There are several methods for data reduction that are effective, but one must bear in mind that they are all imperfect. The implementation of the various methods should preserve the characteristics of the original database. The two main methods are:

- ✓ feature elimination;
- ✓ example elimination.

Feature elimination involves the examination of the various features while attempting to identify those with low predictive potential. Features that are considered to be poor predictors or are redundant in relation others, can usually be discarded. It is also possible to combine two or more feature into one, as long as this aggregation process preserves the essence of the information.

When applying example elimination methods, the representative nature of the database should be preserved. Hence the statistical sampling rules should be implemented.

3.1.3 Data mining

The data mining step is the essence of the KDD process. It involves the application of several data mining techniques to the example database. The output of this step is the knowledge extracted by the data mining algorithm. Each data mining technique produces the knowledge in a different format, such as a decision tree, decision rules, clusters etc. It is often hard to predict which technique will produce the best results, and it is unnecessary. Moreover, sometimes the aggregation of knowledge produced by more that one technique provides the best final solution.

3.1.4 Results interpretation and evaluation

After the data mining techniques are implemented, it is necessary to interpret (post-process) the discovered knowledge. The interpretation is required in terms of description and prediction - the two primary goals of discovery systems in practice. In this stage, as noted by Feelders et al. (2000), it is crucial to involve the expertise in data mining, company data, and the investigated domain. This involvement is needed to correctly assess underlying processes occulted behind the data and to avoid the many pitfalls of data mining.

3.1.5 Using discovered knowledge

The final step is to put discovered knowledge in practical use either by documenting it and reporting it or by embedding it in a computer system. In first sight, this stage might be regarded as trivial and straightforward, but this is not the case. The conclusions drawn from the KDD process often reveal the complex nature of the problem and its solutions. This is not surprising as data mining techniques are not necessary when dealing with simple problem. Hence, the implementation of the new knowledge should often be done in gradually, while continuously monitoring the result achieved and the degree to which they fulfill the expectations.

3.2 Data mining methods

The most widely used methodology in Data Mining is the Cross Industry Standard Process for Data Mining, known as CRISP-DM. The CRISP-DM methodology consists of an iterative

process consisting of the following six phases: business understanding phase, data understanding phase, data preparation phase, modeling phase, evaluation phase, and deployment phase. The Clementine software employed for this research adopts the CRISP-DM methodology to analyze problems.

This section illustrates the data mining techniques applied in the research, 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 behaviors based on mapping of a set of input values to an output value.

K-means and Kohonen networks perform the task of segmenting a heterogeneous population into more than one homogeneous subgroup, thus they fall under the descriptive analysis part of data mining applications. Decision trees, association rules and neural networks examine the data and estimate the outcome values of a dependent variable, consequently they fall under the predictive analysis part of data mining implementations.

3.2.1 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 for a specific execution by the researcher, while “means” refers to the cluster being represented by the mean of observations on the selected variables.

The number of clusters intends to reduce the dimensionality of the problem, but the definition of this number is arbitrary and standard practice suggests trying solutions with different number of clusters, while examining each time the result in order to comprehend which one is most useful.

The implementation of K-means clustering in this research utilizes the “maximin” method to first select cluster centers. Initially, the algorithm positions the first cluster center as the first record of the data file, and the remaining centers are created by searching for positions in an n-dimensional space that are as far as possible from any other cluster centers already generated. The initial cluster centers consequently cover as large a data range as possible.

Then, the algorithm calculates the Euclidean distance between each record and every cluster center and assigns each record to the cluster center with the smallest squared Euclidean distance. After the assignment of all the cases to a cluster group is completed, the location of

each cluster center is recomputed as the average of all the cases within the cluster. This iterative process moves the cluster centers and reiterates until one stopping criteria is reached, namely a change in means below a defined threshold or a maximum number of iterations.

The process in a small dataset with only two input fields is represented graphically in figure 2. In a three cluster solution, three largely spaced records are chosen, distances are calculated and the cluster centers recalculated accordingly. Consequently, the centers move through the two-dimensional space to their final positions. The solid lines represent the boundaries between the clusters.

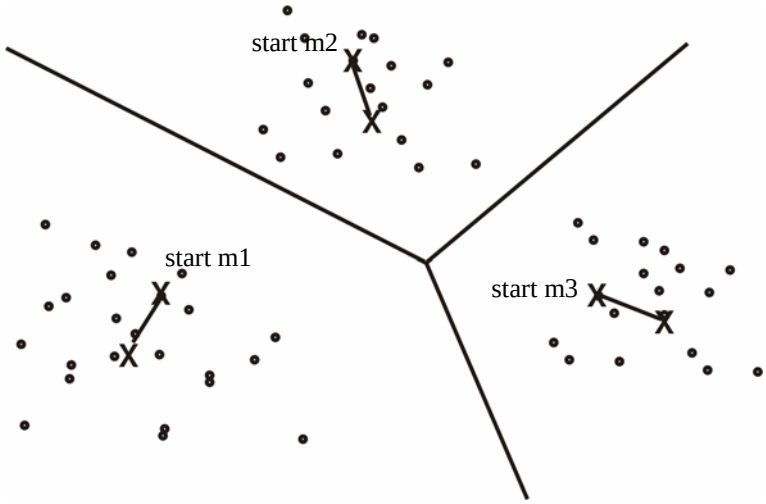

FIGURE 2. Illustration of K-means clustering

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 20 iterations maximum.

3.2.2 Descriptive analysis: 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 basic assumption considers clusters as 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. Each neuron is also connected to the surrounding networks, as illustrated in figure 3.

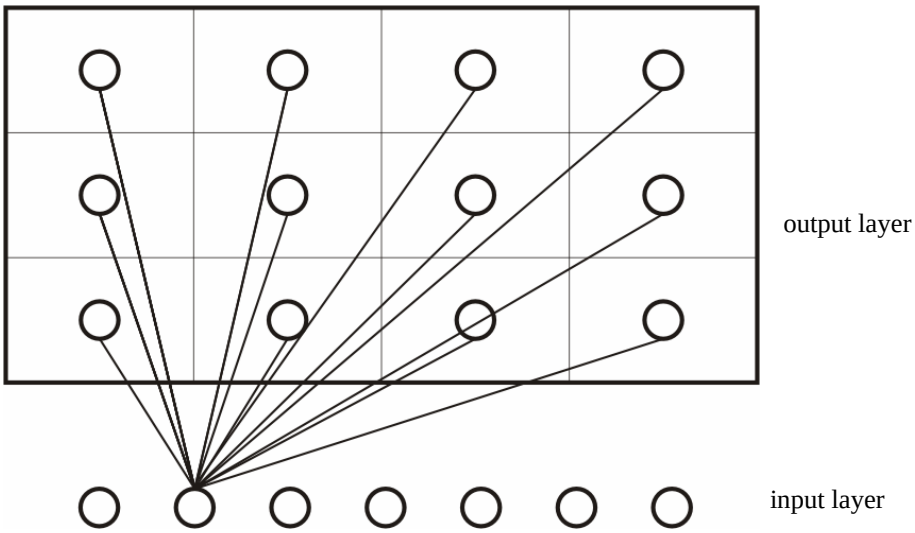

FIGURE 3. Illustration of a Kohonen network

The network trains by presenting cases to the grid. Characteristics of each record are compared with those of all neurons in the grid, after giving random weights initially. The neuron with the most similar pattern “gets” the examined network, and the weights of the artificial neuron are adjusted to be more similar to that of the record just acquired. This enhances the likelihood of similar records to be captured by the same node. The network adjusts the weights of the surrounding neurons as well, as each case enters the examination phase. After the data pass through the network a number of times, the result consists of a map containing cluster of records corresponding to different types of patterns in the data. Similar patterns should be closed in the map than patterns that are dissimilar.

The algorithm works in two phases: a first stage with initial large-scale changes and a second stage with smaller changes in the weights in order to perform a fine-tuning of the map. The algorithm stops when the change in weights between cycles is small or the maximum number of iterations is reached. As in any neural network, a learning parameter *eta* is defined for the network process. Kohonen networks consist of two phases and define a different learning parameter for each phase. Contemporarily, the algorithm requires the number of neighbors around the acquired node for each phase.

Given the training process that involves long iterations with weight adjustments, the Kohonen network takes longer to train than K-means clustering. These networks provide a different and valuable view of patterns in the data, and possibly an alternative one with respect to other clustering methods.

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, neighbors equal to 2 and 20 cycles. The second phase has learning parameter equal to 0.1, neighbors equal to 1 and 100 cycles. Convergence is reached when the learning parameter arrives to a null value.

3.2.3 Predictive analysis: decision trees

Decision trees constitute a method able to forecast or classify future observations according to decision rules. If the information is divided in classes, it is possible to utilize the data to generate rules able to classify previous cases and new cases with absolute precision.

This approach is known as advanced induction rules. The decisional process behind the model appears clear when observing the tree, with a clear advantage with respect to other techniques whose internal logic is difficult to interpret. Further, the process includes automatically in the rule only the attributes relevant to the decision, as the irrelevant attributes are ignored and the data dimensions are reduced.

Decision trees are presented with their actual configuration when the description of the partition and classification of the data is valuable information, as illustrated in figure 4. Decision trees are converted into “if-then” rules to enhance the comprehension of the model when the relationship among the elements of a group is relevant.

Among the different algorithms for constructing decision trees, C5.0 and CHAID are the approaches that split categorical predictors and appear suitable for accident analysis given the nature of the fields in the database.

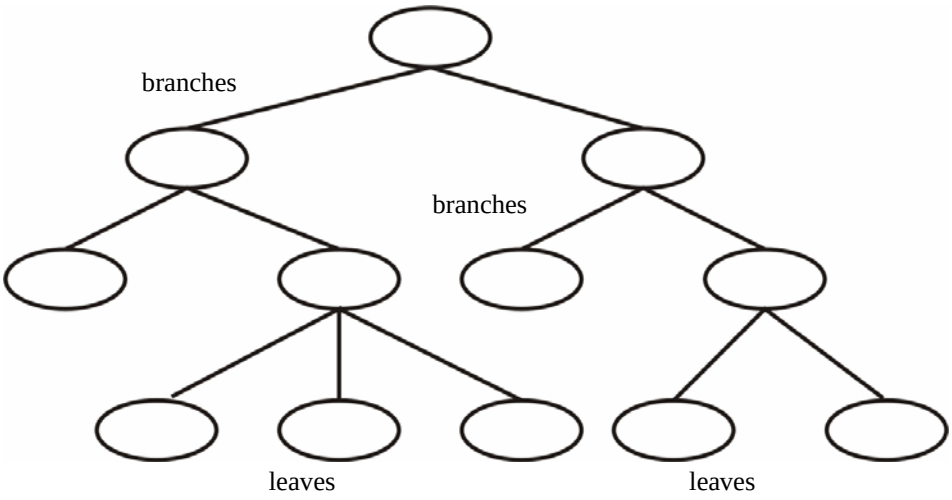

FIGURE 4. Illustration of a decision tree

The algorithm C5.0 divides the records according to the field that yields the maximum information gain. Each subgroup defined at the first division is further examined, generally according to a different field. The process is reiterated 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.

Note that the 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, tends to favor partitions containing large number of outcomes and to present an advantage for a symbolic predictor with many categories over one with few categories. In order to avoid this deficiency, the C5.0 algorithm calculates the gain ratio by dividing the information gain for the potential information generated by partitioning the data into n outcomes, whereas the information gain measures the information relevant to classification.

When the algorithm C5.0 produces a decision tree, each leafs describes a certain subgroup of the training data and each record enters only one single leaf. When the algorithm produces induction rules, these rules 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. With the decision tree, the forecast for each record is unique. With the induction rules on the other hand, the forecast for each record is weighted according to the relevance of the different rules for the case itself.

The CHAID algorithm, acronym of Chi-squared Automatic Interaction Detection, utilizes chi-squared statistics to identify optimal subdivisions of the dataset. Initially, the CHAID method 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 this variable that are yielding similar results, while proceeds with the division in subgroups of the data according to the new categories created. The merge of the categories when the difference among the remaining categories is equal to the difference obtained with the independency test.

In general, the C5.0 models are stable in presence of missing data and large number of input fields, do not require long training time and their interpretation are easier to be interpreted. The CHAID algorithm is efficient in presence of missing data and generates trees for categorical predictors with more than one branch for each subgroup. For predicting purposes, part of the data is used for training and part of the data is used for test. The comparison between the actual and the predicted values provides a measure of goodness-of-fit of the estimated models.

The implementation of the C5.0 algorithm in this research proposes the definition of a 75% pruning and of 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 to favor this union, 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.

3.2.4 Predictive analysis: 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 strengths or weights associated to them. The learning adapts the weights at each iteration and provides the system of a method to learn by example.

The Multi-Layer Perceptron (MLP) is currently the most popular type of neural network. The MLP network is a simplified model of the human mind elaboration process, and works by simulating and elevated number of simple elaboration units that resemble abstract versions of neurons. As illustrated in figure 5, 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.

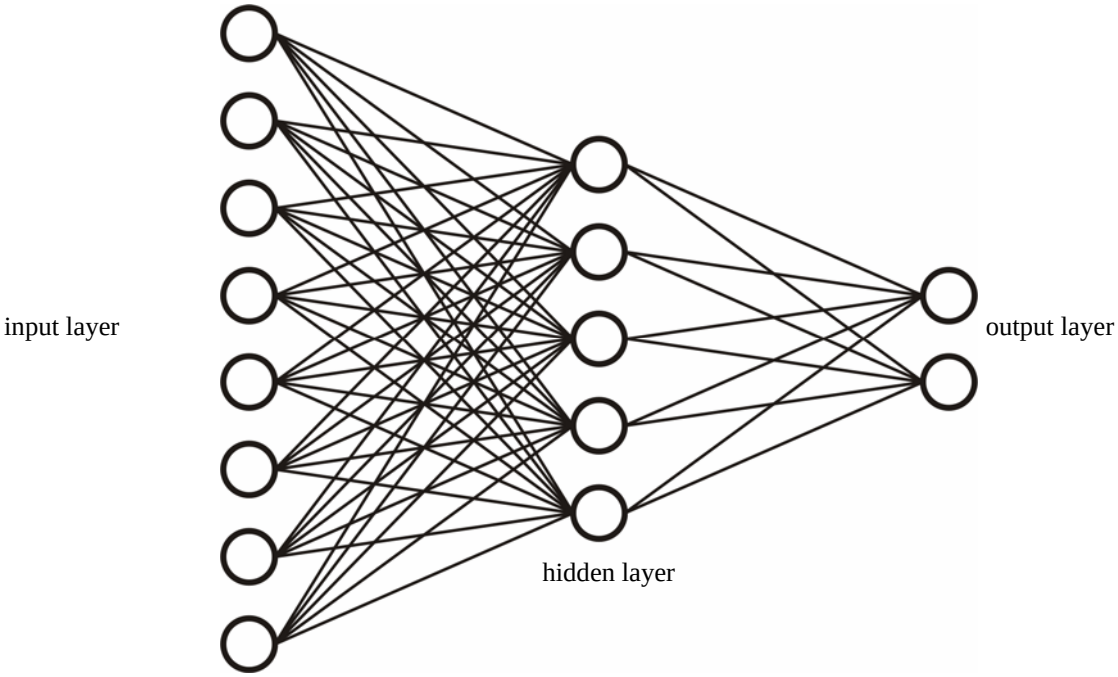

FIGURE 5. Illustration of a MLP neural network

The network learning process consists of the exam of single records, the forecast for each case and the correction of the weights each time a forecast is incorrect. This process reiterates and the network improves its forecasts until one or more interruption criteria are satisfied.

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 and the provided answers are compared to these output, consequently 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.

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.

Several MLP algorithms are available to the analyst. The Quick method creates a network with one hidden layer containing varying number of neurons according to the number and type of input fields. The Dynamic training considers an initial network with two hidden layers of two neurons each that grows dynamically by adding one neuron at each layer until no benefit is given by growing attempts. The Prune method uses a large one or two hidden layer network and removes the weakest neurons according to a sensitivity measure until no benefit is given by removal attempts. The Exhaustive Prune training technique invokes a more exhaustive examination of the network and removes less weak networks at each iteration with respect to the Prune technique, as the criteria appear much stricter.

Following the same concept of the decision trees, also in the neural network applications the dataset is divided into a training and a validation part, as the test part consists of part of the input data used by the algorithm to determine when the training phase reaches the stop. Again, the comparison between the actual and the predicted values provides a measure of goodness-of-fit of the constructed neural networks.

The implementation of MLP in this study uses the 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 *hidden rate* of eliminated neurons is equal to 0.02 at each iteration 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.

3.2.5 Predictive analysis: 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 typically. Unfortunately, the search space for independent rules is exponential with the number of attributes, thus association rule algorithms are computationally expensive.

Given that an association rule consists of some conditions, also named antecedents, that are followed by some conclusions, the evaluation of the rules necessitates two criteria: the support is the percentage of records in the dataset for which the conditions hold, 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 the conclusion, given that the conditions are met.

The Apriori rule discovery algorithm works only with symbolic data that are coded as flag fields and indexes and minimizes 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 the 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.

4 Data description

Assembling, integrating and cleaning the data were the initial task in the accident analysis with the illustrated data mining techniques. This section addresses the description of the structure of the data, by introducing exploration and verification of the data quality and illustrating selection, cleaning and integration of different data sources.

4.1 Data preparation

Data about the accidents occurred in Israel were provided by the Central Bureau of Statistics, on the condition that the crash resulted in the injury of at least one person involved. This research utilized two different databases for accident analysis purposes: the first data source contained detailed information about the accidents, while the second data source provided additional information about the drivers involved, retrieved by the census data.

4.1.1 CBS database

The Central Bureau of Statistics collected information regarding accidents that occurred in Israel and resulted in the injury of at least one person involved in the crash. Every year thousands of accidents were reported and for each year databases consist of lists of records in which each single record corresponds to a single crash.

For each year, three different files recorded every accident with injury: the accident file, the vehicle and driver file, the injured file. Overall, detailed information covered every aspect of the accident, from the location to the characteristics of the infrastructure, from the vehicles to the persons involved, from the weather condition to the traffic light situation, and so on. The following paragraphs present an excursus of the content of the different files.

The first piece of information in the accident file included information regarding when and where the accident occurred: date and time, urban or interurban location, intersection or road section, police district. Then, the accident was classified according to three levels of severity, along with the fact that someone involved in the crash actually died (fatal), sustained severe injuries (severe) or resulted lightly injured (light), and according to the definition of several accident types and the description of both the modality and the cause of the crash. Then, the infrastructure was described: allowed speed, presence and condition of median barrier, traffic light and road signals in general, condition of the surface related also to the weather at the

moment the accident occurred. Last, information involving pedestrian and collided objects was provided whenever necessary for specific typologies of accidents.

The vehicle and driver file included records of each vehicle and driver involved in the accident. Each record corresponded to one vehicle and its driver, and listed generic information such as the type, age, motor, weight and direction of travel of the vehicle, and the gender, age, licensing year and past offences of the driver. Clearly, the limited information provided in this file with respect to the details of the persons involved explained the further enrichment with census data executed in the second data source.

The injured file comprised records of each person injured, consequently added also pedestrian that were not listed in the previous files. Each record counted each injured person, and registered generic information such as gender, age, nationality (including the “aliya” process for Israelis born abroad), population group, residence and of course type of injury sustained. The same limitations applied to the data regarding the drivers.

One of the initial problems in checking the data quality was the change of coding system from year to year by the Central Bureau of Statistics. The consistency of the coding is important for analytical purposes, since the lack of common definitions for the same accident type or the same median barrier identification could lead to 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, accidents occurred in Israel between 2001 and 2004 were selected for the initial runs of the data mining methods. In this period the coding system was consistent and allowed avoiding potential problems related to the heterogeneity of the data.

The second problem consisted of the merge between the three data files. Note that the accident file contained one accident for each record, but the other two files contained multiple records corresponding to the same accident, as long as many drivers, vehicles and injured persons were involved. A unique file was composed by considering as a base file the accident file, and defining as vehicle 1, vehicle 2 and so on the vehicles. The same applied to the drivers and to the injured. At the end, a unique database listed as many records as accidents, each with its number of vehicles and injured involved. Note that the limitations of the number of columns in the SPSS program, used according to the format provided by the Central Bureau of Statistics, advised to eliminate the records with more than eight vehicles and consequently drivers involved.

The third problem involved the verification of the data quality and consequently the cleaning of the data files. The reported accidents contained missing values, either in the form of actual missing reported information or in the form of not applicable categorical variable to some accidents. Further, the listed records contained also typo errors (for example relative to drivers that were too young, such as 2 or 3 years old) that needed to be transformed in missing values before processing the data. Data mining techniques actually allow considering missing data, as the patterns of information from the large amount of records are not biased by the absence of some variable values for some records. The first data source for accident analysis consisted of 72,056 records, and the main characteristics of the database are summarized in section 4.2.1.

4.1.2 Enlarged database

The enlarged database included traffic accident data occurred in the period between 1996 and 2000, matched with information retrieved from the 1995 Israeli Census data.

The accident data file was different from the one described in the previous section, as some of the information was lost as the price to be paid for matching the census data. In particular, the condition of the infrastructure was described less thoroughly, the time at which the accident occurred was specified less precisely, and any information about pedestrians was removed, as well as the exact location of the crash. Information regarding median and number of ways was present for around 20% of the records, days and months were summarized for groups and not detailed as for the period between 2000 and 2004, and accordingly information regarding the season could not be defined accurately.

The 1995 Israeli Census included a "short questionnaire" and an "enlarged questionnaire". The "short questionnaire" was filled by the entire population, and contained general demographics variables, such as gender, age, marital status, religion etc. Twenty percent of the households were randomly chosen to answer the "enlarged questionnaire" to represent the entire Israeli household population. The "enlarged questionnaire" included details about the household, such as number of rooms, apartment ownership etc., and several personal details, such as number of children, work status, occupation, education, income etc.

The file containing the information about the traffic accidents was linked to the census file at the individual level by matching the Identification Number ("Teudat Zehut") of the drivers who were involved in car accidents. In this way each driver who was implicated in a car

accident from 1996 to 2000 was matched to the personal data from the 1995 census. The average linkage percent between accident data and the census data was about 92%, and when the quality of the linkage was tested by comparing driver's age and gender in the census data file to those in the accident data file, it was found that for 97.6% of the merged cases the age and gender were identical in both databases, and on additional 2.3% only the gender or the age were identical. These findings suggest that the merging procedure was fairly correct and that for each driver which was involved in a car accident we succeeded to find his or her data on the Census file.

The enlarged database was created by merging the accident and the census data, and each record corresponded to an accident with its corresponding characteristics, personal data of the drivers involved and details about their vehicles. Given the difficulties of the matching process, only single-vehicle and two-vehicle accidents were included in the analysis, which accounted for about 90% of all the accidents. Given the large amount of missing data, caused by the fact that only 20% of the population actually filled the enlarged census questionnaire, the information was elaborated in order to obtain some categorical variables for data mining applications.

Vehicles were categorized and the number of private, public, light commercial, heavy commercial vehicles implicated in each crash, as well as the number of motorcycles and bicycles involved, were defined in as many fields. The same procedure was followed for the gender and the age of the drivers, the population group, the religion, the place of birth: the variables were constructed as differences existing between the drivers (for example if they were both Jewish, both born in Israel, both belonging to the same generation that meant their age difference was below 10 years) and the content of the constructed categorical fields is detailed as input variables for descriptive and predictive analysis of this data source in section 5.1.

4.2 Data variables

The frequency analysis of the constructed data sources provides a first insight into the variables that are analyzed through the data mining techniques.

4.2.1 CBS database

The number of accidents analyzed in the period between 2001 and 2004 was equal to 72,056, distributed across the period according to figure 6. This graph suggests that there has not been a significant decrease in the number of accidents with reported injuries in the four year period, with practically an identical number of accidents in 2003 and 2004.

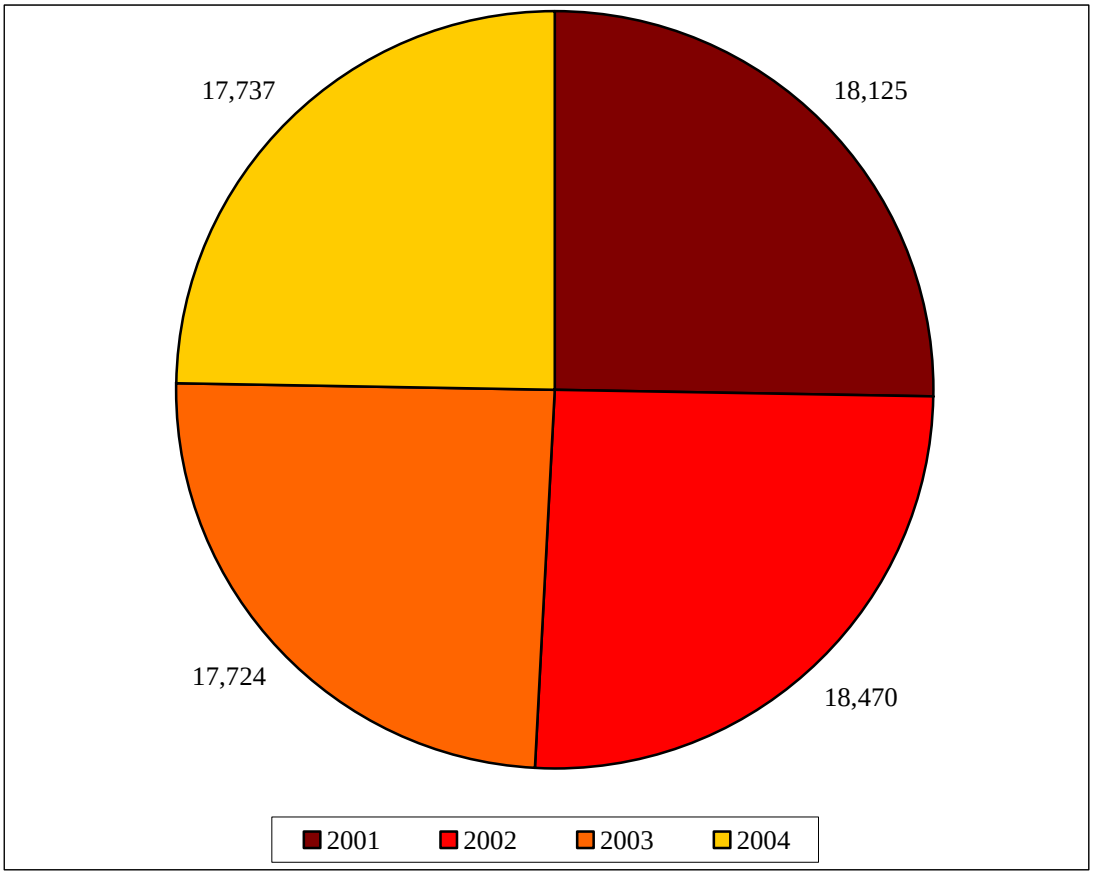

FIGURE 6. Number of accidents per year in Israel with reported injury

The analysis of the severity illustrates that less than 15% of the crashes resulted in a severe or fatal injury, as illustrated in figure 7. Nonetheless, fatal accidents decreased during the four year period, with a more clear cut decrease between the first two years and the two last years: from 471 fatalities in 2001 and 452 fatalities in 2002, the number went down to 415 and 425 accidents in 2003 and 2004.

Among the almost twenty typologies of collisions, slightly more than half (51.5%) occurred when the front of a vehicle hits the side of another vehicle, 11.3% were front-to-back crashes, 5.9% and 4.6% were respectively side-to-side and front-to-front accidents. Significantly, more

than one tenth of the crashes (14.7%) involved a pedestrian, who was typically crossing on zebras without a traffic light (43.6% of these crashes) or out of zebras and far from an intersection (34.7% of these crashes). Single-vehicle accidents constituted 23.7% and two-vehicles collisions corresponded to the 66.1% of the total records.

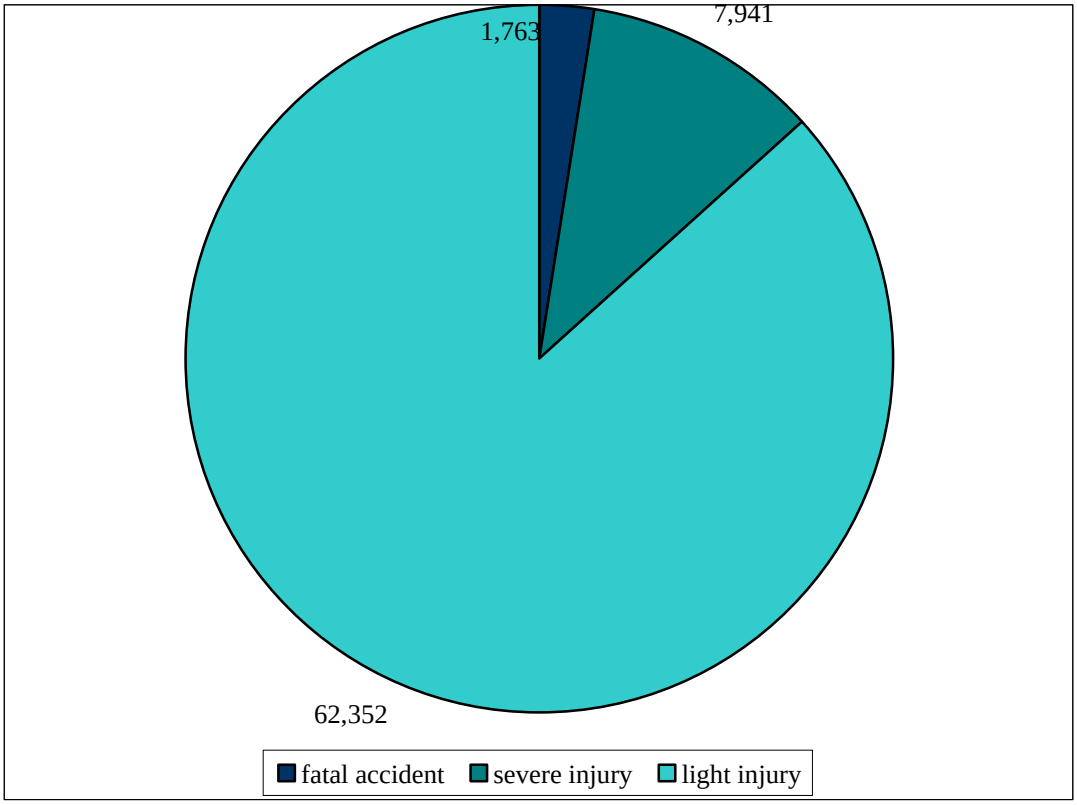

FIGURE 7. Number of accidents per level of injury

As shown in figure 8, more than 70% of the accidents happened in urban areas. Accordingly, the majority of the crashes concentrated in the Tel Aviv area (30.2%), the Haifa area (13.2%), the Jerusalem area (8.3%) and the Hasharon area (11.2%). Moreover, accidents were almost equally divided between intersections and sections.

The condition of the infrastructure was reported as good in more than 90% of the cases. Accordingly road signals were depicted as good in more than 85% of the records, light malfunctioning of road lights and traffic lights is described in around 1% of the accidents. Further, the good weather conditions in 90% of the crashes implied good road surface and not surprisingly more than 90% of the accidents were imputed to the drivers' fault. Considering that the records registered reported cases, it is possible that there is a tendency to impute the

fault to the drivers rather than to the infrastructure conditions. On the other hand, the average good conditions of the weather and the roads suggest that these are not the motives causing the accidents. Among the possible offenders, slightly more than 8% of the drivers had previous offences for speed exceeding and less than 2% reported problems of drugs or alcohol at the moment of the crash.

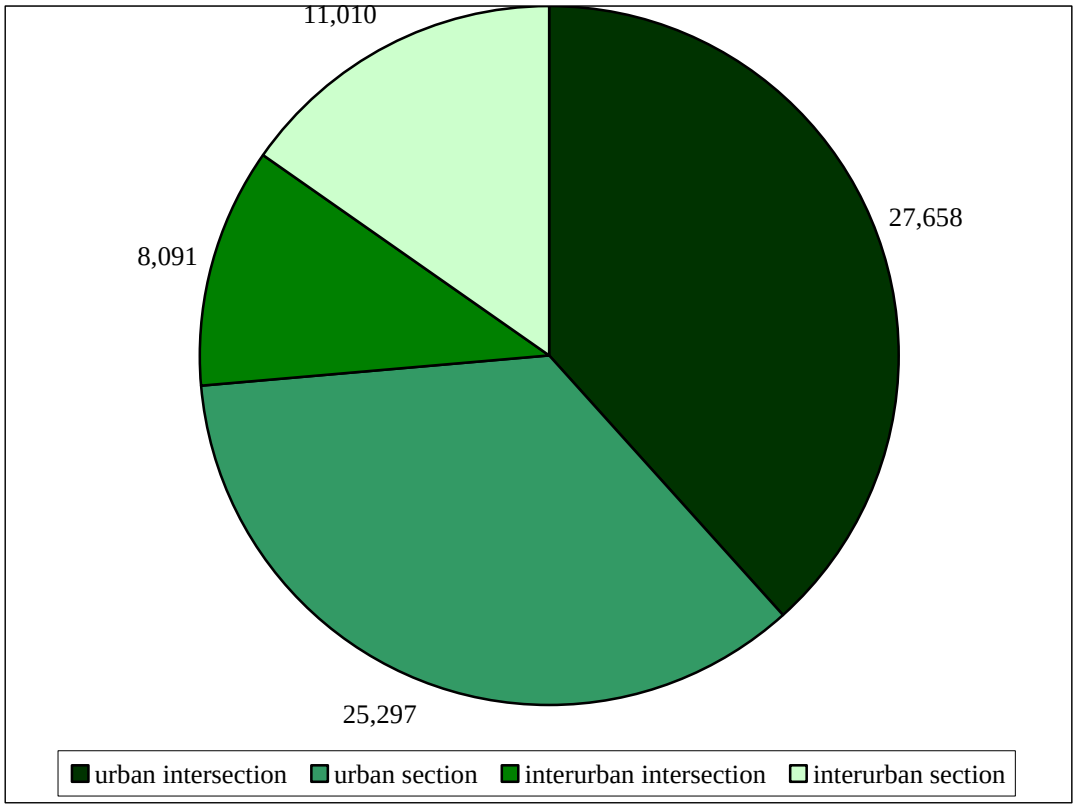

FIGURE 8. Number of accidents per location type

As illustrated in figure 9, more than half of the accidents occur in the evening and at night. The evening peak counts more crashes than the morning peak, and apparently during the day there is a trend of increasing cases apart from the evening period with less traffic. During the week the distribution of the accidents is equally divided across all days, except from Saturday when less crashes take place. During the year the distribution of the collisions is equally divided across all seasons, without exceptions.

With respect to the vehicles in the accidents, more than 10% of the collisions had a public vehicle involved, more than 20% had a light commercial vehicle implicated and almost 15% had a motorcycle drawn in. At least a man was involved in 90% of the accidents, and no

woman was implicated in 60% of the crashes, suggesting that men drive more than women and likely appear more aggressive.

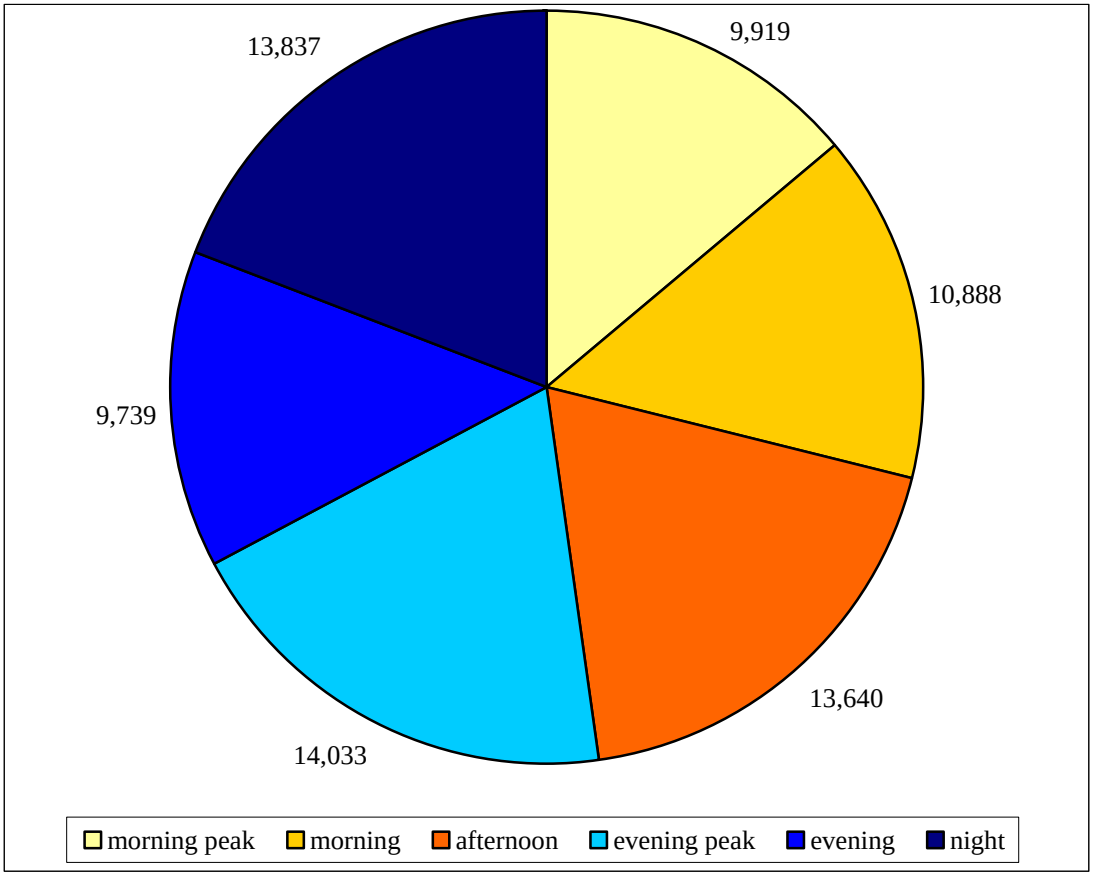

FIGURE 9. Number of accidents per periods during day

4.2.2 Enlarged database

The number of accidents analyzed in the period between 1996 and 2000 was equal to 105,812, distributed across the period according to figure 10. This graph illustrates the decrease in the number of accidents with reported injuries across the five year period, with stability in the first three years and a sensible diminishment in the other two years.

The analysis of the severity shows that less than 14% of the collisions resulted in a severe injury or a fatality, as illustrated in figure 11. The number of fatalities decreased proportionally to the number of total accidents, as the initial three years of this period were characterized by over 400 fatalities per year (with a peak in the year 1998) and the remaining two years saw this number decreasing to 375 and 388 deaths in 1999 and 2000 respectively.

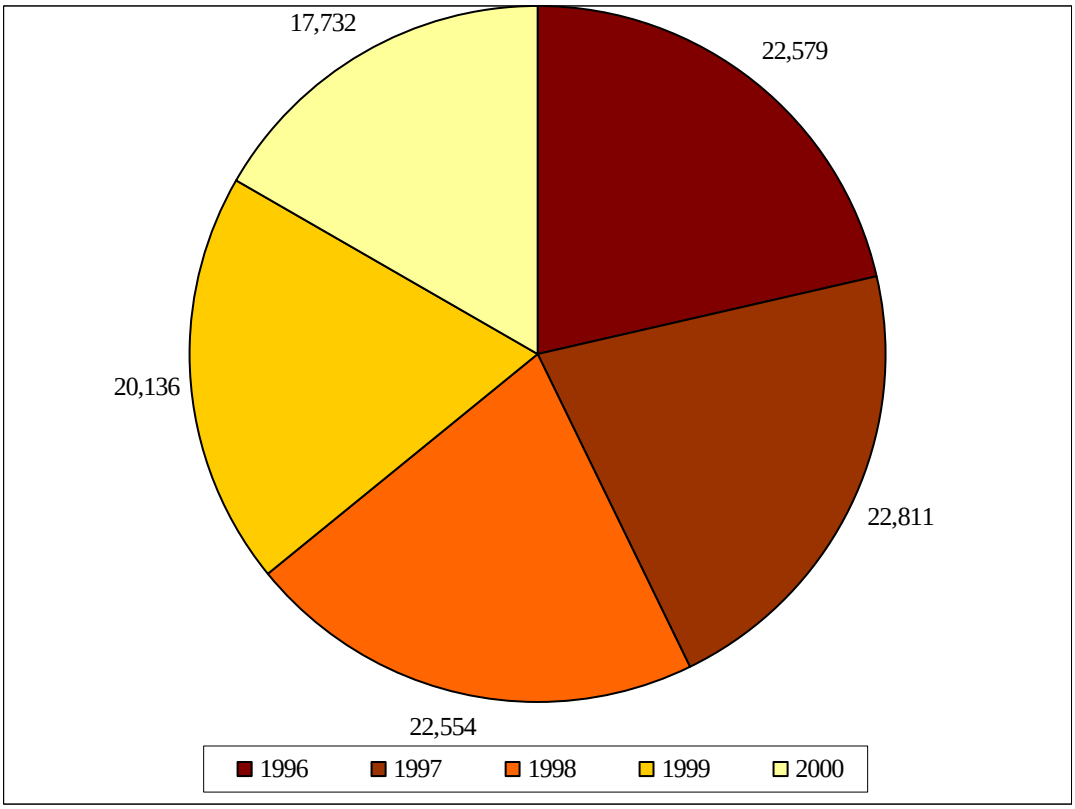

FIGURE 10. Number of accidents per year in Israel with reported injury

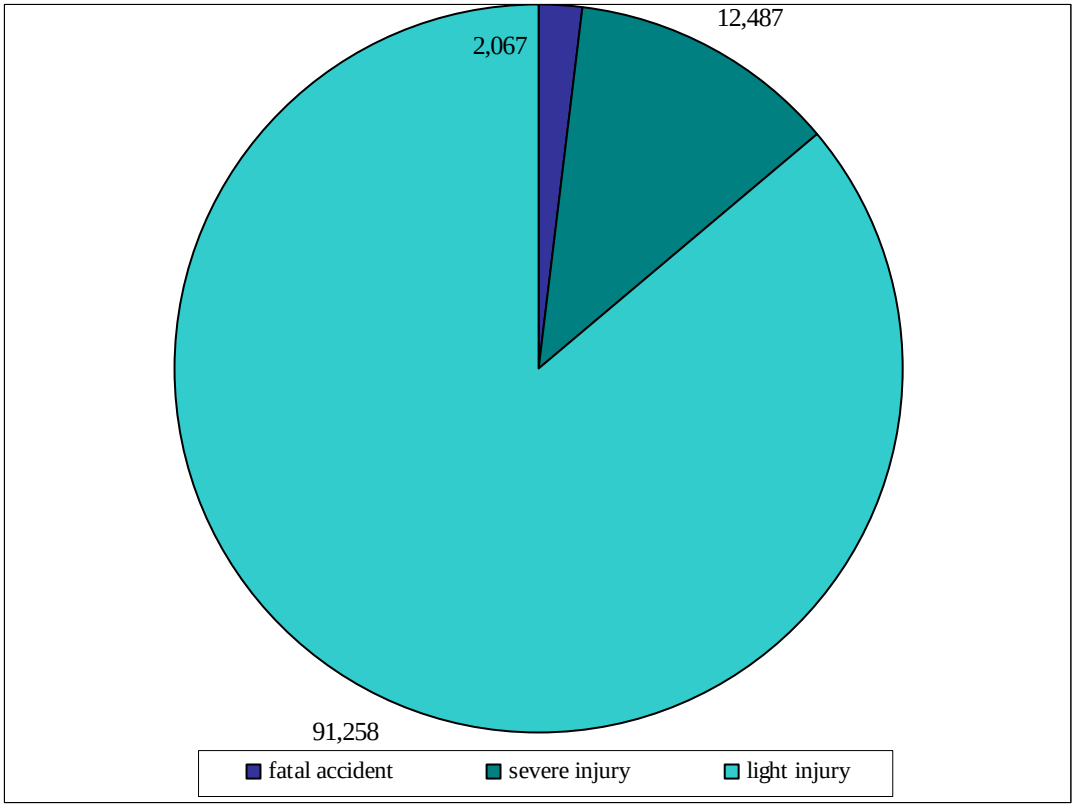

FIGURE 11. Number of accidents per level of injury

Among the different type of crashes, almost half (49.6%) consisted of the front of a vehicle hitting the side of another vehicle, 8.8% were front-to-back crashes, 5.7% and 4.1% were respectively side-to-side and front-to-front accidents. Significantly, almost one fifth of the crashes (18.1%) involved a pedestrian, and note that any information regarding the position or the movement of the pedestrians was not detailed in the database. Reminding that only single-vehicle and two-vehicle crashes were considered in the analysis, due to the restrictions on the merging process, single-vehicle accidents were 28.7% of the total crashes analyzed.

As illustrated in figure 12, more than 80% of the accidents took place in urban areas. Accordingly, the majority of the collisions concentrated in the Tel Aviv area (31.9%), the Haifa area (11.2%), the Jerusalem area (10.0%) and the Hasharon area (10.3%). While in urban areas crashes were almost equally divided between intersections and sections, outside these urban areas accidents were prevalently happening far from intersections.

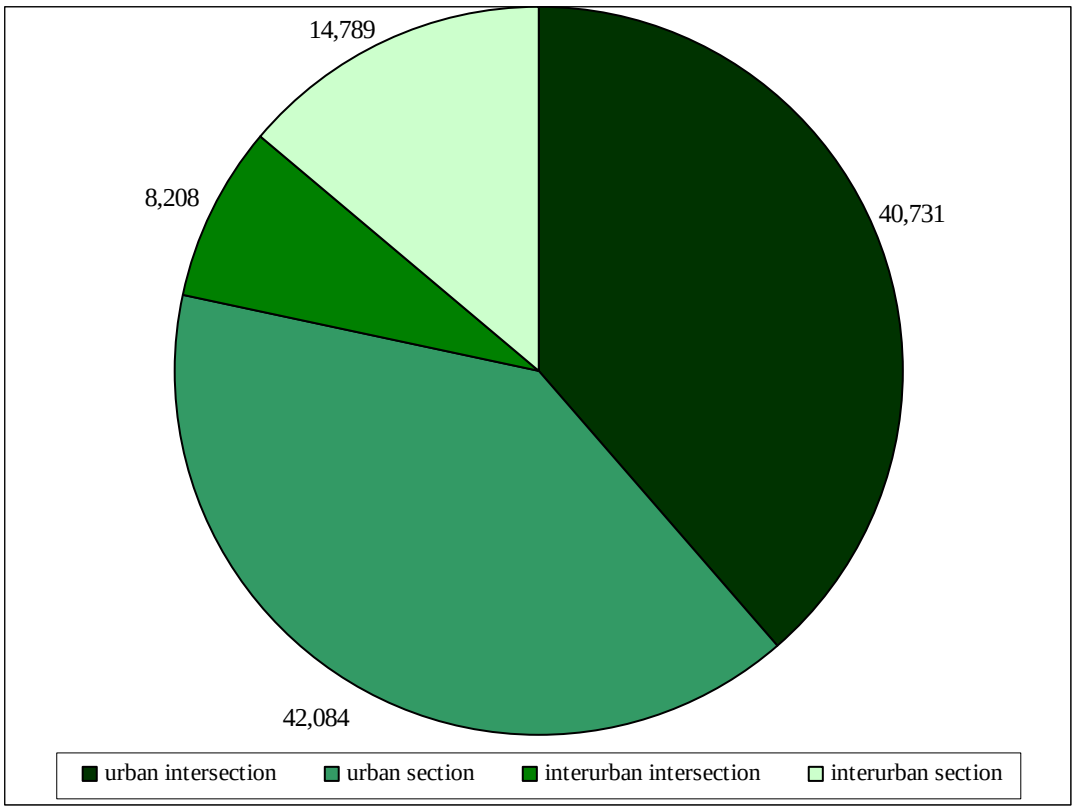

FIGURE 12. Number of accidents per location type

With respect to the other data source, the condition of the infrastructure was not presented, while the condition of the median and the number of ways was not reported for around 80%

of the records. Good weather conditions were described for 92.2% of the records, and accordingly good surface was described in more than 90% of the cases. The cause of the accident was described as fault of the driver in almost 90% of the crashes, and among the possible offenders, around 4% had previous offences for speed violations.

With respect to the vehicles in the accidents, more than 6% of the crashes had a public vehicle implicated, more than 20% had a light commercial vehicle and almost 15% had a motorcycle involved. At least a man was involved in 90% of the accidents, and no woman was implicated in 60% of the crashes, suggesting that men drove more than women and likely appeared more aggressive, with percentages that are absolutely the same seen for the following four year period. As expected from the distribution of the population, over 80% of the drivers were Jews, while a limited number was Moslem and only a few were Christian.

5 Model elaboration

Data mining methods described in section 3 were applied to the data sources illustrated in section 4 to evaluate descriptive and predictive capabilities of data mining techniques in the accident analysis research field. Results focusing on both data sources provide insight into different characteristics of accidents.

Each modeling technique was implemented in Clementine software for data mining: the software works according to streams as the one illustrated in figure 13, where the file containing the data is read, the types for each variable are defined, the input fields for each analysis are selected by filtering the read database and the model options are selected in the model nodes. The program executes the model as a stream from the accident data files to the data mining technique and produces results that can be visualized in tables, analyzed with confusion matrices for predictive purposes and represented with graphical options. Confusion matrices measure the prediction accuracy by comparing predicted versus observed outcomes.

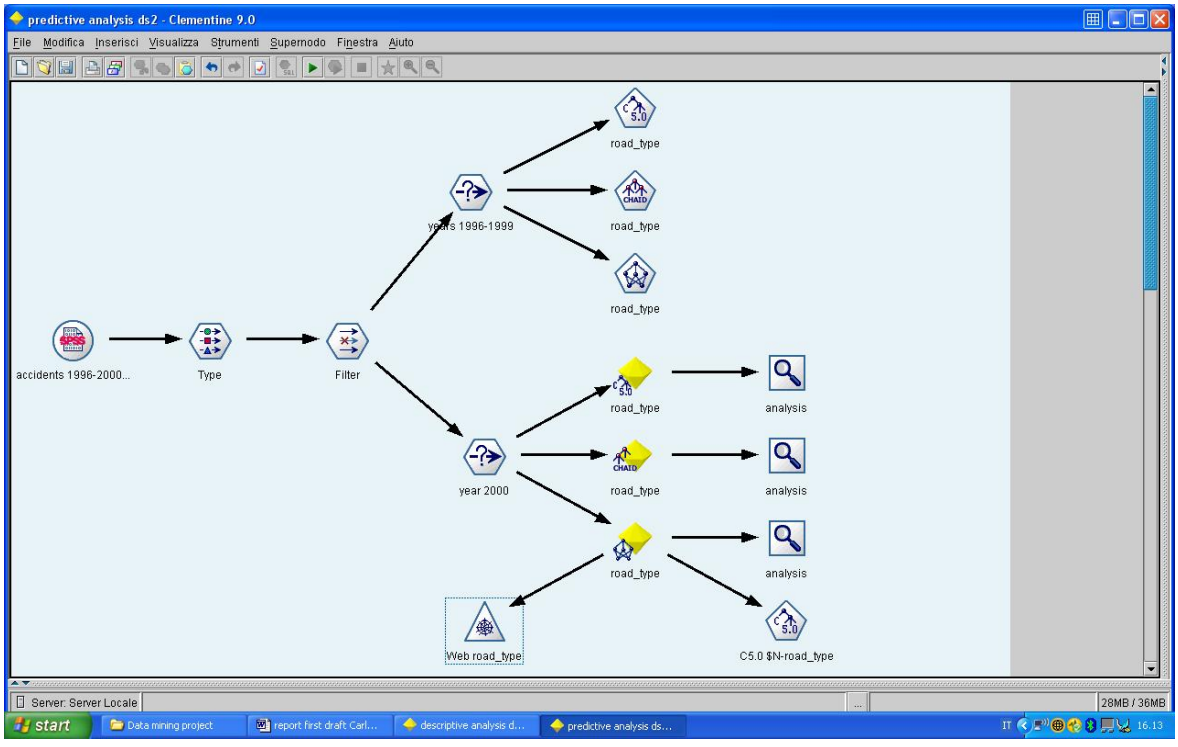

FIGURE 13. Example of Clementine stream

For each technique, databases were elaborated in order to satisfy the data requirements and variables were inserted with some general criteria: for descriptive purposes all the variables in

the data sources were considered in order to find patterns of similarity among the records; categorical variables in which one category was clearly more frequent over the others could not constitute a reliable output variable, as the prediction rates could have been high only because all the models would have predicted always the dominant category, with the exception of the accident severity given the importance of this variable; several variables were excluded from consideration when their link to the predictors or to other variables was too obvious (for example day and hour, rainy weather and wet surface) and when their nature constituted an outcome rather than a predictor (for example the injured persons were a consequence of the accident and could not be considered to predict the outcome of the crash).

5.1 CBS database

As previously stated, the present results elaborated the accidents between 2001 and 2004. Note that for predictive purposes the dataset was divided into a training set, containing the accidents occurred between 2001 and 2003, and a test set, containing the crashes happened in 2004. The nature of the neural network elaboration, that randomly divides the input dataset into training and test sets, considers the database accounting for the collisions taken place in 2004 as the validation set.

The variables considered for descriptive and predictive analysis are illustrated in table 2. Among these variables, four fields were considered as valuable categorical dependent variables: day / night, accident severity, accident location and accident type. The first output variable was considered for illustrative purposes, especially in the phase of comprehending the options of the different data mining techniques and choosing the most suitable methods for the analysis of the accidents.

The three remaining output variables provided valuable information about the classification possibilities of accidents according to their location, type and outcome in terms of injuries. Accident location appeared the most suitable dependent variable, since the records were quite distributed over the categories and since the characteristics of crashes occurring either in urban areas or close to intersections were of interest for the analysis. Accident severity contained one prevalent category, as light injury was the most frequent outcome, but the predictive techniques defined rules for fatal accidents as well. Accident type accounted for several categories, but the predictive methods produced good forecast results despite the high number of possible outcomes.

VARIABLES	CATEGORIES
accident severity	1. fatal accident – 2. severe injury – 3. light injury
type of accident	1. pedestrian – 2. front/side crash – 3. front/rear crash – 4. side/side crash – 5. front/front crash – 6. collision with stopped car – 7. collision with parked car – 8. collision with object – 9. rolling/slipping – 10. fire – 11.other crashes
accident modality	1. entrance of intersection – 2.exit of intersection – 3. parking or gas station – 4. slope – 5. curve – 6. bridge or tunnel – 7. railway crossing – 8. straight road or junction – 9. other
cause of the accident	1. offense of the driver – 2. pedestrian action – 3. passenger behavior – 4. cyclist behavior – 5. car malfunctioning – 6. other
location of the accident	1. urban intersection – 2. urban section – 3. interurban intersection – 4. interurban section
allowed speed	1. 50 km/h – 2. 60 km/h – 3. 70 km/h – 4. 80 km/h – 5. 90 km/h – 6. 100 km/h
day / night	1. day – 2. night
day of the week	1. Sunday – 2. Monday – 3. Tuesday – 4. Wednesday – 5. Thursday 6. Friday – 7. Saturday
season of the accident	1. spring – 2. summer – 3. autumn – 4. winter
weather conditions	1. clear – 2. rainy – 3. hot – 4. foggy – 5. other
number of ways on the road	1. one way – 2. two ways with separation line – 3. two ways without separation line – 4. other
median on the road	1. painted line – 2. safety rail – 3. no safety rail – 4. non built separation – 5. other
shoulders of the road	1. good condition – 2. bad condition – 3. rough road – 4. bad condition and rough road
width of the road	1. up to 5 m. – 2. 5 to 7 m. – 3. 7 to 10.5 m. – 4. 10.5 to 14 m. – 5. over 14 m.
regulation of intersection	1. no control – 2. functioning traffic light – 3. malfunctioning traffic light – 4. blinking yellow – 5. stop sign – 6. right of way sign – 7. other
illumination on the road	1. normal daylight – 2. limited visibility because of the weather – 3. night with lighting – 4. night without lighting – 5. malfunctioning lighting – 6. unknown night conditions
surface conditions of the road	1. dry – 2. wet from water – 3. wet from slippery material – 4. covered with mud – 5. covered with sand – 6. other
location of crossing pedestrians	1. crossing on zebras with traffic light – 2. crossing on zebras without traffic light – 3. crossing out of zebras next to an intersection - 4. crossing out of zebras far from an intersection – 5. not specified crossing position
location of standing pedestrians	1. pedestrian standing on the road - 2. pedestrian standing on the median - 3. pedestrian standing on the sidewalk or shoulders - 4. pedestrian playing on the road - 5. pedestrian in the traffic direction - 6. pedestrian against the traffic direction
type of collision with objects	1. with street signal – 2. with safety rail – 3. with building – 4. with bridge – 5. with light or phone pole – 6. with tree – 7. with other object
distance of colliding objects	1. up to 1 m. – 2. up to 3 m. – 3. object on the road – 4. object on the median – 5. unknown position of the object
vehicles involved	1. one vehicle – 2. two vehicles – 3. three vehicles – 4. fore or more vehicles

TABLE 2. Categorical variables for descriptive and predictive analysis – CBS database

VARIABLES	CATEGORIES
speed offences	1. at least one driver with previous speed violations – 2. no driver with previous speed violations
alcohol or drugs offences	1. at least one driver with previous alcohol or drug violations – 2. no driver with previous alcohol or drug violations
private vehicles	0. no private vehicle involved - 1. one private vehicle involved – 2 two private vehicles involved – 3 three private vehicles involved – 4. four or more private vehicles involved
public vehicles	0. no public vehicle involved - 1. one public vehicle involved – 2 two public vehicles involved – 3 three or more public vehicles involved
light commercial vehicles	0. no light commercial vehicle involved - 1. one light commercial vehicle involved – 2 two light commercial vehicles involved – 3 three or more light commercial vehicles involved
heavy commercial vehicles	0. no heavy commercial vehicle involved - 1. one heavy commercial vehicle involved – 2 two heavy commercial vehicles involved – 3 three or more heavy commercial vehicles involved
motorcycles	0. no motorcycle involved - 1. one motorcycle involved – 2 two motorcycles involved – 3 three or more motorcycles involved
bicycles	0. no bicycle involved - 1. one bicycle involved – 2 two bicycles involved – 3 three or more bicycles involved

TABLE 2. Categorical variables for descriptive and predictive analysis – CBS database (continued)

5.1.1 Cluster analysis

The definition of the considered number of clusters constituted a compromise between a small number, which would have given problems in terms of excessive dimension of the clusters, and a large number, which would have caused difficulties in terms of semantic interpretation of the clusters. Further, different techniques were applied with the same number of clusters, in order to evaluate whether the implementation of different data mining methods had an effect on the results.

K-means clustering was applied by testing solutions with 5, 6 and 7 clusters. Kohonen networks were constructed by experimenting linear maps with 5 and 6 clusters for method comparison, as well as bi-dimensional maps. Unfortunately, these maps did not converge to a clustering solution for every dimension chosen. For example maps 2×4 or 3×4 showed the tendency during the algorithm processing of all the records to appear strongly connected to only part of the clusters because of the impossibility to define groups with the similarity property that characterizes neighboring clusters in the Kohonen structure.

Table 3 illustrates the number of records assigned to each cluster for the three runs of K-means algorithm and the three runs of Kohonen networks. As detailed in the following paragraphs, increasing the number of clusters had a different effect on clusters generated by

K-means or by Kohonen networks. Even though the increase could provide benefit in terms of interpretation, it is visible from the nine clusters solution obtained with a Kohonen 3×3 map that there are three clusters containing less than 3000 records that most likely introduce problems in terms of general description.

NUMBER OF RECORDS	K-MEANS	K-MEANS	K-MEANS	KOHONEN	KOHONEN	KOHONEN
	5 CLUSTERS	6 CLUSTERS	7 CLUSTERS	1×5 MAP	1×6 MAP	3×3 MAP
cluster 1	23510	15695	15471	19650	15822	13226
cluster 2	6806	3736	3696	6364	6884	3362
cluster 3	15441	10402	10349	17573	11873	14370
cluster 4	8515	7837	6383	9946	12249	4880
cluster 5	17784	21954	9616	18523	7243	279
cluster 6		12432	12235		17985	2740
cluster 7			14306			15226
cluster 8						2057
cluster 9						15916

TABLE 3. Number of records per cluster – CBS database

The semantic interpretation of the clusters constitutes a difficult task, and related to the frequency of each category of the input variables in the records belonging to each cluster.

The first solution with 5 groups obtained with the K-means algorithm defined the following clusters:

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 and not in road junctions (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 zebras without traffic lights or crossing not in 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 divided the first cluster into accidents that happen during day and accidents that happen during night. The solution with 7 clusters maintained the groups obtained with 6 clusters and split the third cluster again according to crashes occurring during day versus crashes occurring during night.

The first solution from the Kohonen network with 1×5 map defined 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 not in road junctions at night, both on two-way roads where the allowed speed was 50 km/h and on two-way roads where the allowed speed was over 80 km/h (6,364 cases);

3. accidents that occurred not in road intersections 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, was not verified with the Kohonen networks. The Kohonen

network with 1×6 map actually defined different clusters with respect to the previous 1×5 map:

1. front to side accidents that happened during day in urban intersections, either in 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 or 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 where the median could be constituted by 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 mainly constituted by 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 white line to separate the two ways (17,985 cases).

The solution from the Kohonen network with 3×3 map is far more complicated to interpret, especially since at least four clusters present similar characteristics without a clear cut distinction among one another. The remaining five clusters appear more similar to the first solution rather than to the second, also in accordance to the fact that the initial solution resulted by a 1×5 map.

Note that the results from the Kohonen clusters gave evidence to 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 the non-linear maps did not converge for every dimension, and by the fact that the solutions of this clustering technique appear to be not strictly interrelated among one another, especially in the 3×3 map where the

least significant clusters were positioned in the middle of the map and separated the remaining resulting groups.

Also note that the clusters obtained with different techniques have different characteristics: notably, there is not a group containing all pedestrian accidents in the Kohonen clustering. Given the differences in the algorithms, it should not be surprising that the two clustering technique worked into two different directions: K-means created clusters based on the location and the typology of the accident, while Kohonen networks generated clusters based on the day or night attribute and the location.

Given the fact that Kohonen networks did not work properly for bi-dimensional maps, and that K-means divided the crashes by typology of the accident rather than the more trivial day or night variable, results from the implementation of the K-means algorithm appear more interesting from the accident analysis perspective. Further, Kohonen networks were far more expensive from the computational perspective, with at least half an hour necessary to converge when K-means algorithm processes the same amount of data in less than two minutes.

5.1.2 Decision trees

Decision trees produced hundreds of rules that helped classifying the accidents according to the chosen dependent variables. The following sections detail the most interesting results for each output variable considered in the first phase of the study.

5.1.2.1 Day / night

As previously explained, this first dependent variable was considered for illustrative and testing purposes when examining the data mining techniques to be applied throughout the research.

According to the C5.0 algorithm, the most relevant node for the construction of the tree was the involvement of heavy commercial vehicles in collisions that occurred prevalently during day. According to the CHAID algorithm, the most important node to classify the same records was the season in which the crash actually took place. The similarity with the other technique was evident when considering that the second most significant node for the C5.0 algorithm was also the season, and that the following most relevant node was the type of accident for both algorithms. The predictive ability for C5.0 was equal to 66.5% and for CHAID was

equal to 65.6%, both algorithms missing the majority of the target for the night accidents. Accordingly to these preliminary findings, the two algorithms perform in a similar way with respect to the first dependent variable.

Some rules are illustrated in tables 4 and 5 for both algorithms, after a selection that resulted from pruning the tree, considering the highest confidence levels and not accounting for the most trivial findings if not for illustrative purposes.

IF an accident occurs with no heavy commercial vehicle involved, AND the accident happens in the spring, AND the accident occurs with at least one pedestrian involved THEN the likelihood of the accident occurring during day is 77.4%
IF an accident occurs with no heavy commercial vehicle involved, AND the accident happens in the summer, AND the allowed speed is 50 km/h, AND no driver involved committed a previous speed violation, AND the accident occurs with at least one pedestrian involved THEN the likelihood of the accident occurring during day is 74.7%
IF an accident occurs with no heavy commercial vehicle involved, AND the accident happens in the spring, AND the accident is a front to side collision, THEN the likelihood of the accident occurring during day is 71.7%
IF an accident occurs with no heavy commercial vehicle involved, AND an accident occurs with at least one light commercial vehicle involved, AND the accident happens in the winter, AND the road surface is wet, AND the accident is a front to back collision THEN the likelihood of the accident occurring during day is 70.4%
IF an accident occurs with no heavy commercial vehicle involved, AND the accident happens in the winter, AND the accident happens with clear weather, AND the allowed speed is 50 km/h, AND the accident occurs with at least one pedestrian involved THEN the likelihood of the accident occurring during day is 67.9%
IF an accident occurs with no heavy commercial vehicle involved, AND the accident happens in the winter, AND the accident is a collision with an object THEN the likelihood of the accident occurring during night is 69.3%
IF an accident occurs with no heavy commercial vehicle involved, AND the accident occurs with at least one private vehicle involved, AND the accident happens in the winter, AND the accident is a car rolling THEN the likelihood of the accident occurring during night is 68.3%
IF an accident occurs with no heavy commercial vehicle involved, AND the accident happens in the winter, AND the accident is a collision with a parked vehicle THEN the likelihood of the accident occurring during night is 67.6%
IF an accident occurs with no heavy commercial vehicle involved, AND the accident happens in the winter, AND the accident happens in rainy weather, AND the accident occurs with at least one pedestrian involved THEN the likelihood of the accident occurring during night is 64.2%

TABLE 4. Rules for C5.0 tree with CBS data - day / night

Given the prevalence of day accidents over night accidents, it was not surprising noticing the highest number of rules with a higher confidence level for crashes taking place during day. On one hand, the characteristic of the CHAID algorithm to merge the categories was evident from the results and helped increasing the level of confidence of the rules. On the other hand, the C5.0 algorithm appeared easier to interpret because of the clear-cut definition of the rules.

IF the accident happens in the spring, AND the accident is either a front to side, a front to back or a front to front collision, AND the traffic light is not functioning or there is a either a stop or a right of way sign, AND the accident occurs with at least one private vehicle involved, AND the accident occurs with at least one woman involved THEN the likelihood of the accident occurring during day is 83.7%
IF the accident happens in the summer, AND the accident is either a front to side or front to back collision or at least a pedestrian is involved, AND the accident occurs with at least two women involved THEN the likelihood of the accident occurring during day is 83.6%
IF the accident happens in the spring, AND the accident is either a front to side or a front to back collision, AND the accident occurs without any private vehicle involved, AND the accident occurs with no woman involved THEN the likelihood of the accident occurring during day is 78.5%
IF the accident happens in the spring, AND the accident is either a front to side or a front to back collision, AND the accident occurs with one private vehicle involved, AND the accident occurs with no woman involved THEN the likelihood of the accident occurring during day is 72.8%
IF the accident happens in the winter, AND the accident is either a front to side or a front to front collision, AND the traffic light is not functioning or there is a either a stop or a right of way sign, AND the accident occurs with one private vehicle involved, AND the accident occurs with at least one woman involved THEN the likelihood of the accident occurring during day is 72.8%
IF the accident happens in the autumn, AND the accident is either a back to side or a side to side collision or at least a pedestrian is involved, AND the accident occurs with at least a woman involved THEN the likelihood of the accident occurring during day is 68.4%
IF the accident happens in the winter, AND the accident is either a collision with an object or with a parked vehicle THEN the likelihood of the accident occurring during night is 68.6%
IF the accident happens in the autumn, AND the accident is either a collision with an object or with an animal THEN the likelihood of the accident occurring during night is 64.6%

TABLE 5. Rules for CHAID tree with CBS data - day / night

The most interesting rules for classification of daily accidents concern the hit of pedestrians in winter with bad weather, the presence of light commercial vehicles in front to back collisions and the possibility of the influence of the rain in autumn and winter. The most relevant rules to classify night crashes concern the collision with objects, parked vehicles or pedestrians. For the former accidents, the inadequate distance maintained between vehicles could explain part

of the problem. For the latter accidents, the actual lack of illumination on the streets at night could explain the issue, even though the decision tree did not determine if the limited visibility is responsible for the accident.

5.1.2.2 Accident severity

Considering as categorical dependent variable the severity of the accident, the decision trees from the implementation of the two applied algorithms resulted extremely different. From the C5.0 technique, the initial nodes were the number of vehicles implicated, the presence of a bicycle or a commercial vehicle in the collision, the existence of previous speed violations by at least one of the drivers and the involvement of a pedestrian. From the CHAID technique, the initial nodes were the type of accident, the location of the crash and the presence of private vehicles. The predictive ability for C5.0 was equal to 87.2% and for CHAID was equal to 86.6%, but significantly only the C5.0 algorithm was able to predict fatal accidents and consequently performed better than the CHAID algorithm.

Some of the rules for fatal and severe accidents are summarized in table 6 and table 7 for both algorithms, and these tables emphasize that the C5.0 tree performed better the classification of the accidents according to the severity of the collision. In fact, no rule for fatal accidents and only one rule for crashes resulting in severe injuries were produced by the CHAID method.

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 resulting 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 resulting fatal is 75.0%
IF the accident occurs with only one vehicle involved, AND the involved vehicle is a public vehicle, AND the allowed speed is 100 km/h, THEN the likelihood of the accident resulting fatal is 64.4%
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, AND the accident happens in clear weather THEN the likelihood of the accident resulting in a severe injury is 70.8%

TABLE 6. Rules for C5.0 tree with CBS data - accident severity

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 and crossing the road, AND the only light in the location of the accident is the natural night light, AND the accident happens in the autumn THEN the likelihood of the accident resulting in a severe injury is 65.1%
IF the accident occurs with only one vehicle involved, AND the involved vehicle is a commercial vehicle, AND the accident occurs with at least one pedestrian involved, AND the allowed speed is 50 km/h, AND the accident occurs in an evening before the holidays THEN the likelihood of the accident resulting in a severe injury is 64.2%
IF the accident occurs with more than one vehicle involved, AND the accident occurs with at least one bicycle involved, AND the accident occurs with at least one commercial vehicle involved, AND the accident takes place inside an urban area and not in an intersection THEN the likelihood of the accident resulting in a severe injury is 63.6%

TABLE 6. Rules for C5.0 tree with CBS data - accident severity (continued)

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%

TABLE 7. Rules for CHAID tree with CBS data - accident severity

Fatal accidents mainly involved pedestrians that crossed roads where the allowed speed was high or the width was broad, most likely a highway or one of the major roads of the country. This phenomenon is frequently seen in Israel, especially where villages are extremely close to these arterials. Further, the involvement of bicycles and commercial vehicles increased 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 pointed out that single-vehicle crashes produced the most severe outcomes in terms of injuries.

5.1.2.3 Accident location

Considering the location of the accident as the dependent variable, the generated decision trees were not similar, but also not totally different. For the C5.0 tree, the most significant variables consisted of the allowed speed, followed by the typology of the accident, the existence of previous speed violations by at least one of the involved drivers and the number of vehicles implicated in the crash. For the CHAID algorithm, the most relevant fields consisted of the regulation of the intersections, followed by the allowed speed, the number and the type of vehicles involved and the condition of the median. The prediction accuracy

was also significantly different, as for the C5.0 tree was equal to 67.1% and for the CHAID tree was equal to 87.8%.

The confusion matrices in table 8 and table 9 provide insight into the different predictive performances between the two algorithms. Note that the rows represent the actual observations and the columns the predicted values, consequently the elements in the diagonal constitute the correct predictions for accidents occurred in 2004, based on the models estimated with the accidents taken place between 2001 and 2003.

	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 8. Confusion matrix for C5.0 tree with CBS data - accident location

	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

TABLE 9. Confusion matrix for CHAID tree with CBS data - accident location

The C5.0 algorithm performed extremely well when predicting crashes that took place outside an urban area, but confused the collisions inside urban areas as the majority of the incorrect forecasts consisted of urban accidents that occurred in road junctions and were predicted to take place in a section and vice versa. The CHAID algorithm did not present the same problem, and the only errors consisted of an excess in predicting accidents in urban areas.

Some of the rules with higher confidence level are presented in tables 10 and 11. The classification of the accidents with respect to the location provided evidence to some aspects. Accidents in interurban intersections occurred mainly in conditions of limited visibility, for example at night with only the natural night light available, and when the median was not constructed. This leads to think about obvious problems related to excessive speed. This concept was confirmed by the rules regarding crashes in interurban sections, for example with single-vehicle accidents where a car was getting off the road. In urban areas, pedestrians, cyclists and motorcyclists were often involved: in this case supposedly sometimes their behavior caused the crashes that resulted immediately in severe consequences.

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 two vehicles involved, AND the accident is a front to side collision, AND the light in the location of the accident is the natural daily light, AND the median is not constituted by a safety rail THEN the likelihood of the accident occurring in an interurban intersection is 66.8%
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 80 km/h, AND the accident occurs with more than one vehicle involved, AND the accident is a front to front collision, THEN the likelihood of the accident occurring in an interurban section is 75.6%

TABLE 10. Rules for C5.0 tree with CBS data - accident location

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 is a front to side collision, AND the road signs in the location of the accident are in poor conditions, AND the accident results in light injuries, AND the width of the road where the accident occurs is between 5 and 7 m. THEN the likelihood of the accident occurring in an urban intersection is 71.4%
IF the speed allowed is 50 km/h, AND the accident occurs with at least one pedestrian involved and crossing the road, AND the road signs in the location of the accident are in good conditions, AND the accident results in severe injuries, AND the cause of the accident is the fault of the driver, AND the width of the road where the accident occurs is between 7 and 10.5 m. THEN the likelihood of the accident occurring in an urban intersection is 64.4%
IF the speed allowed is 50 km/h, AND the accident is a front to side collision, AND at least one of the drivers involved has a previous speed violation, AND the accident occurs with at least one motorcycle involved, AND the road signs in the location of the accident are in good conditions THEN the likelihood of the accident occurring in an urban intersection is 63.5%
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%
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, AND the road shoulders in the location of the accident are in poor conditions THEN the likelihood of the accident occurring in an urban section is 76.7%

TABLE 10. Rules for C5.0 tree with CBS data - accident location (continued)

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%

TABLE 11. Rules for CHAID tree with CBS data - accident location

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%

TABLE 11. Rules for CHAID tree with CBS data - accident location (continued)

Interestingly, with different dependent variables the two techniques showed different predictive capabilities: with accident location the CHAID tree outperformed the C5.0, while exactly the opposite happened with accident severity.

5.1.2.4 Accident type

Using accident type as output field could have raised questions about the predictive ability of the applied algorithms, as the high number of categories could affect forecast precision. Both algorithms considered the number of vehicle involved as the most significant variable, but while for the C5.0 tree the following relevant fields were the presence of pedestrians, bicycles and objects, for the CHAID tree were not only on the presence of pedestrians, but also the location of the accident and the conditions of both traffic lights and medians. The prediction accuracy for the C5.0 tree was equal to 76.2% and for the CHAID tree was equal to 73.0%, satisfying results when considering the number of categories. In particular, good predictive ability was shown for front to side, front to back and object collisions, as well as crashes with pedestrians involved.

Some of the rules for accident type predictions are summarized in table 12 and table 13 for some of the most frequent crashes. Some similarities are retrievable from the tables, only the confidence level of the C5.0 tree is higher and explains the better prediction accuracy. Not surprisingly, the absence of traffic lights sees the majority of the front/side collision for the likely not respect of the road signs, the collisions with pedestrians involve single vehicles in almost all the cases, and previous speed violations appear related with colliding with an object. This dependent variable does not establish which method appears more suitable than

the other, and adds to the previous findings in which each technique alternatively performed better than the alternative algorithm.

IF the accident occurs with only one vehicle involved, AND the accident takes place inside an urban area and not in an intersection, AND at least one of the drivers involved has a previous speed violation THEN the likelihood of the accident being a collision with an object is 76.4%
IF the accident occurs with more than two and up to four vehicles involved, AND the accident takes place inside an urban area and in an intersection, AND the traffic light in the location of the accident is functioning properly, AND the width of the road where the accident occurs is up to 5 m., AND the accident results in light injuries THEN the likelihood of the accident being a front/back collision is 72.6%
IF the accident occurs with more than two and up to four vehicles involved, AND the accident takes place inside an urban area and in an intersection, AND the road junction is regulated according to a “stop” sign, AND the accident results in light injuries THEN the likelihood of the accident being a front/side collision is 92.4%
IF the accident occurs with more than two and up to four vehicles involved, AND the accident takes place inside an urban area and in an intersection, AND the road junction is regulated according to a “right of way” sign THEN the likelihood of the accident being a front/side collision is 83.4%
IF the accident occurs with more than two vehicles involved, AND the accident occurs with only two private vehicles involved, AND the width of the road where the accident occurs is between 5 and 7 m., AND the allowed speed is 50 km/h THEN the likelihood of the accident being a front/side collision is 75.8%
IF the accident occurs with only one vehicle involved, AND the accident occurs with one pedestrian involved and crossing the road, AND the accident occurs with one private vehicle involved THEN the likelihood of the accident being a collision with a pedestrian is 99.9%

TABLE 12. Rules for C5.0 tree with CBS data - accident type

IF the accident occurs with only one vehicle involved, AND the allowed speed is 50 km/h, AND at least one of the drivers involved has a previous speed violation THEN the likelihood of the accident being a collision with an object is 70.6%
IF the accident occurs with more than two vehicles involved, AND the accident takes place inside an urban area and not in an intersection, AND the road where the accident occurs is one way THEN the likelihood of the accident being a front/back collision is 69.2%
IF the accident occurs with two vehicles involved, AND the accident takes place outside an urban area and in an intersection, AND the road junction is regulated according to a “stop” sign THEN the likelihood of the accident being a front/side collision is 94.4%
IF the accident occurs with two vehicles involved, AND the accident takes place outside an urban area and in an intersection, AND the traffic light in the location of the accident is either blinking or not working, AND the accident happens in the afternoon or in the evening THEN the likelihood of the accident being a front to side collision is 91.3%

TABLE 13. Rules for CHAID tree with CBS data - accident type

IF the accident occurs with two vehicles involved, AND the accident takes place inside an urban area and in an intersection, AND the road junction is regulated according to a “right of way” sign THEN the likelihood of the accident being a front/side collision is 89.7%
IF the accident occurs with two vehicles involved, AND the accident takes place inside an urban area and in an intersection, AND the road traffic light in the location of the accident is functioning properly, AND the accident occurs with one motorcycle involved THEN the likelihood of the accident being a front/side collision is 78.1%
IF the accident occurs with only one vehicle involved, AND the accident occurs with one pedestrian involved and crossing the road, AND the accident occurs with one private vehicle involved THEN the likelihood of the accident being a pedestrian collision is 99.9%

TABLE 13. Rules for CHAID tree with CBS data - accident type (continued)

5.1.3 Neural networks

Neural networks generated connections between input variables and output predictors, and processed the same database used for the construction of the decision trees. As neural networks are more complex to interpret with respect to decision trees, the relative importance of the input variables helps understanding the network model.

5.1.3.1 Day / night

The constructed neural network determined that the type of accident, the existence of previous alcohol or drug violations by at least one driver involvement of pedestrians, the type of control in the intersections as well as the type of vehicles involved were the most relevant input variables to predict whether the accidents happened during day or night. The network generated two hidden layers between the input and output layers according to the exhaustive prune algorithm. The estimated precision during the training and test phase reached 65.8%, and this value was confirmed by the value of the prediction accuracy equal to 66.4% when the estimated model was applied to crashes who took place in the year 2004, with an excess of accidents that actually occurred at day that were assigned at night.

Note that these results were different from the results obtained with the decision tree algorithms, where the season in which the collision happened and the type of accident were the most relevant fields. With this respect, table 14 orders the most relevant input variables for the estimated model, with the importance of the input variables varying between 0 (no relevance) and 1 (complete explanation).

VARIABLES	RELATIVE IMPORTANCE
type of accident	0.2679
alcohol or drugs offences	0.2089
regulation of intersection	0.2009
season of the accident	0.1733
private vehicles	0.1589
heavy commercial vehicles	0.1506
light commercial vehicles	0.1248
motorcycles	0.1133
other variables	< 0.1000

TABLE 14. Relevant input variables for MLP network with CBS data - day / night

Since this illustrative example, it appears clear that the neural network produced similar results in terms of predictive ability, and the cost in terms of computational time is sensibly higher, as the convergence of the network took around two hours while the convergence of a decision tree takes less than two minutes. Further, the interpretation of the neural network seems far more complicated than the interpretation of the decision trees, as the graphical representation of figures 14 and 15 exemplifies.

Figure 14 illustrates the circular representation of the neural network, in which input and output variables are collocated in a circle and the connections from the input to the output variables are represented with various degree of thickness directly proportional to the strength of the connections.

Figure 15 presents the reticular representation of the neural network, in which input variables are collocated in the design space according to their distance from the output variables. The connections are thicker when they are stronger, and if possible the interpretation of the results appears even more difficult than with the circular representation.

This problem of the interpretation of the results in neural network implementation is opposed to the relative easiness of interpretation of the rules generated with the decision trees, giving the edge to the latter techniques for predictive purposes.

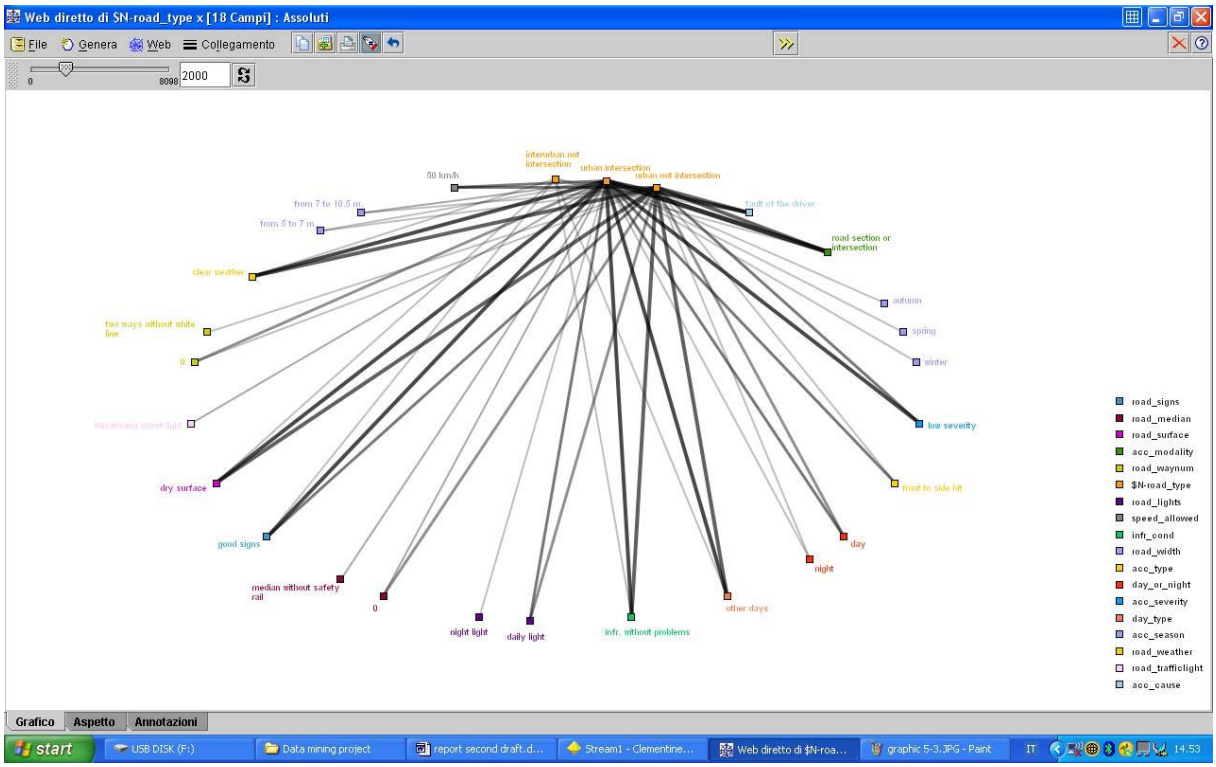

FIGURE 14. Example of neural network with circular representation

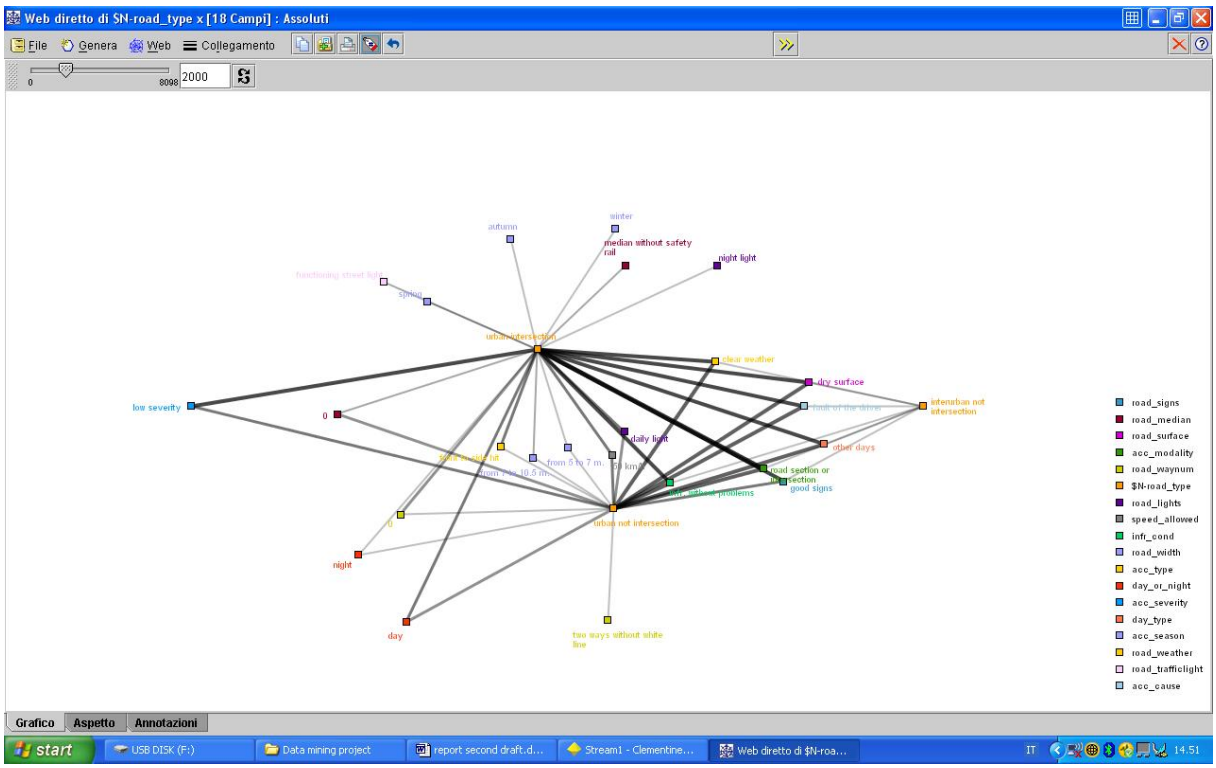

FIGURE 15. Example of neural network with reticular representation

5.1.3.2 Accident severity

The MLP network estimated that the outcome of the accident in terms of severity was explained mainly by the type of accident and the involvement of bicycles, motorcycles, commercial vehicles. The network constructed two hidden layers and reached 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.

This high predictive ability does not actually indicate that the model is good, since the actual confusion matrix indicates that the model predicts almost all accidents as resulting in light injuries. This means that the previous model that employed the predictor day / night was a better model, despite the 20% incorrect forecasts in excess with respect to this one. This problem, encountered also with the CHAID algorithm but at a lesser extent, remarks that the evaluation of the goodness of the model is not only related to the prediction accuracy, but also to the actual interpretability of the model itself.

Table 15 details the most relevant variables to classify whether accidents resulted in fatalities or severe or light injuries.

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

TABLE 15. Relevant input variables for MLP network with CBS data - accident severity

Again, the significant higher computational cost did not produce any advantage in terms of model fit, providing more insight into the evaluation of the decision trees as better methods than neural networks.

5.1.3.3 Accident location

When analyzing the location of the accident, two input variables were largely more significant than any other to explain where the accident occurred: the type of control of the intersections and the allowed speed. Table 16 details the most relevant input variables in predicting the location of the accidents.

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

TABLE 16. Relevant input variables for MLP network with CBS data - accident location

This is the first dependent variable to produce results similar to one of the decision tree algorithms, precisely the one constructed by CHAID method. The estimated precision in training was equal to 81.1%, and the prediction accuracy reached 81.4% when the forecasted results were compared to the observed locations for crashes taken place in the year 2004.

The confusion matrix in tables 17 provides further evidence of the problem described in the previous section: the high predictive ability does not reflect the goodness of the model, as any accident was predicted to be located in an intersection outside an urban area. The first column emphasizes that the model did not predict any accident to happen in an intersection outside an urban area, and almost all the actual crashes that occurred in similar locations were forecasted to intersection, but in urban areas. The remaining part of the matrix shows great accuracy in the predictions.

Note that the inaccuracy is somewhat different with respect to the previous case. The accident severity was probably influenced by the fact that one category, crashes that resulted in light injuries, is dominant over the others. 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.

	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

TABLE 17. Confusion matrix for MLP network with CBS data - accident location

5.1.3.4 Accident type

The last dependent variable analyzed with the neural networks is the type of accident. The most relevant input fields for this categorical variable with numerous categories were the involvement of pedestrians, either crossing the road or simply standing close to the road, the number and the type of vehicles involved as well as the existence of previous speed or alcohol violations by at least one driver. Table 18 summarizes the relative importance of the input fields.

VARIABLES	RELATIVE IMPORTANCE
location of crossing pedestrians	0.1094
number of vehicles involved	0.1054
location of standing pedestrians	0.0890
private vehicles	0.0617
speed offences	0.0385
alcohol or drugs offences	0.0298
public vehicles	0.0291
light commercial vehicles	0.0279
other variables	< 0.0250

TABLE 18. Relevant input variables for MLP network with CBS data - accident type

The estimated precision during the training and test phase was equal to 74.0% and the prediction accuracy with respect to collisions that happened in the year 2004 was 74.5%. As for the decision trees, the neural network model predicted correctly in particular front/side and front/back collisions, as well as collisions with objects and pedestrians. This is actually the first variable that the neural network forecasted with accuracy, even though at a slightly lower level than decision trees and at an extremely higher computational cost.

5.1.4 Association rules

Association rules were applied with a different approach with respect to the other data mining techniques, at least with respect to the data perspective. The analysis focused on black spots, which are defined as specific locations in the road network where the frequency of fatalities results higher than the expected average.

For this reason the list of the black spots defined for the Israeli road network was matched to the accident data in order to create an additional field. This “black spot” variable was actually a binary field, where 1 indicated that the accident occurred in a black spot and 0 otherwise, and was assumed to be the categorical predictor for the analysis. Given the definition and the construction of the black spot data and the strict correlation with the severity of the accidents, not surprisingly the results were not enlightening as fatal accidents resulted to happen in black spots.

Given the difficulties of working with this variable, the database was divided into two parts: accidents occurred in black spots and accidents taken place in other network sections. The severity of the accidents was considered as categorical dependent variable and an association rule algorithm with very restrictive parameters, described in the methodological section as a confidence rule equal to 90% and a support equal to 5%, was tested. The Apriori algorithm produced around 5,000 rules for both parts of the database, and interestingly the lift of these rules was highly comparable and for every rule was around one.

This meant that any of the rules was clearly discernible among the thousands generated, and that any attempt to synthesize the results would have been based purely on personal criteria, rather than objective. Considering that the literature exhibited the very same problem of selecting rules among the thousands, the existing papers in the specific subject did not enlighten with this respect, the association rules algorithm were discarded under the

motivation that the results were not interpretable and consequently the method was not suitable for further analysis.

5.2 Enlarged database

As previously stated, the present analysis elaborated the accidents between 1996 and 2000. The variables considered for descriptive and predictive analysis are illustrated in table 19, and include some information retrieved from the census about the drivers involved in the accident. The amount of missing data and the necessity of understanding the relative characteristics of drivers and vehicles involved suggested the creation of new variables according to the principles described in section 4.2.

Note that for predictive purposes the dataset was divided into a training set, containing the accidents occurred between 1996 and 1999, and a test set, including the crashes happened in 2000. The nature of the neural network elaboration, that randomly divides the input dataset into training and test sets, considers the database accounting for the collisions taken place in 2000 as the validation set.

Among the variables considered, three dependent variables were considered suitable: accident location, accident severity and accident type. The day / night variable was not considered in this phase, given the illustrative and testing purpose of the model estimation with this predictor in the first phase of the research.

VARIABLES	CATEGORIES
accident severity	1. fatal accident – 2. severe injury – 3. light injury
accident type	1. pedestrian – 2. front/side crash – 3. front/rear crash – 4. side/side crash – 5. front/front crash – 6. collision with stopped or parked car – 7. collision with object – 8. rolling/slipping – 9. other crashes
accident modality	1. entrance of intersection – 2.exit of intersection – 3. parking or gas station – 4. slope – 5. curve – 6. bridge or tunnel – 7. railway crossing – 8. straight road or junction – 9. other
accident cause	1. offense of the driver – 2. pedestrian action – 3. passenger behavior – 4. cyclist behavior – 5. car malfunctioning – 6. other
accident location	1. urban intersection – 2. urban section – 3. interurban intersection – 4. interurban section
allowed speed	1. 50 km/h – 2. 60 km/h – 3. 70 km/h – 4. 80 km/h – 5. 90 km/h – 6. 100 km/h
day / night	1. day – 2. night
type of day	1. Sunday to Thursday - 2. Friday, Saturday and holidays

TABLE 19. Categorical variables for descriptive and predictive analysis – Enlarged database

season of the accident	1. spring – 2. summer – 3. autumn – 4. winter
weather conditions	1. clear – 2. rainy – 3. hot – 4. foggy – 5. other
number of ways on the road	1. one way – 2. two ways with separation line – 3. two ways without separation line – 4. other
median on the road	1. painted line – 2. safety rail – 3. no safety rail – 4. non built separation – 5. other
shoulders of the road	1. good condition – 2. bad condition – 3. rough road – 4. bad condition and rough road
width of the road	1. up to 5 m. – 2. 5 to 7 m. – 3. 7 to 10.5 m. – 4. 10.5 to 14 m. – 5. over 14 m.
signals on the road	1. no control – 2. functioning traffic light – 3. malfunctioning traffic light – 4. blinking yellow – 5. stop sign – 6. right of way sign – 7. other
lights on the road	1. normal daylight – 2. limited visibility because of the weather – 3. night with lighting – 4. night without lighting – 5. malfunctioning lighting – 6. unknown night conditions
surface conditions of the road	1. dry – 2. wet from water – 3. wet from slippery material – 4. covered with mud – 5. covered with sand – 6. other
type of collision with objects	1. with street signal – 2. with safety rail – 3. with building – 4. with bridge – 5. with light or phone pole – 6. with tree – 7. with other object
distance of colliding objects	1. up to 1 m. – 2. up to 3 m. – 3. object on the road – 4. object on the median – 5. unknown position of the object
Jewish drivers	1. at least a Jewish driver involved – 2. no Jewish driver involved
Moslem drivers	1. at least a Moslem driver involved – 2. no Moslem driver involved
Christian drivers	1. at least a Christian driver involved – 2. no Christian driver involved
drivers of the same religion	1. same religion – 2. different religion
gender of the drivers	1. two male drivers – 2. male and female drivers – 3. female drivers
origin of the drivers	1. two drivers born in Israel – 2. one driver born in Israel and one driver born abroad – 3. two drivers born abroad
age of the drivers	1. same generation – 2. one generation of difference – 3. two generations of difference – 4. three or more generations of difference
experience of the drivers	1. at least one inexperienced driver – 2. two averagely experienced drivers – 3. at least one very experienced driver
speed offences	1. at least one driver with previous speed violations – 2. no driver with previous speed violations
private vehicles	0. no private vehicle involved - 1. one private vehicle involved – 2 two private vehicles involved
public vehicles	0. no public vehicle involved - 1. one public vehicle involved – 2 two public vehicles involved
light commercial vehicles	0. no light commercial vehicle involved - 1. one light commercial vehicle involved – 2 two light commercial vehicles involved
heavy commercial vehicles	0. no heavy commercial vehicle involved - 1. one heavy commercial vehicle involved – 2 two heavy commercial vehicles involved
motorcycles	0. no motorcycle involved - 1. one motorcycle involved – 2 two motorcycles involved
bicycles	0. no bicycle involved - 1. one bicycle involved – 2 two bicycles involved

TABLE 19. Categorical variables for descriptive and predictive analysis – Enlarged database (continued)

Note that drivers were considered belonging to the same generation if their age difference was less than 15 years, with two or more generations defined accordingly to multiple of this value. Drivers were considered inexperienced if they obtained their license in the three years prior to the accident and very experienced if driving for more than 15 years.

5.2.1 Clustering

The determination of the number of clusters followed the same logic explained for the other data source, and resulted from the elaboration of the previous data source supported this choice. Accordingly, K-means clustering was applied by generating solutions with 5, 6 and 7 clusters, and Kohonen networks were constructed by experimenting linear maps with 5 and 6 clusters for method comparison, as well as additional bi-dimensional maps. The non-linear maps exhibited again problems of convergence, related to the incapacity the algorithm of creating clusters satisfying the necessary property of adjacent clusters to share characteristics.

Table 20 shows the number of cases assigned to each cluster for each K-means run and each Kohonen network map. From the perspective of the number of records included in each cluster, the two methods provided sensibly different results.

NUMBER OF RECORDS	K-MEANS	K-MEANS	K-MEANS	KOHONEN	KOHONEN	KOHONEN
	5 CLUSTERS	6 CLUSTERS	7 CLUSTERS	1×5 MAP	1×6 MAP	3×3 MAP
cluster 1	31300	31273	31255	43797	29963	19165
cluster 2	18142	11630	11605	926	13858	2978
cluster 3	11666	19563	7342	25160	6902	30599
cluster 4	18657	18413	18117	5049	20181	2362
cluster 5	26047	11371	11365	30880	4170	1046
cluster 6		13562	12939		30738	2262
cluster 7			13189			29050
cluster 8						236
cluster 9						18114

TABLE 20. Number of records per cluster – Enlarged database

The Kohonen networks concentrated the majority of the records in part of the clusters, even in the solution with 5 clusters one does not contain even 1000 cases, and this tendency is even more evident in the 3×3 map, where only four clusters include more than 90% of the records.

The K-means algorithm distributed the cases homogeneously among the clusters, and exhibited the property already discussed of dividing bigger clusters in smaller parts when their number was increased.

The semantic definition of the clusters constituted again a difficult task and relied on the frequency analysis of the categories for the input variables.

The first solution with 5 clusters generated by the K-means algorithm defined the following groups:

1. accidents that involved pedestrians, mostly in urban areas and not in intersections, on narrow roads without safety rail (31,300 cases);

2. accidents between two private vehicles, mostly in urban areas and in intersections regulated either with traffic light, right of way or stop sign (18,142 cases);

3. accidents that involved a motorcycle, mostly in urban areas between drivers with similar experience and similar age (11,666 cases);

4. accidents between a private vehicle and either a light commercial vehicle, a heavy commercial vehicle, a motorcycle, mostly in urban areas and in narrow roads (18,657 cases);

5. front to side accidents that happened in urban areas, mostly during day in junctions between drivers with similar age, same religion and same experience (26,047 cases).

The solution with 6 clusters divided the fifth cluster into accidents that occurred during day and accidents that occurred during night. In addition to this division, the solution with 7 clusters split the second group according to the same principle, collisions happening during day versus collisions happening during night.

The first solution originated from the Kohonen network with a linear map 1×5 identified the following clusters:

1. front to side accidents that happened in urban areas, mostly in intersections in narrow roads between drivers of similar age and same religion (43,797 cases);

2. front to side accidents that occurred during day, mostly in intersections either inside or outside urban areas (926 cases);

3. accidents that took place in urban areas, between two private vehicles or a private vehicle and light commercial vehicles, with drivers of different religions (25,160);

4. accidents that happened not in intersections between a private vehicle and either a public vehicle, a light commercial vehicle or a motorcycle (5,049 cases);

5. accidents that involved pedestrians, mostly in urban areas and in narrow roads without safety rail, and between Jewish drivers one of which was born in Israel and the other was born abroad (30,738 cases).

The second solution generated from the Kohonen network with a linear map 1×6 exhibited the same property shown by the K-means algorithm, as the first group of the five-clusters solution was split into two groups, the first containing the accidents happening at day and the second including the crashes occurring at night.

The solution from the Kohonen network with 3×3 map was far more complicated to interpret, especially since at least four clusters present similar characteristics without a clear cut distinction among one another and a fifth group contains only a couple hundred cases. The remaining four clusters were similar to the groups in the initial solution from the 1×5 map, excluding the cluster constituted by accidents that involved pedestrians.

Note that the results from the Kohonen networks exhibited the same property already faced, consisting in the fact that the importance of the vicinity property caused non-linear maps not to converge because similarities among adjacent clusters were not found and all the cases were linked to a restricted number of clusters. Again, in the 3×3 map the least significant clusters were positioned in the middle of the map and separated the remaining resulting groups.

Note also that the clusters from different techniques have different characteristics, but this difference appears less accentuated with the analysis of this data source, as for example both methods defined a cluster with pedestrian involved in the collisions. K-means also

distinguished a cluster with motorcycles implicated in the crash, and in general exhibited the characteristic that the typology of the vehicle played a significant role in generating the groups, as well as the typology of collision. Again, given also the difficulties of Kohonen networks to converge with bi-dimensional map configuration, K-means algorithm appeared more appropriate from the accident analysis perspective.

5.2.2 Decision trees

Decision trees generated hundreds of rules that allowed classifying the accidents according to the selected predictors. The following section details the most relevant results for each dependent variable considered in this second phase of the study.

5.2.2.1 Accident severity

The implemented decision trees produced different results when the categorical output variable was the severity of the accidents. Interestingly, the C5.0 tree reached only three levels of depth and determined as the most significant input variables the cause, the typology and the location of the crash. The CHAID tree was much deeper and the most relevant fields were the typology and the location of the collision, as well as the illumination on the road and the type of vehicles involved. The prediction accuracy for C5.0 and CHAID was equal to 85.6% and both techniques failed to forecast fatal accidents when the predicted values were compared to the actual locations for collisions happened in the year 2000.

For this reason the only rules for crashes resulting in severe injuries are summarized in tables 21 and 22. Note that not only two rules were produced for C5.0 method and one rule was generated for CHAID technique, but also the likelihood of accidents resulting in severe injuries were below 50%, to emphasize the problems regarding the analysis of a categorical predictor in which one category is dominant over the others.

IF the cause of the accident is the action of a pedestrian, AND the accident occurs with at least one pedestrian involved, AND the accident occurs outside an urban area and not in an intersection THEN the likelihood of the accident resulting in a severe injury is 49.5%
IF the cause of the accident is the action of a pedestrian, AND the accident occurs with at least one pedestrian involved, AND the accident occurs outside an urban area and in an intersection THEN the likelihood of the accident resulting in a severe injury is 46.9%

TABLE 21. Rules for C5.0 tree with enlarged data - accident severity

IF the accident occurs with at least one pedestrian or a car rolling involved, AND the accident takes place outside an urban area THEN the likelihood of the accident resulting in a severe injury is 42.8%

TABLE 22. Rules for CHAID tree with enlarged data - accident severity

Among the rules, the involvement of pedestrians in the accident or the rolling of a vehicle appeared more relevant to the outcome of the accident, partly confirming the previous finding that single-vehicle accidents produced more severe outcomes in terms of injuries.

5.2.2.2 Accident location

Considering the location of the accident as the categorical dependent variable, the generated decision trees resulted not completely different. For the C5.0 tree, the most relevant variables to explain the location were the typology of accident, the illumination conditions on the road, the width of the road and the involvement of motorcycles. For the CHAID tree, the most significant fields were the regulation of the intersections, the typology of the accident, the width of the road and the involvement of motorcycles. The prediction accuracy for the C5.0 algorithm was equal to 56.9% and for the CHAID algorithm was equal to 77.0%.

Note that the same difference was found for the analysis of the previous data source, and the confusion matrices in tables 23 and 24 provide insight into the different predictive performances between the two techniques. Remind that the rows represent the actual observations and the columns the predicted values, consequently the elements in the diagonal constitute the correct predictions for accidents happened in 2000, based on the models estimated with the accidents occurred between 1996 and 1999.

	INTERURBAN INTERSECTION	**INTERURBAN SECTION**	**URBAN INTERSECTION**	**URBAN SECTION**
INTERURBAN INTERSECTION	35	119	1053	399
INTERURBAN SECTION	22	962	634	1062
URBAN INTERSECTION	27	101	4737	1970
URBAN SECTION	15	357	1892	4347

TABLE 23. Confusion matrix for C5.0 tree with enlarged data - accident location

	INTERURBAN INTERSECTION	INTERURBAN SECTION	URBAN INTERSECTION	URBAN SECTION
INTERURBAN INTERSECTION	0	0	1606	0
INTERURBAN SECTION	0	747	14	1919
URBAN INTERSECTION	0	0	6835	0
URBAN SECTION	0	466	78	6067

TABLE 24. Confusion matrix for CHAID tree with enlarged data - accident location

Both algorithms failed to predict crashes occurring outside urban areas in intersections, and even though the prediction accuracy of the CHAID tree was superior, this method completely ignored this category of collisions. The CHAID algorithm almost perfectly forecasted the location of accidents in urban areas to either intersections or sections, but failed to allocate outside urban areas the crashes while exactly inserting them in either road junctions or road sections. The errors of the C5.0 method were more distributed and did not exhibit a specific pattern.

Some rules with higher confidence level are illustrated in table 25 and 26. Rules for accidents taking place in interurban junction do not show confidence levels over 40% and are not reported. The only rule with confidence level over 60% for accidents occurring in interurban section generated with CHAID algorithm is presented for illustrative purposes. Accidents in interurban sections occurred mainly at night when artificial illumination was not present and the road was averagely wide, and raining conditions had an effect as well. In urban areas, pedestrians and motorcyclists were often involved in crashes that happened mainly in narrow roads at day, and collisions with objects occurred at night when artificial illumination was not present.

IF the accident is a front to side collision, AND the accident occurs in a curve, AND the accident occurs at night without lighting THEN the likelihood of the accident occurring in an interurban section is 90.0%
IF the accident is a car rolling out of the road, AND the accident happens during night without artificial lighting, AND the width of the road where the accident occurs is between 7 and 10.5 m. THEN the likelihood of the accident occurring in an interurban section is 86.0%

TABLE 25 Rules for C5.0 tree with enlarged data - accident location

IF the accident is a car rolling out of the road, AND the accident happens during day, AND the width of the road where the accident occurs is between 7 and 10.5 m., AND the accident happens in rainy weather, AND the surface of the road is wet THEN the likelihood of the accident occurring in an interurban section is 68.3%
IF the accident is a front to side collision, AND the accident results in light injuries, AND the accident occurs with one motorcycle involved, AND the accident happens during night, AND the width of the road where the accident occurs is between 7 and 10.5 m. THEN the likelihood of the accident occurring in an urban intersection is 78.1%
IF the accident is a front to side collision, AND the accident results in light injuries, AND the accident occurs with one motorcycle involved, AND the accident happens during day, AND the width of the road where the accident occurs is between 5 and 7 m. THEN the likelihood of the accident occurring in an urban intersection is 75.2%
IF the accident is a collision with an object, AND the accident happens during night without artificial lighting, AND the width of the road where the accident occurs is up to 5 m. THEN the likelihood of the accident occurring in an urban section is 74.6%
IF the accident is a front to side collision, AND the accident results in light injuries, AND the accident occurs with one motorcycle involved, AND the accident happens during day, AND the width of the road where the accident occurs is up to 5 m. THEN the likelihood of the accident occurring in an urban section is 67.4%
IF the accident occurs with at least one pedestrian involved, THEN the likelihood of the accident occurring in an urban section is 67.4%

TABLE 25. Rules for C5.0 tree with enlarged data - accident location (continued)

IF the accident is a front to side collision, AND the accident results in light injuries, AND the accident occurs with one motorcycle involved, AND the accident happens during night, AND the width of the road where the accident occurs is between 7 and 10.5 m. THEN the likelihood of the accident occurring in an urban intersection is 78.1%
IF the accident is a front to side collision, AND the accident results in light injuries, AND the accident occurs with one motorcycle involved, AND the accident happens during day, AND the width of the road where the accident occurs is between 5 and 7 m. THEN the likelihood of the accident occurring in an urban intersection is 75.2%
IF the accident is a collision with an object, AND the accident happens during night without artificial lighting, AND the width of the road where the accident occurs is up to 5 m. THEN the likelihood of the accident occurring in an urban section is 74.6%
IF the accident is a front to side collision, AND the accident results in light injuries, AND the accident occurs with one motorcycle involved, AND the accident happens during day, AND the width of the road where the accident occurs is up to 5 m. THEN the likelihood of the accident occurring in an urban section is 67.4%
IF the accident occurs with at least one pedestrian involved, THEN the likelihood of the accident occurring in an urban section is 67.4%

TABLE 26. Rules for CHAID tree with enlarged data - accident location

IF the regulation of the intersection is missing AND the width of the road where the accident occurs is between 7 and 10.5 m., AND the accident is a car rolling or slipping out of the road THEN the likelihood of the accident occurring in an interurban section is 62.1%
IF the regulation of the intersection is blinking yellow, AND the accident occurs with at least one pedestrian involved THEN the likelihood of the accident resulting in an urban intersection is 96.4%
IF the regulation of the intersection is blinking yellow, AND the accident is a front to front collision, AND the accident occurs with at least one motorcycle involved THEN the likelihood of the accident resulting in an urban intersection is 95.5%
IF the regulation of the intersection is a “right of way” sign, AND the width of the road where the accident occurs is over 7 m., AND the accident occurs with at least one woman involved THEN the likelihood of the accident resulting in an urban intersection is 91.1%
IF the regulation of the intersection is blinking yellow, AND the accident is a front to front collision, AND the drivers involved are both Jewish, AND the drivers involved are a man and a woman THEN the likelihood of the accident occurring in an urban intersection is 87.5%
IF the regulation of the intersection is missing AND the width of the road where the accident occurs is between 7 and 10.5 m., AND the accident is a front to side collision, AND the accident occurs with at least one motorcycle involved THEN the likelihood of the accident occurring in an urban section is 90.7%
IF the regulation of the intersection is missing AND the width of the road where the accident occurs is between 5 and 7 m., AND the accident results in fatalities or severe injuries THEN the likelihood of the accident occurring in an urban section is 90.4%
IF the regulation of the intersection is missing AND the width of the road where the accident occurs is over 10.5 m., AND the accident occurs with at least one pedestrian involved THEN the likelihood of the accident occurring in an urban section is 89.1%
IF the regulation of the intersection is missing AND the width of the road where the accident occurs is up to 5 m., AND the accident is a front to side collision, AND the drivers involved are both Jewish THEN the likelihood of the accident occurring in an urban section is 88.7%

TABLE 26. Rules for CHAID tree with enlarged data - accident location (continued)

5.2.2.3 Accident type

When using the typology of accidents as the dependent categorical variables, the two algorithms produced different trees. For the C5.0 tree, the most significant fields were the cause of the accident, the surface conditions, the existence of previous speed violations by one of the drivers and the location of the accident. For the CHAID tree, the most relevant variables were the involvement of motorcycles, the regulation of the intersection, the population group of the drivers and the width of the road. The prediction accuracy for the C5.0 algorithm was equal to 58.6% and for the CHAID algorithm was equal to 73.8%, with an interesting disparity with respect to the predictive ability measured with the first dataset.

The C5.0 algorithm forecasted accurately front to side collisions and accidents with pedestrians involved, even though part of the latter was classified as the former. The CHAID technique predicted accurately front to side collisions, crashes with pedestrians implicated and accidents where cars rolled or slipped out of the road.

Some of the rules for accident type forecasts are illustrated in table 27 and 28, and since the confidence levels for most of the accident types reached values under 50%, the reported rules are some of the most significant ones.

IF the accident is caused by the driver behavior, AND the accident takes place inside an urban area and not in an intersection, AND the accident occurs in a curve, AND the accident happens during day, AND the surface is wet from water THEN the likelihood of the accident being a front to front collision is 53.4%
IF the accident is caused by the driver behavior, AND at least one driver involved has previous speed violations, AND the accident takes place inside an urban area and in an intersection, AND the surface condition is dry, AND the accident results in light injuries, AND the regulation of the intersection is a “stop” sign, THEN the likelihood of the accident being a front to side collision is 77.5%
IF the accident is caused by the driver behavior, AND at least one driver involved has previous speed violations, AND the accident takes place inside an urban area and in an intersection, AND the surface condition is dry, AND the accident results in light injuries, AND the regulation of the intersection is a “right of way” sign, THEN the likelihood of the accident being a front to side collision is 72.3%
IF the accident is caused by the driver behavior, AND the accident takes place inside an urban area and in an intersection, AND the accident results in fatalities, AND the regulation of the intersection is a working traffic light, AND the accident happens during day THEN the likelihood of the accident being a collision with a pedestrian is 73.1%

TABLE 27. Rules for C5.0 tree with enlarged data - accident type

The absence of traffic lights appears related to the majority of the front/side collisions, while pedestrians are involved in crashes at intersection in urban areas Basically front/side and pedestrian collisions were the typologies with higher confidence levels, and it appears that pedestrian are involved in accidents in urban areas, in junctions with narrow roads and large roads.

The prediction accuracy is higher for the CHAID method, as it was for the other predictors, and with this data source it appears that this algorithm outperforms systematically the C5.0 technique. The same conclusion could not be drawn while analyzing the other data source.

IF the accident occurs with no motorcycle involved, AND the regulation of the intersection is either blinking yellow or a “right of way” sign, THEN the likelihood of the accident being a front to side collision is 93.2%
IF the accident occurs with at least one motorcycle involved, AND the regulation of the intersection is either blinking yellow or a “stop” sign, THEN the likelihood of the accident being a front to side collision is 92.9%
IF the accident occurs with at least one motorcycle involved, AND the regulation of the intersection is either blinking yellow or a “right of way” sign, THEN the likelihood of the accident being a front to side collision is 88.6%
IF the accident occurs with no motorcycle involved, AND the accident occurs with two private vehicles involved, AND the regulation of the intersection is a malfunctioning traffic light, AND both drivers involved are Jewish, THEN the likelihood of the accident being a front to side collision is 82.4%
IF the accident occurs with no motorcycle involved, AND the accident takes place inside an urban area and in an intersection, AND the width of the road where the accident occurs is between 5 and 7 m. or between 10.5 and 14 m. THEN the likelihood of the accident being a collision with a pedestrian is 91.7%
IF the accident occurs with no motorcycle involved, AND the accident takes place in an urban area and in an intersection, AND the width of the road where the accident occurs is up to 5 m. THEN the likelihood of the accident being a collision with a pedestrian is 73.8%

TABLE 28. Rules for CHAID tree with enlarged data - accident type

5.2.3 Neural networks

The links between input and output variables were constructed by the MLP networks that processed the same data source used for the elaboration of the decision trees. The relative importance of the input fields was considered to interpret the network model.

5.2.3.1 Accident severity

The neural network evaluated that the severity of the accident was explained mainly by the type of accident, the location of the crash and the existence of previous speed offences by one of the drivers involved. The network generated two hidden layers and reached a precision of 86.1% during the training phase, while the prediction accuracy was slightly different (85.4%) when compared to the outcome of accidents taken place in the year 2000. Table 29 summarizes the most relevant variables when the classification of the accidents according to the severity is performed.

As seen for the decision trees, the actual confusion matrix indicates that the fatal accidents are not correctly forecasted, even though slightly more severe accidents are predicted. Most important, the substantially higher computational cost did not produce any significant advantage in terms of predictive ability, consequently confirming that decision trees are better and more efficient methods than neural networks.

VARIABLES	RELATIVE IMPORTANCE
type of accident	0.1727
location of the accident	0.0990
speed offences	0.0612
shoulders on the road	0.0608
private vehicles	0.0576
lights on the road	0.0575
width of the road	0.0573
public vehicles	0.0505
other variables	< 0.0500

TABLE 29. Relevant input variables for MLP network with enlarged data - accident severity

5.2.3.2 Accident location

When analyzing the location of the accident, the most significant variables were the typology of accident, the illumination on the road, the width of the road and the regulation of the intersections. Table 30 details the most relevant input variables in predicting the location of the accidents.

The estimated precision during the training stage was equal to 80.4%, and the predictive ability was equal to 79.7% when the predicted values were confronted with the observed locations for accident occurred in the year 2000.

The confusion matrix in table 31 illustrates further the problem described in the previous sections: the high prediction accuracy does not necessarily represent a good model, as any accident that actually took place in a junction was allocated to an urban area and the same happened to any crash that actually occurred far from a junction. At least for interurban

sections half of the collisions were correctly allocated, while for the lower right part of the matrix the precision accuracy is almost perfect.

VARIABLES	RELATIVE IMPORTANCE
accident type	0.1423
lights on the road	0.1140
width of the road	0.0852
accident severity	0.0821
regulation of the intersection	0.0816
motorcycles	0.0701
collision with an object	0.0529
weather	0.0467
other variables	< 0.0400

TABLE 30. Relevant input variables for MLP network with enlarged database - accident location

	INTERURBAN INTERSECTION	INTERURBAN SECTION	URBAN INTERSECTION	URBAN SECTION
INTERURBAN INTERSECTION	84	0	1522	0
INTERURBAN SECTION	0	1434	0	1246
URBAN INTERSECTION	99	0	6736	0
URBAN SECTION	0	727	3	5881

TABLE 31. Confusion matrix for MLP network with enlarged data - accident location

5.2.3.3 Accident type

The last dependent variable considered for neural network analysis is the typology of accident. The most relevant input variables for this predictor were the object with which there was the collision, the cause and the location of the accident and the condition of the road surface. Table 32 illustrates the relative importance of the input fields.

The estimated precision during the training stage was equal to 73.2% and the prediction accuracy with respect to accidents that took place in the year 2000 was 74.8%. As for the decision trees, the neural network model predicted correctly in particular front/side collisions,

as well as crashes with objects and pedestrians. Again, neural networks produced comparable results with respect to decision trees, but at much more expensive computational costs.

VARIABLES	RELATIVE IMPORTANCE
object collided	0.2266
cause of the accident	0.1218
location of the accident	0.1063
condition of the road surface	0.0482
regulation of the intersection	0.0439
other variables	< 0.0400

TABLE 32. Relevant input variables for MLP network with enlarged data - accident type

As for the other dependent variables, the most relevant input variables were the characteristics of the accident and of the vehicles, rather than the characteristics of the drivers. The decision trees produced rules in which some of the latter variables were relevant, but most likely the amount of missing data caused the analysis to rely on the variables most frequent in all the records. For this reason the collision characteristics probably resulted more relevant for the accident analysis.

6 Conclusions and further research

This study focused on one hand on the search for the most promising data mining techniques for accident analysis, and on the other hand on the analysis of accident data to provide some classification and predictive rules to understand accident occurrence.

Accordingly, the presentation of the conclusions and the proposition of further research develop in three mainstreams that constitute the main elements of this research: the data mining techniques, the accident data and the safety recommendations from the application of the former to the latter.

6.1 Data mining techniques

This research studied initially several data mining techniques and focused on the understanding of both descriptive and predictive methods. The initial selection of eligible techniques for this study was based on literature survey findings, which proposed the test implementation of some among the large number of alternative techniques available.

Namely, the descriptive techniques that were suitable for categorical data were K-means clustering and Kohonen networks, while the predictive methods that were suitable for the same type of data were CHAID and C5.0 algorithms for the construction of decision trees, the MultiLayer Perceptron neural networks and the Association Rules algorithms.

Descriptive techniques were useful to classify the large amount of analyzed accidents, and between the two techniques considered, K-means clustering appeared to be more effective. Clusters from the K-means algorithm appear to be more clear-cut defined and classification is less trivial than the one obtained with the Kohonen networks, where the occurrence of the accident during day or during night seemed to be the most relevant factor to identify similarities among the collisions. The limit of descriptive techniques is that for safety analysis they do not provide indications different from the mere description of the main characteristics of the accidents, and for this reason predictive methods have to be considered more suitable.

Among the different techniques analyzed, the less satisfying has been the search for Association Rules. The number and the quality of the hundreds of rules generated did not allow to provide significant results and this technique was not considered extensively in the research for these difficulties of interpretation of the findings of the rules.

Neural networks presented a similar problem of interpretation of the results, especially in terms of understanding the relative importance of the inputs and of the intermediate neurons that constituted the two hidden levels connecting inputs and outputs. The definition of the outputs was a general problem in the prediction phase, as neural networks and decision trees require the apriori definition of the output variables and consequently highly depend on this variable choice.

Decision trees performed better than neural networks in terms of definition of clear rules that illustrate the most relevant variables for the outcome of an accident according to the different categories of the output variable (for example the location of the accident). According to the experience developed in this research, decision trees with CHAID algorithm produced the most promising results in terms of predictive accuracy and interpretability of the rules. C5.0 algorithm also performed satisfactorily, but failed to predict some of the categories for some of the output variables.

From a general perspective, once a problem is assigned and an output variable is accordingly defined, decision trees appear the most promising techniques to investigate the problem and provide rules able to explain the phenomenon. The extraction of the rules is extremely simpler than with neural networks, and the different algorithms help treating any different type of variables (nominal, ordinal, categorical etc.), just as in this research the CHAID algorithm was the most suitable to the available data.

6.2 Accident data

This research exploited two different databases. On one hand a database was constituted by information concerning the collision, from its nature to its location, from the infrastructure to the atmospheric conditions, from some general characteristics of the drivers to some general description of the vehicles involved. On the other hand a database was constituted by detailed information concerning the drivers and the vehicles that took part into the collision.

The first problem concerning the data is related to the fact that, when the two sources were merged, some of the information from the first database was removed to account for the increase of detail at the driver and vehicle level. This trade-off in terms of accuracy had a high cost in terms of missing data: for example the exact location of the accident was removed from the file in order to provide details of the drivers. This implied a loss of information especially regarding some of the infrastructure characteristics, such as the median condition,

that appeared relevant when analyzing only the accident characteristic. As expected, analysis of data with large number of missing values had implications on the quality of the results.

The second problem concerning the data is related to the nature of the information itself. Most of the variables included in the database are categorical, and they are pre-defined without the possibility of working on the definition of the categories themselves for the different variables. Accordingly, the analysis is influenced by the constrained definition of the categories, which are used in the data mining implementation as they were pre-defined. Accident location for example were classified according to four categories (urban vs. interurban and intersection vs. section), and results show that this classification was too general.

The conclusion of this research is that the importance of the data is extremely crucial when data mining and accident analysis are involved. Further, more independence should be given to the analyst, in the sense that data should be as general as possible in the description of the information, in order for the analyst to work on the definition of the classification. Of course, more accuracy in the collection of the information regarding the accidents would be welcome, as not only missing data are frequent, but also errors in the data gathering.

6.3 Safety recommendations

This study focused on different data mining techniques and allowed to understand which methods are more suitable for analyzing accident categorical data. The results of the implementation of these techniques, presented in chapter 5, exhibit quite complex relationships between the explanatory variables and the chosen dependent variables. Accordingly, it results extremely difficult to pinpoint clear safety recommendations from the rules defined for example with the decision trees or the neural networks.

This study does not provide a ranking or a quantitative measure of the relative importance of the explanatory variables. Nevertheless, some variables were found more significant than others and appeared more frequently in the clusters and in the rules, according to the techniques implemented and the output variables considered.

From the analysis of the results, it appears evident that there are safety issues with respect to accidents involving pedestrians, as their number is quite high and typically involve pedestrians not in intersections that cross roads where the allowed speed is around 90 km/h, or

pedestrians that cross roads in winter or with bad weather. The first typology suggests problems in the behavior of the pedestrians that perhaps underestimate the time necessary to cross, while the second typology suggests problems in the behavior of the drivers that perhaps have limited visibility also because of lack of prudence in difficult driving conditions. From the analysis of the accident rules, it appears also clear that the regulation of the traffic plays a significant role. Not only respecting signals or traffic lights would reduce the likelihood of front to side accidents, the most recurring collisions in the country, but also checking for their correct functioning as non-working traffic lights appear related to high likelihood of accident occurrence.

The accident severity seems to be influenced mainly by the typology of vehicles involved. Note that when bicycles, motorcycles and heavy commercial vehicles take part into the collision, the outcome results more severe than when only private vehicles are implicated. According to the same principle, the involvement of pedestrians produces more severe injuries and fatalities in crashes. Also relevant for the severity of the accident are the conditions of the road and the illumination on the infrastructure.

The accident location looks as if it is 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 occur 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 car getting off the road. In urban areas, pedestrians, cyclists and motorcyclists are often involved, and in this case supposedly sometimes their behavior causes the crashes that result immediately in severe consequences.

As clearly noticed throughout the research, this study did not concentrate on a single issue and focused on the analysis of the potential of data mining techniques. Given the absence of focus on a more specific safety issue, the broad amount of results is more difficult to interpret and clear safety recommendations are more difficult to provide.

6.4 Further research

According to the presented conclusions, this report intends to provide some guidelines for further research in the same three mainstreams that constitute the main elements of the study.

In data mining, this study focused mainly on methods that exploited pre-specified dependent variables and tested several models that work with these variables. Predictive models that do not require apriori output variable, such as association rules, did not produce satisfactory results and at the same time leaved the door open to further investigation.

Most likely association rules require additional investigation and contemporaneously more work on the data format in order to produce a reasonable number of logical and interpretable rules. The further study of the implementation of association rules algorithm, not largely explored in literature, is a possible development of this research.

With respect to the data gathering, data need refinement both from a quantitative and a qualitative perspective. From the quantitative perspective it would be ideal not to have a trade-off between enrichment and accuracy of the databases. From the qualitative perspective it appears necessary a better definition of some elements that are of interest when analyzing accidents.

For example, the attempted analysis of the black spot influence on the accident occurrence gave evidence that the definition of the black spots needs reconsideration (or actual recalculation) as the data mining techniques were not able to classify accidents in black spots differently from accidents not in black spots. This was a consequence of the definition of black spots not as critical points in the network, but as sections between intersections, even of few kilometers.

With respect to safety recommendations, this research focused on several accident issues rather than on specific problems. Some models analyzed accidents according to their location, other models according to their severity, but any specific issue (for example pedestrian accidents) was considered extensively, also because this was an exploratory study about the potential of data mining techniques. Accordingly, conclusions in terms of safety indications are partly missing the mark, even though the research was successful in terms of comprehending the potential of data mining techniques and the requirements of the databases for obtaining meaningful results.

The main recommendation for further research is to specify a clear safety question, and through the specification of this issue to analyze the problem with both suitable data mining techniques and data definition.

References

Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining association rules between sets of items in large databases, Proceedings of ACM SIGMOD Conference on Management of Data, Washington D.C., USA, May 26-28, pp. 207-216.

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

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

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

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

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

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

Feelders, A. Daniels, H., and Holsheimerc, M. (2000). Methodological and practical aspects of data mining. Information and Management, 37, pp. 271-281.

Geurts, K., Wets, G., Brijs, T. and Vanhoof, K. (2003). Profiling High frequency accident locations using association rules. Proceedings of the 82th Annual Meeting of the Transportation Research Board, Washington, January 12-16, USA.

Geurts K., Wets G., Brijs T. and Vanhoof, K. (2003) Clustering and profiling traffic roads by means of accident data. Proceedings of the European Transport Conference, Strasbourg, France, October 8-10.

Laube, P. (2001). A classification of analysis methods for dynamic point objects in environmental GIS, GI in Europe: integrative, interoperable, interactive. In Proceedings of the 4th AGILE Conference, Konecny, M., Ed., pp. 121-134.

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

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

Smith, B.L., Scherer, W.T. and Hauser, T.A. (2001). Data mining tools for the support of traffic signal timing plan development. Transportation Research Record No. 1768, National Research Council, Washington D.C., pp. 141-147.

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

Tu Bao, H.O. Introduction Knowledge Discovery and Data Mining, Internet Publication.

Witten, I.H. and Frank, E. (2005) Data mining: practical machine learning tools and techniques. Elsevier, Amsterdam.

Yamamoto, T., Kitamura, R. and Fujii, J. (2002). Drivers' route choice behavior: analysis by Data mining algorithms. Transportation Research Record No. 1807, National Research Council, Washington D.C., pp. 59-66.

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