MAPPING PATTERNS AND CHARACTERISTICS OF FATAL ROAD ACCIDENTS IN ISRAEL

Carlo G. Prato, Technical University of Denmark

Victoria Gitelman, Technion – Israel Institute of Technology

Shlomo Bekhor, Technion – Israel Institute of Technology

ABSTRACT

This paper intends to provide a broad picture of traffic accidents in Israel by uncovering their patterns and determinants in order to answer an increasing need of designing preventive measures, addressing particular situations and targeting specific social groups. The analysis focuses on 1,793 fatal accidents occurred during the four-year period between 2003 and 2006, and applies data mining techniques with the objective of extracting from the data relevant information about accident patterns and major factors without a-priori assumptions about the expected outcome of the study. Kohonen neural networks reveal five accident patterns: (i) single-vehicle accidents of young drivers; (ii) multiple-vehicle accidents between young drivers; (iii) accidents involving either motorcycles or bicycles; (iv) accidents where elderly pedestrians crossed in urban areas; (v) accidents where mostly young children and teenagers cross roads in small villages. Feed-forward back-propagation neural networks validate clustering results and indicate that demographic characteristics of both victims and drivers are the most relevant determinants. Other significant factors are road conditions, time of day, traffic control systems and degree of urbanisation at the location of the accident.

Keywords: accident patterns, cluster analysis, Kohonen networks, feed-forward back-propagation neural networks.

INTRODUCTION

Traffic accidents are one of the leading causes of injury and trauma death in the general population in Israel, where yearly over 400 people are killed and over 35,000 people are injured. More than 30,000 people were killed in traffic accidents since the establishment of Israel over the last sixty years. The identification of accident patterns and the individuation of their determinants constitute an increasing need in order to design preventive measures with the ultimate objective of reducing the annual number of traffic fatalities and accidents.

This study addresses the challenge of providing a broad multi-faceted picture of traffic accidents in Israel, with the aim of identifying particular situations and specific population groups in order to design specialized preventive measures, to plan drivers' educational programs and to define correct targets for safety campaigns. For this purpose, this study focuses on recent data of fatal accidents occurred in Israel during the four-year period between 2003 and 2006. The temporal scope is chosen with the objective of having a long enough period to limit random fluctuations in the accident counts and a short enough period to limit changes in road and traffic conditions.

This study also discusses methodological issues concerning accident analysis. Specifically, this study implements data mining techniques with the objective of extracting from the data relevant information about accident patterns. In the accident analysis area, the advantage of data mining techniques is two-fold with respect to regression analysis. First, data mining techniques do not require any pre-defined underlying assumptions regarding the relationship between target (dependent variable) and predictors (independent variables). While regression analysis has been the most popular technique in traffic safety analysis because the relationship between accidents and their determinants can be clearly identified, its disadvantage is that they may lead to biased estimation of accident likelihood and erroneous interpretation of safety implications in case that the underlying assumptions are violated. Second, data mining techniques allow the classification of cases from datasets in which a large number of variables with an even larger number of categories is considered, overcoming limitations of regression techniques with respect to the number of variables and categories considered. A possible limitation of data mining techniques is the detection of trivial patterns of accidents such as "accidents occur in conditions of rainy weather, wet surface and slippery roads". Indeed, previous studies about the applicability of data mining techniques to accident analysis confirm the existence of difficulties in obtaining non-trivial results in terms of safety recommendations (see Prato et al., 2010). In order to overcome this limitation, this study proposes a methodology that enables to attain less predictable than obvious accident patterns. The methodology consists of partitioning the dataset with respect to main accident characteristics and comparing the results in order to uncover consistent accident patterns that represent major non-trivial safety issues that need to be prioritized.

The remainder of this paper is structured as follows. The next section presents the literature about accident pattern recognition from the methodological perspective. Then, the illustration of the dataset about fatal accidents is followed by the description of the applied data mining techniques. Last, accident patterns and major determinants are presented and the main findings of the study are summarized.

LITERATURE REVIEW

The literature shows a growing interest in the individuation of accident patterns and the individuation of their determinants. This section examines the main factors characterizing accident patterns and analyzes the suitability of the applied methodologies to accident pattern recognition.

Accident determinants

The literature in accident pattern recognition allows identifying main determinants of crash occurrence, as variables that concur to determine different types of accidents are also considered relevant crash predictors. Results from existing studies illustrate difficulties in obtaining meaningful and insightful conclusions regarding the determinants of traffic accidents, and accordingly the design of effective preventive measures based on general conclusions is difficult at best.

For example, the differentiation of accidents according to their modality (e.g., Retting et al. 1995; Preusser et al., 1995; Retting et al., 2000; Fleury and Brenac, 2001; Laapotti and Keskinen, 2004) is an example of general finding from which appears difficult to design effective prevention measures. In fact, all these studies pre-determine that accidents are different according to hypothesized modalities of occurrence, a logical and straightforward assumption that does not answer the question about determinants that bring to the occurrence of a single-vehicle accident rather than a front-to-side collision.

Other examples are the individuation of the relevance of alcohol consumption (e.g., Fontaine and Gourlet, 1997; Laapotti and Keskinen, 2004), speeding (e.g., Berg et al., 2004; Laapotti and Keskinen, 2004; Wang et al., 2008; Skyving et al., 2009), urban or rural location of the crash (e.g., Berg et al. 2004; Wang et al. 2008; Skyving et al. 2009), and rainy conditions (e.g., Geurts et al., 2005; Chang and Chen, 2005; Skyving et al., 2009). Clearly, these general findings are of limited importance as further details regarding the vulnerability of different social groups affected by these factors and the severity impact in particular situations are necessary in order to design effective safety measures. Less common findings are scarce in the aforementioned literature. The studies of Hasselberg et al. (2005) and Tseng et al. (2005) are unique examples relating accident occurrence with social patterns and socio-economic status, and inattention and physical-mental conditions respectively.

Accident pattern recognition methods

Finding non-trivial solutions motivates the search of an appropriate methodology for accident pattern recognition and crash determinants individuation.

Some studies do not rely on statistical techniques, but rather on a-priori definition of accident types that are then compared with the collected data to measure the validity of the assumptions (e.g., Preusser et al., 1995; Retting et al., 1995; Retting et al., 2000; Fleury and Brenac, 2001; Laapotti and Keskinen, 2004). The idea is similar to clustering, only the clusters are defined prior to the data analysis rather than obtained after the application of a statistical technique to the data. Knowledge and experience of the researcher play a major role in the a-priori definition of accident patterns, hence this technique seems subjective rather than objective in its implementation.

Some studies apply factorial analysis of correspondence, often combined with hierarchical ascendant classification (e.g., Fontaine and Gourlet, 1997; Berg et al., 2004; Hasselberg et

al., 2005; Wang et al., 2008; Skyving et al., 2009). A limitation of these studies is that these largely used techniques require a limited number of input variables to work properly. Typically, these studies consider four to six predictors and a maximum of twelve, which means a large amount of assumptions is required prior to the analysis in terms of decisions about the variables to be considered. Clusters are likely affected by the assumptions subjectively selected by the analyst.

Wong and Chung (2007) describe the rough set theory for accident classification, a non-parametric method intended to allow researchers to analyze accidents in multiple dimensions and to model accident occurrence as factor chains. Wong and Chung (2008) implement the theory to a sample of single-vehicle accidents, but the implementation of rough set theory does not look feasible, as the technique is not even described in detail and results suggest that a large number of small clusters with low importance might be identified.

Some studies implement data mining techniques to cluster accidents and perform comparative analyses between different accident types (e.g., Geurts et al., 2003; Geurts et al., 2005; Chang and Chen, 2005; Tseng et al., 2005). In particular, the application of Kohonen neural networks appears suitable to obtain accident patterns through self-organized learning, since they answer the requirements of treating large numbers of variables, assessing their importance in order to refine the analysis, and obtaining correlation patterns without any predefined assumption. Also feed-forward back-propagation neural networks seem suitable to comprehend major determinants through the organization of the different variables in order of relative importance for the discernment among accident patterns. Accordingly the present study applies Kohonen and feed-forward back-propagation neural networks to uncover accident patterns and their determinants.

DATA DESCRIPTION

The fatal accident dataset is prepared on the basis of data gathered by the Israel Central Bureau of Statistics, which collects information regarding every accident resulting in the injury of at least one person involved in the crash. Every year thousands of accidents are reported and organized in three different files that record their characteristics, namely the accident file, the vehicle and driver file, the injured person file.

The accident file includes details regarding each crash: severity of the accident (i.e., fatal, severe, light), date and time, geographical location, distinction between urban and rural as well as section or intersection location. Information regarding date and time is further elaborated into details about season, day of the week and period of the day of the accident. Then, the accident file contains details about the infrastructure: allowed speed, width of the road, presence and condition of median barrier, traffic lights and road signals in general, conditions of the surface related also to the weather at the moment the accident happened. Last, the accident file comprises information concerning pedestrians and collided objects for specific types of accidents. The vehicle and driver file includes records of each vehicle and driver involved in the accident. Each record corresponds to one vehicle and its driver, and lists vehicle features (e.g., type, age, motor, weight and direction of travel of the vehicles)

and driver characteristics (e.g., gender, age, licensing year and birth location). Information contained in this file is further elaborated in order to classify age and social groups of the drivers, driving experience in terms of years from licensure, and age of the vehicles. The injured person file comprises records of each person injured including pedestrians. Each record corresponds to one injured person, and registers generic information (e.g., gender, age, birth location, social group, place of residence and type of injury sustained).

The data processing begins with the merging of the three data files. As the accident file contains one accident for each record and the other two files contain multiple records corresponding to different drivers, vehicles and persons involved in each accident, a unique file is composed by considering as a base file the accident file, and adding progressively vehicles, drivers and persons involved. Then, the consistency of the coding system is verified in the databases provided by the Central Bureau of Statistics. The selection of a four-year period between 2003 and 2006 allows having a consistent coding system that allows avoiding potential problems related to inconsistencies in the variable definition. Last, data are cleaned by examining the possible removal of records with missing values, even though data mining techniques allow considering missing data. The selection of fatal accidents as the focus of the study allows avoiding problems related to the correctness of the data, as most of the errors are reported in accidents that resulted into light injuries, while fatal accidents appear meticulously described and are error free from the missing data perspective.

The fatal accident dataset consists of 1,793 records and the categorical variables considered for data analysis are summarized in table 1.

Table 1 - Categorical variables for data analysis

Variables	Categories
year of the accident	2003 – 2004 – 2005 – 2006
season of the accident	spring – summer – autumn – winter
day of the week	sunday – monday – tuesday – wednesday – thursday – friday – saturday
day / night	day – night
period of the day	morning peak – morning/afternoon off-peak – afternoon peak – evening/night
type of day	normal – pre-festive – festive – inner-festive
type of accident	single-vehicle – front-to-front – front-to-side – front-to-rear – side-to-side – pedestrian
cause of the accident	driver behavior – passenger behavior – pedestrian behavior – motorcyclist behavior – cyclist behavior – vehicle malfunctioning – other
location of the accident	urban intersections – urban sections – rural intersections – rural sections

Variables	Categories
allowed speed	50 km/h – 60 km/h – 70 km/h – 80 km/h – 90 km/h – 100 km/h
weather conditions	clear – rainy – hot – foggy – not specified
road surface conditions	dry – wet from water – wet from slippery material – covered with mud – covered with sand – not specified
number of ways	one-way road – two-way road with separation line – two-way without separation line – not specified
median	with separation fence – without separation fence – not specified
shoulders of the road	good conditions – bad conditions – rough road – bad conditions and rough road
width of the road	up to 5 m. – 5 to 7 m. – 7 to 10 m. – 10 to 14 m. – over 14 m.
illumination on the road	daylight – night without illumination – night with illumination
regulation of intersections	no control – working traffic light – failing traffic light – blinking yellow – stop sign – priority sign – not specified
location of crossing pedestrians	on crosswalks with traffic light – on crosswalks without traffic light – outside crosswalks next to an intersection – outside crosswalks far from an intersection – not specified
crossing pedestrians	suddenly – from hidden places – normally – not specified
type of victim	car driver – car passenger – motorcycle driver – motorcycle passenger – cyclist – pedestrian
age of the victims	less than 14 years old – 15 to 19 years old – 20 to 24 years old – 25 to 34 years old – 35 to 44 years old – 45 to 54 years old – 55 to 64 years old – more than 65 years old
gender of the victims	male – female
social groups of the victims	Jewish – non-Jewish
Jewish victims	born in Israel – from alya
type of vehicles involved	car – light truck – heavy truck – public transport – motorcycle – bicycle – not specified
size of the motor of the vehicles	less than 1000 cc. – 1000 to 1300 cc. – 1300 to 1600 cc. – 1600 to 2000 cc. – over 2000 cc.
age of the vehicles	up to 2 years old – 2 to 5 years old – 6 to 10 years old – 11 to 15 years old – more than 15 years old
safety on board	seatbelts – helmets – child-seat – not used
situation of the vehicles	regular property – rented – stolen
property of the vehicles	private – army – police – other

Variables	Categories
age of the drivers	17 to 19 years old – 20 to 24 years old – 25 to 34 years old – 35 to 44 years old – 45 to 54 years old – 55 to 64 years old – more than 65 years old
licensing years for the drivers	up to 2 years – 2 to 5 years – 6 to 10 years – 11 to 20 years – more than 20 years
gender of the drivers	male – female
social groups of the drivers	Jewish – non-Jewish
Jewish drivers	born in Israel – from alya
type of city	more than 200000 inhabitants – 100000 to 200000 inhabitants – 30000 to 100000 inhabitants – 10000 to 30000 inhabitants – less than 10000 inhabitants

METHODOLOGY

The methodology proposed in the present study is developed with the aim of recognizing accident patterns and their determinants from the analysis of the dataset rather than restricting the analysis to arbitrary predictors, while avoiding results either too difficult to explain or too simple to be relevant. In order to achieve this aim, the current study proposes to apply a sequence of two prominent data mining techniques on partitioned data sets. Specifically, at first Kohonen networks are applied to obtain accident patterns through self-organized learning, since these networks answer the requirements of treating large numbers of variables, assessing their importance to refine the analysis, and obtaining correlation patterns without any predefined assumption. Then, feed-forward back-propagation neural networks are implemented to validate cluster results and classify accident determinants through the organization of different predictors in order of relative importance for discerning among accident patterns. Details about data mining techniques and their implementation in the present study are provided in the following sub-sections.

Data mining techniques

Kohonen networks

A Kohonen neural network (Kohonen, 1982; Kohonen, 2001) is a self-organizing map that is trained in an unsupervised mode, namely the network is presented with data and the corresponding output is not pre-specified. Kohonen networks are very practical because they are relatively simple to construct, can be trained very rapidly and can be applied to linearly separable problems. The structure of a Kohonen network is illustrated in figure 1.

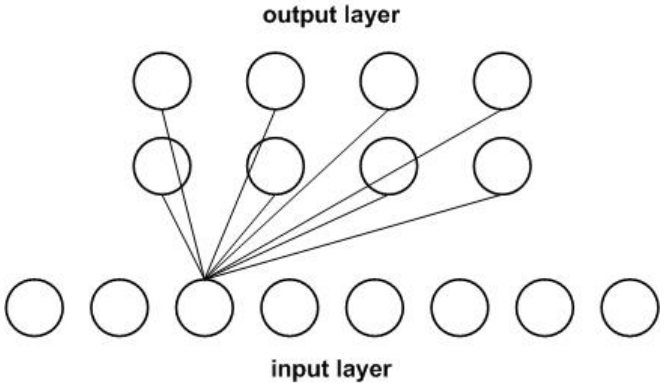

Figure 1 - Structure of a Kohonen neural network

The input is given by neurons that correspond to categories of the input variables. The input neurons cause a reaction from the output neurons by being presented to the network during a training process that involves stepping through several epochs until the error of the network is below an acceptable level. Since the Kohonen network is trained in an unsupervised fashion, the traditional definition of error as the difference between predicted and observed outputs does not apply. However, since the purpose of the network is to classify the input into clusters, the error for the Kohonen network is able to measure how well the network is classifying the cases and also to compute how each of the predictors is relevant to the classification process.

Initially, the network presents cases of the input layer to neurons of the output layer and assigns random weights to the connections between the layers. Characteristics of each record are compared with those of all neurons in the output grid, and the output neuron with the most similar characteristics to the input case is the “winner”. Thus, the weight of this “winner” output neuron is adjusted to be more similar to that of the record just acquired in order to enhance the likelihood of similar records to be captured by the same node. Then, the network adjusts the weights of the neighbouring neurons as well, as each case enters the examination phase. After the data pass through the network, the result consists of a map containing clusters of records corresponding to different patterns in the data. Similar patterns should be closer in the map than patterns that are dissimilar.

The algorithm works in two phases, namely 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. These two phases are defined through different learning parameters and different numbers of neighbours to be modified when the “winner” neuron is identified. The learning parameter is a constant used by the learning algorithm during each epoch and must be a positive number less than 1, typically ranging around 0.4 or 0.5 in the initial phase and around 0.1 or 0.2 in the fine-tuning phase of the network. Setting the learning rate to a large value causes the training to progress faster, but also generates the risk that the network will never converge because the oscillations of the weight vectors will be too great for the classification patterns to ever emerge. For this reason the algorithm is usually programmed at a relatively high learning rate in the first phase and at a decreased rate in the second phase, as training progresses toward the fine-tuning of the output layer. Analogously, the number of neighbours to be modified is higher in the initial phase, when weights are modified

for several clusters to capture similarities among adjacent output neurons, and lower in the last phase, when weights are modified for one or two clusters to underline differences also among adjacent output neurons.

Feed-forward back-propagation neural networks

A feed-forward back-propagation neural network is a network that is trained in a supervised mode (see for example Reed and Marks, 1999). The structure of this type of neural network consists of an input layer corresponding to the categories of the input variables, an output layer corresponding to the categories of the expected output variable, and one or more hidden layers to connect input and output, as illustrated in figure 2.

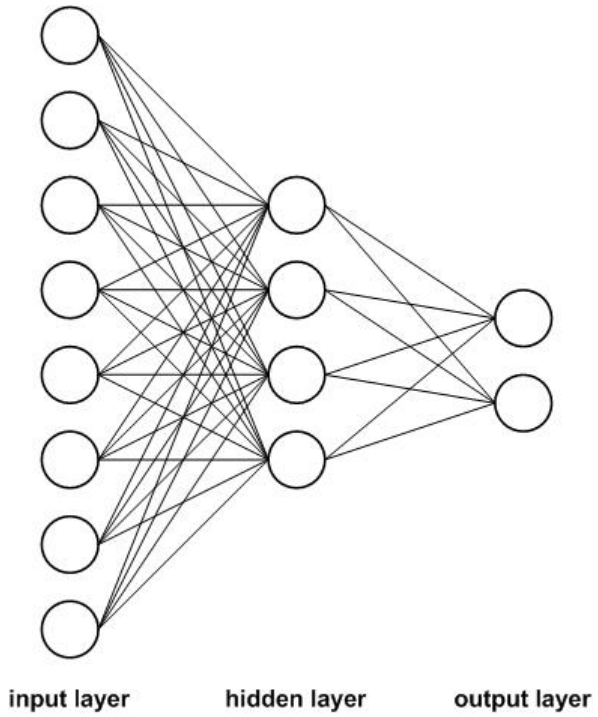

Figure 2 - Structure of a feed-forward back-propagation neural network

The term “feed-forward" describes that the neural network connects neurons only forward, thus each layer of the network contains only links to the next layer (for example from the input to the hidden layer) without any connection back. The term “back-propagation” describes that the neural network has supervised training, thus the network must be provided with sample inputs and anticipated outputs that will be compared against the actual outputs from the neural network. Using these anticipated outputs, the "back-propagation" training algorithm calculates the error and adjusts the weights of the various layers backwards from the output layer all the way back to the input layer.

Initially, the network learning process consists in assigning random weights and obtaining random answers without apparent sense. Then, the network examines single records, forecasts the output for each case and corrects the weights each time a forecast is incorrect with respect to the expected outcome. 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. This learning process reiterates and the network improves its forecasts until one or more interruption criteria are satisfied.

Typically, most neural networks have only one hidden layer and rarely have more than two hidden layers. The structure of the network depends on both the number of neurons inside input and output layers, and the number of hidden layers and their neurons. Ultimately, the selection of the structure comes down to the selection among three processes. The “forward selection method” begins by selecting a small number of hidden neurons with which the neural network is trained and tested. Then, the number of hidden neurons is increased and the process is repeated as long as the overall results of the training and testing improved. The “backward selection method” begins by using a large number of hidden neurons with which the neural network is trained and tested. Then, this process continues until about the performance improvement of the neural network is no longer significant. The “pruning method” involves evaluating the weighted connections between the layers. If the network contains any hidden neurons which contains only zero weighted connections, the algorithm removes them from the network.

Proposed implementation procedure for the data mining techniques

The proposed implementation procedure for the data mining techniques consists of three steps: partition of the dataset, cluster analysis and determinants classification.

In the first step, the fatal accidents dataset is partitioned according to three main characteristics of the accidents. Specifically, when data mining techniques are applied to the overall dataset, typical results illustrate that main characteristics of the accident (e.g., location, crash type, victim type) tend to guide the clustering process toward the division of the records according to predictable patterns such as “accidents with rainy weather, wet road surface and slippery conditions” that do not require complex statistical methods to be uncovered. Instead, partitioning the dataset according to the main characteristics of the accidents prior to the implementation of the data mining techniques limits the possibility of defining only trivial clusters and allows verifying the consistency of the clusters obtained from different initial perspectives. Four data partitions are defined in accordance with three accident characteristics: type of accident (i.e., pedestrian, single-vehicle, multiple-vehicle), accident location (i.e., urban or rural, sections or intersections) and type of victim (i.e., pedestrians, car drivers, car passengers, two-wheel vehicle users).

In the second step, Kohonen networks are applied to the partitions of the accident dataset. For each partition, the number of clusters in the Kohonen neural networks is determined

through a trial and error process, which exploits the property by Kohonen networks of assigning progressively the data to different clusters by dividing bigger clusters into smaller ones. The initial number of clusters in each partition is assumed equal to three and the increase of one cluster at the time stops when the additional cluster either does not contain a significant number of cases or the error of the network does not diminish. At convergence, the comparison between clusters obtained for different partitions allows individuating accident patterns for fatal accidents in Israel during the study period.

In the third and final step, feed-forward back-propagation neural networks with exhaustive pruning are applied to the entire dataset with respect to the same variables used for the construction of the partitions (i.e., type of collision, location of the crash and type of victim), in order to validate the cluster results and to obtain a classification of the accident determinants. The number of input neurons corresponds to the number of categories of the input variables, the number of output variables equals the categories of the accident characteristics chosen to partition the data, and the optimal number of neurons in the hidden layer is determined through the exhaustive pruning process. Note that the output neurons are three for accident type (i.e., pedestrian, single vehicle and multiple vehicle), two for area of accident location (i.e., urban and rural), two for type of accident location (i.e., sections and intersections), and four for victim type (i.e., pedestrians, car drivers, car passengers, two-wheel vehicle users). Moreover, over-fitting is avoided by considering randomly 50% of the dataset for training and the remaining 50% for test before the validation. Accordingly, the neural network minimizes the error by comparing the prediction following the training of the model with 50% of cases with the actual outcomes of the remaining 50% of cases.

DATA ANALYSIS

This section illustrates accident patterns and major determinants resulting from the implementation of Kohonen and feed-forward back-propagation neural networks to the illustrated partitions of the fatal accident dataset.

Cluster analysis for accident pattern recognition

Cluster analysis per type of accident

Three types of fatal accidents are considered for partitioning the dataset: pedestrian (603 cases), single-vehicle (416 cases) and multiple-vehicle (774 cases) accidents.

Five clusters are obtained for pedestrian fatal accidents, as described in figure 3. The most important variables uncovered by the Kohonen networks are the accident location, the width and the presence of a physical median barrier in road sections, the modality and the period of the crash, and the age and social group of the victims. Considering similarities between neighbouring clusters, major differences are characterized in terms of age, as the first cluster contains accidents with elderly while the last cluster includes accidents with children and teenagers, social groups, as the initial clusters refer to Jewish and the last one to non-Jewish

pedestrians, and geographical location, as the map moves from metropolitan areas for the first two clusters to small villages for the last one.

pedestrian urban elderly	• 185 records • pedestrians more than 65 years old • pedestrians crossing on zebras far from intersections • accidents in the Tel Aviv metropolitan area
pedestrian two-wheel vehicles	• 90 records • pedestrians crossing outside zebras suddenly • two-wheel vehicles involved • accidents in urban areas
pedestrian rural night distraction	• 133 records • pedestrians crossing outside zebras suddenly at night • accidents in rural sections
pedestrian youngsters night distraction	• 77 records • pedestrians less than 35 years old • pedestrians crossing outsize zebras suddenly at night • accidents in wide sections
pedestrian urban teenagers	• 118 records • pedestrians less than 14 years old • accidents in urban sections • accidents in small villages

Figure 3 - Clusters of pedestrian accidents

Four clusters are found for single-vehicle fatal accidents, as represented in figure 4. The most important variables are the level of seatbelt law compliance, the location in urban or rural areas, the age of the vehicle and the age and social group of the drivers involved. Examining the similarities between neighbouring clusters, major differences are defined according to the social group of the drivers, as the first cluster relates to non-Jewish while the last one to Jewish drivers, and the geographical location, as the first cluster contains accidents in small villages and the last cluster includes crashes in larger urban areas.

Four clusters are uncovered for multiple-vehicle fatal accidents, as shown in figure 5. The most relevant variables are the level of driving experience of the drivers, the crash location in either a section or an intersection, the period of the day, the age of the vehicle and the age and social group of the drivers. Observing the map of the clusters and the similarities between neighbours, major differences are defined according to social groups, as the first cluster includes accidents involving only non-Jewish and the last one only Jewish drivers, while passing through a cluster where both are contemporarily involved. These accident patterns between Jewish and non-Jewish drivers, as well as between experienced and inexperienced drivers, confirm a theory proposed by Factor et al. (2008) about inter-groups differences being significant traffic accident determinants in Israel.

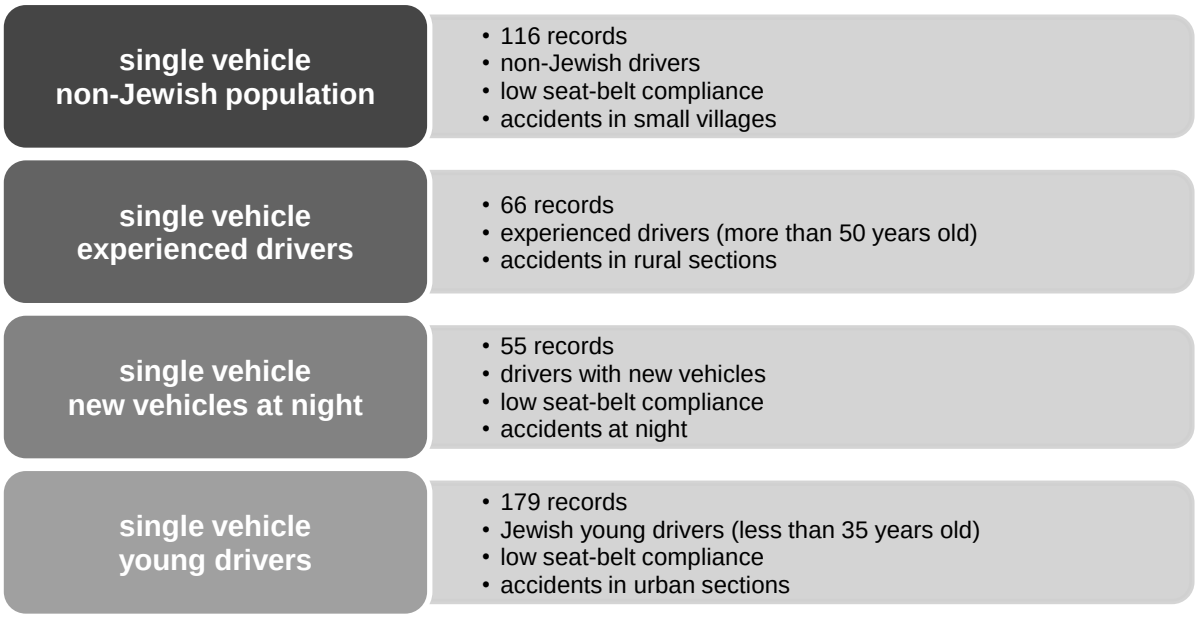

Figure 4 - Clusters of single-vehicle accidents

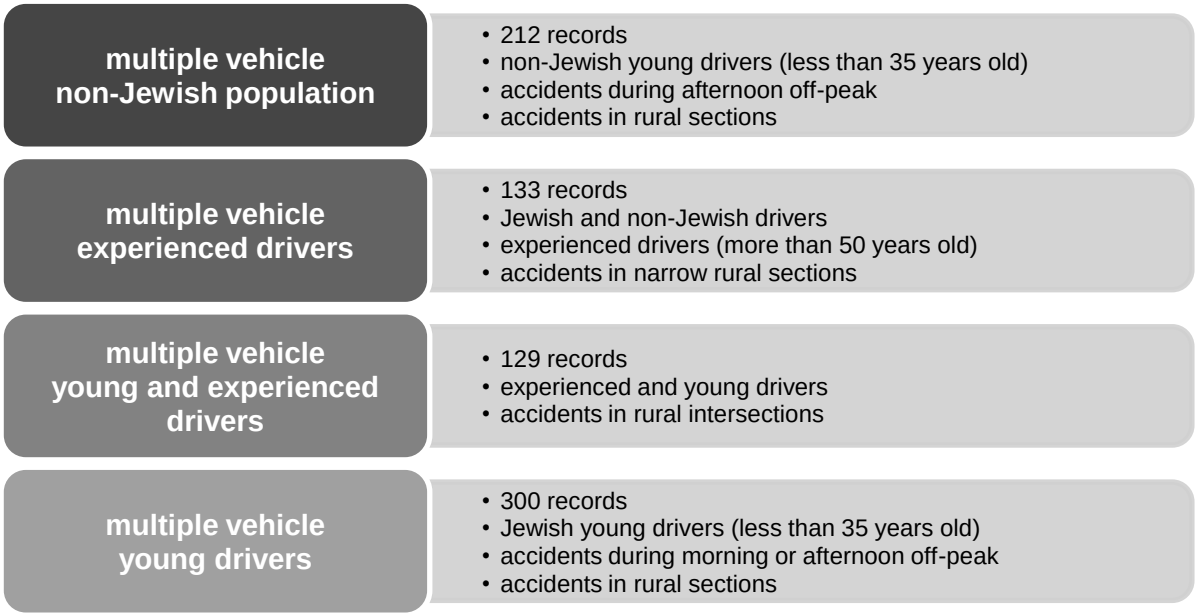

Figure 5 - Clusters of multiple-vehicle accidents

Cluster analysis per location of the accident

Two different types of classification of accident locations are considered for partitioning the dataset: initially, results illustrate clusters for accidents in intersections (451 cases) versus accidents in sections (1342 cases), and then clusters for urban crashes (749 cases) versus rural crashes (1044 cases).

Four clusters are obtained for fatal accidents that occurred in intersections, as described in figure 6. The most important variables are the location in either urban or rural areas, the period of the day, the involvement of pedestrians and the age of drivers and victims. Considering the similarities between neighbouring output neurons in the Kohonen map,

major differences are observed with respect to the type of accident, as multiple-vehicle collisions are opposed to pedestrian crashes, and the age of the victims, as young drivers in the initial clusters are opposite to elderly pedestrians in the last ones.

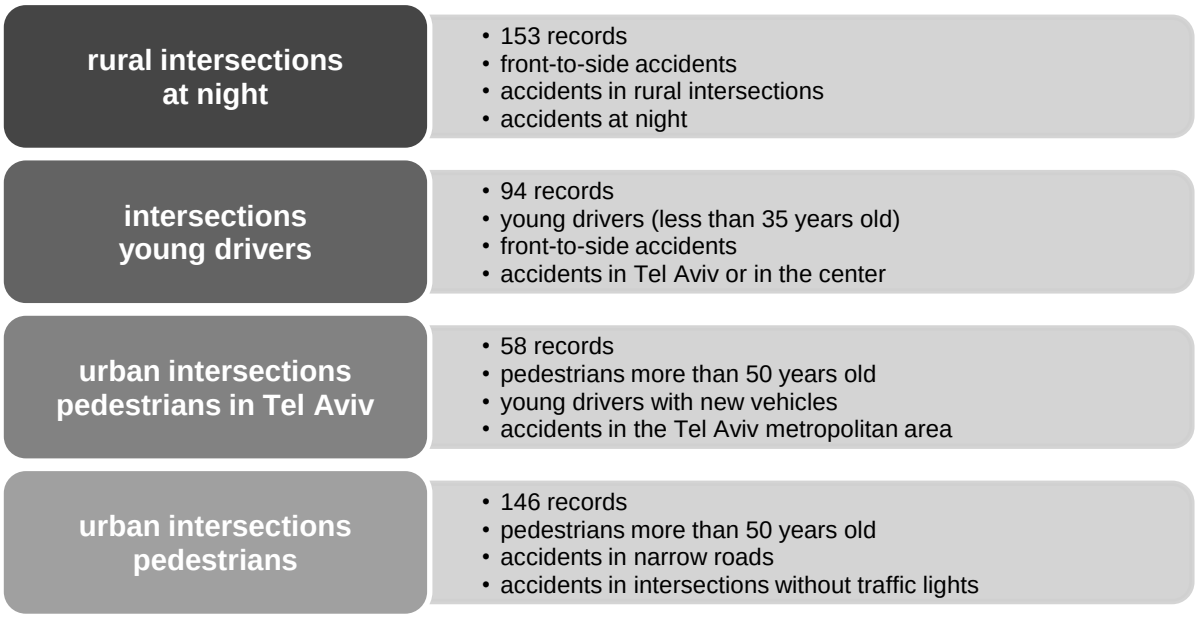

Figure 6 - Clusters of accidents in intersections

Six clusters are found for fatal accidents that took place in road sections, as illustrated in figure 7. The most important variables are the location in either urban or rural areas, the period of the day and the related existence of artificial illumination on the road, the type of accident, the type of vehicles involved, and the age and social groups of drivers and victims. Examining similarities between neighbours in the Kohonen map, major differences are observed in terms of social groups of the drivers, as opposite clusters contain crashes in which either only non-Jewish or only Jewish drivers were involved, period of the day, as the first clusters focus on night accidents without artificial illumination and the last clusters concentrate on day crashes, and type of accident, as initially the clusters refer to multiple-vehicle and then to single-vehicle and pedestrian accidents.

Five clusters are attained for fatal accidents that happened in urban areas, as depicted in figure 8. The most important variables are the location in either metropolitan areas or small villages, the location in either sections or intersections, the period of the day, the width of the road, the compliance with the seatbelt law, and the age and social groups of drivers and victims. Reading the map of output neurons, major differences are observed with respect to the age of the victims, as young drivers in the first cluster is opposite to elderly pedestrians, the period of the day, as initial clusters concentrate on night crashes and the last ones on day crashes, and the geographical location, with different clusters varying between small villages and large metropolitan areas.

rural sections at night non-Jewish population	• 229 records • non-Jewish young drivers (less than 35 years old) • accidents at night without illumination • accidents in rural sections
rural sections at night Jewish population	• 246 records • Jewish young drivers (less than 35 years old) • accidents at night without illumination • accidents in rural sections
sections single vehicle young drivers	• 337 records • young drivers (less than 35 years old) • single vehicle accidents • accidents at night
sections two-wheel vehicles	• 107 records • motorcycles and bicycles involved • accidents in the Tel Aviv metropolitan area
urban sections elderly pedestrians	• 261 records • pedestrian more than 65 years old • accidents in Tel Aviv metropolitan area
urban sections teenage pedestrians	• 162 records • pedestrians less than 14 years old • pedestrians crossing suddenly from hidden places • accidents in small villages

Figure 7 - Clusters of accidents in sections

urban young drivers	• 246 records • young drivers (less than 35 years old) • low seat-belt compliance • accidents at night
urban new vehicles	• 69 records • drivers with new vehicles • low seat-belt compliance • accidents at day
urban sections teenage pedestrians	• 126 records • pedestrians less than 14 years old • accidents in narrow urban roads • accidents in small villages
urban sections pedestrians	• 110 records • pedestrians more than 50 years old • accidents in small villages
urban intersections pedestrians	• 198 records • pedestrians more than 65 years old • pedestrians crossing on zebras • accidents in the Tel Aviv metropolitan area

Figure 8 - Clusters of accidents in urban areas

Five clusters are also uncovered for fatal accidents that occurred in rural areas, as shown in figure 9. The most important variables are the geographical location, the period of the day and the related existence of artificial illumination, the width of the road and the presence of median barrier, the modality of the crash, and the age and social groups of drivers and victims. Considering the similarities between neighbouring clusters, major differences are drawn by the social groups of the drivers, as non-Jewish and Jewish drivers are at the opposite sides of the Kohonen map, and the type of accident, as multiple-vehicle, single-vehicle and pedestrian crashes differentiate the output neurons.

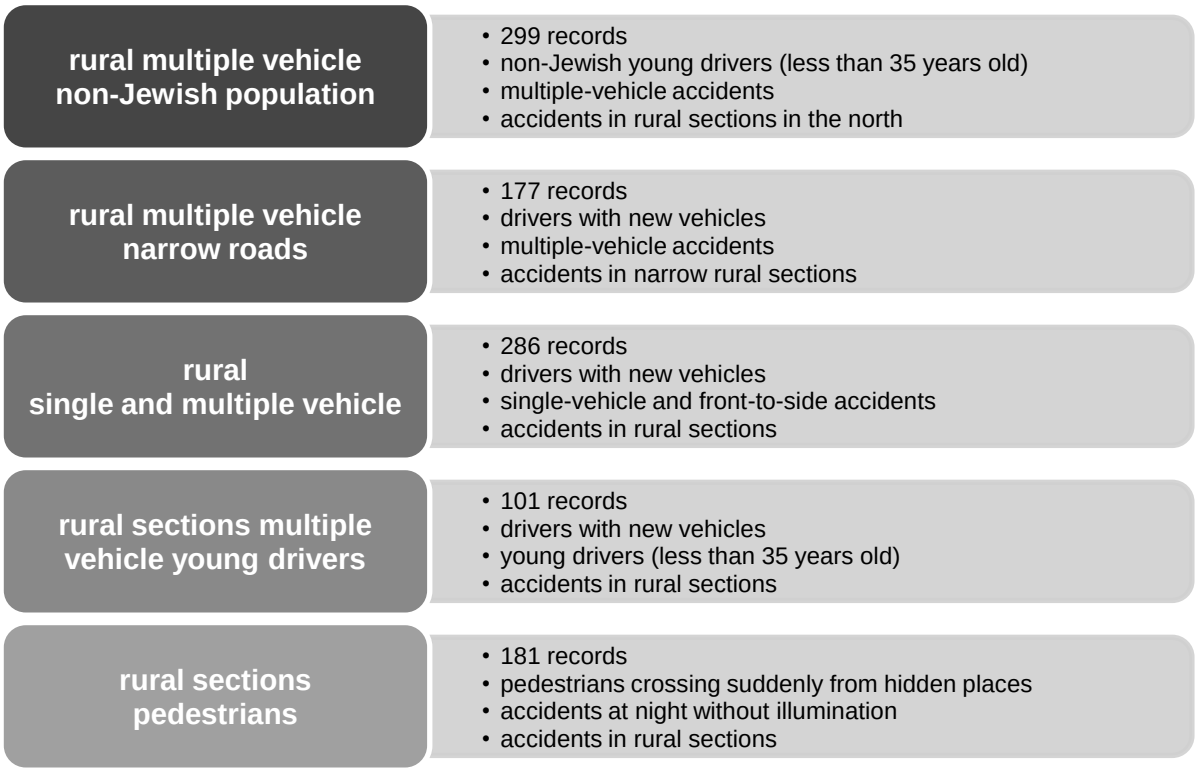

Figure 9 - Clusters of accidents in rural areas

Cluster analysis per type of victim

Four types of victims have been considered for partitioning the dataset: pedestrians (611 cases), car drivers (528 cases), car passengers (433 cases) and two-wheel vehicle users (221 cases).

Five clusters are obtained for fatal accidents involving the death of a pedestrian and are represented in figure 10. As expected, the definition of the clusters for pedestrian victims proposes the same solution of the determination of the clusters for pedestrian accidents, only with the inverted order in the numeration of the output neurons of the Kohonen map. Note that the number of pedestrian victims is slightly larger than the number of pedestrian accident, and indicates that some accidents had multiple pedestrian fatalities.

pedestrian urban teenagers	• 142 records • pedestrians less than 14 years old • accidents in narrow urban roads • accidents in small villages
young pedestrian night distraction	• 76 records • pedestrians less than 35 years old • pedestrians crossing outsize zebras suddenly at night
pedestrian rural night distraction	• 122 records • pedestrians crossing outside zebras suddenly at night • accidents in rural sections
pedestrian urban distraction	• 77 records • pedestrians more than 50 years old • pedestrians crossing outside zebras suddenly • accidents in urban areas
pedestrian urban elderly	• 194 records • pedestrians more than 65 years old • pedestrians crossing on zebras far from intersections • accidents in the Tel Aviv metropolitan area

Figure 10 - Clusters of accidents with pedestrian victims

Four clusters are found for fatal accidents resulting in the death of at least one of the car drivers involved, as illustrated in figure 11. The most important variables are the accident location, the type of accident, the period of the crash, and the age and social group of the victims. Observing the Kohonen map of clusters, major differences are determined mainly by the social group of the drivers, as non-Jewish and Jewish drivers are at opposite sides of the map.

car drivers new vehicles	• 176 records • Jewish drivers with new vehicles • single- and multiple-vehicle accidents • accidents in rural sections
young car drivers rural multiple vehicle	• 130 drivers • young drivers (less than 35 years old) • accidents at night • accidents in rural sections
young car drivers rural sections at night	• 50 records • Jewish experienced and young drivers • accidents at night • accidents in rural sections
young car drivers Arabic population	• 172 records • non-Jewish young drivers (less than 35 years old) • single- and multiple-vehicle accidents • low seat-belt compliance

Figure 11 - Clusters of accidents with car driver victims

Four clusters are attained for fatal accidents resulting in the death of at least one of the passengers of the cars involved, as summarized in figure 12. The most important variables are the accident location, the type of accident, the period of the crash, and the age of the drivers and the victims. Examining the similarities between neighbouring neurons of the Kohonen map, major differences are observed with respect to the period of the day, as night or morning or afternoon characterize the different clusters, and the geographical location of the accidents, as crashes occurred in the south or the center are at opposite sides of the map with respect to crashes occurred in the north.

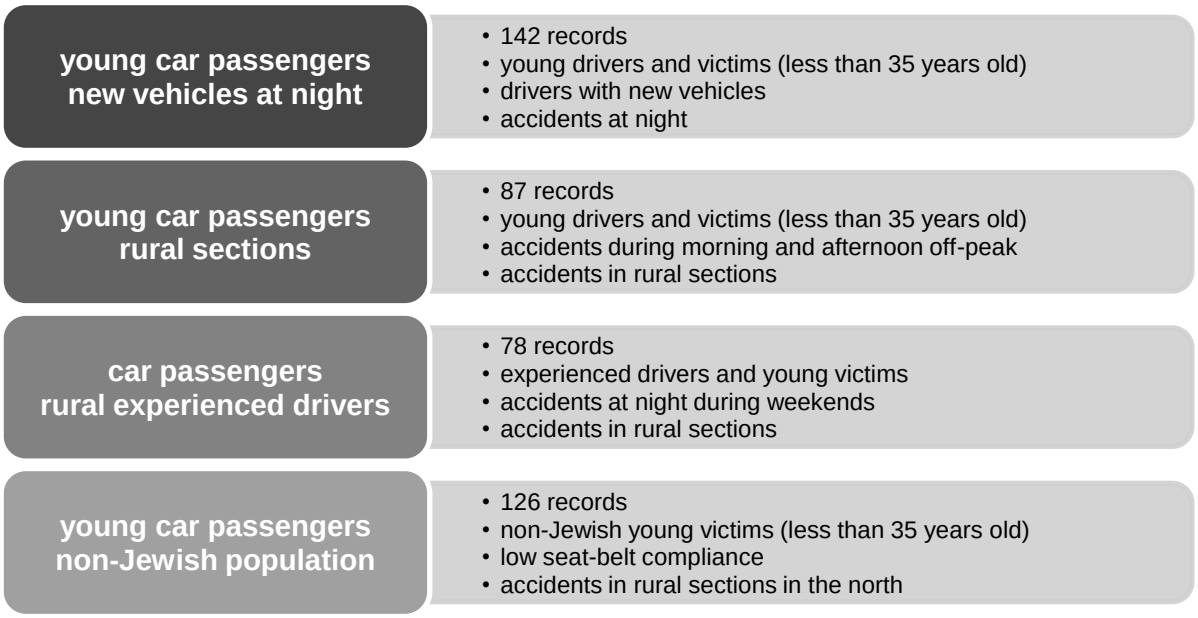

Figure 12 - Clusters of accidents with car passenger victims

Three clusters are uncovered for fatal accidents involving the death of at least one road user on either a motorcycle or a bicycle, as presented in figure 13. The most important variables are the type of two-wheel vehicle, the accident location, the modality of the collision and the age of the riders. Considering the similarities between neighbouring clusters, major differences are related to the type of two-wheel vehicle, as motorcycle accidents are distinguished from bicycle ones.

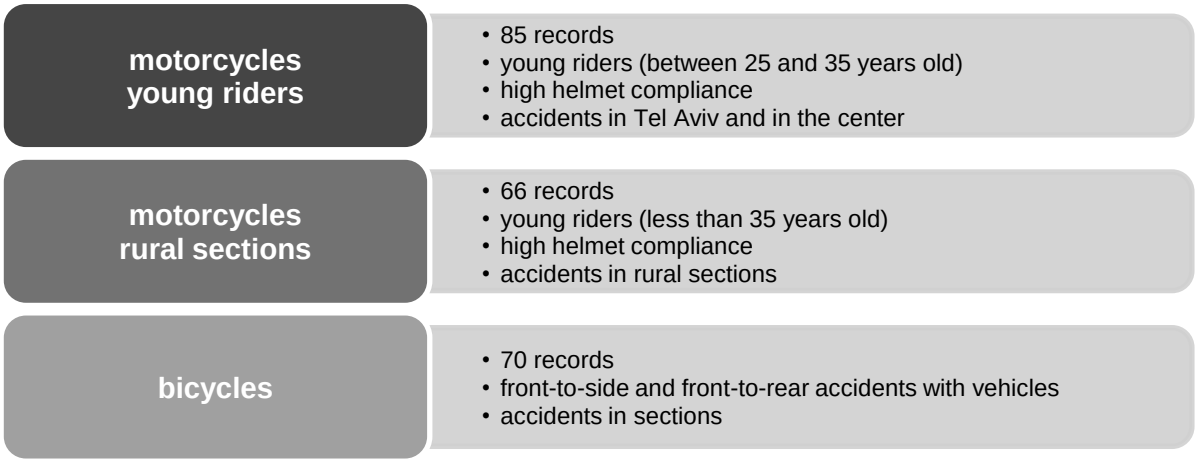

Figure 13 - Clusters of two-wheel vehicle victims

Accident patterns

The large number of clusters obtained from analyzing the different partitions is examined in order to comprehend whether the clustering solutions are consistent, namely whether clusters present similarities once they are obtained from different starting points in the analysis. As some similar patterns are individuated among the clusters, five accident patterns are recognizable regardless of the dataset partition considered.

The first cluster contains single-vehicle accidents of young drivers who tend not to wear seatbelts and to lose control of their vehicles, while driving at night in rural sections without artificial illumination. The second cluster includes multiple-vehicle accidents between young drivers behind the wheel of relatively new vehicles, who have front-to-front and front-to-side collisions in both urban and rural areas at night where artificial illumination is missing. The third cluster consists of accidents involving either motorcycles or bicycles in urban sections. The fourth cluster comprises accidents where elderly pedestrians crossed on crosswalks far from intersections in urban areas, mainly in the Tel Aviv metropolitan area. The fifth cluster includes accidents where mostly young children and teenagers crossed narrow roads in small villages.

Neural networks for factor determination

Feed-forward back-propagation neural networks emphasize the relative importance of each variable with respect to all the other variables in the dataset and intend to further validate the results obtained with the Kohonen networks. Note that for implementation of feed-forward back-propagation networks the dataset is not partitioned, but rather the layer of output neurons has as many elements as the categories of the variable that is considered for partitioning the dataset during the cluster analysis. The following sub-sections detail the results of the feed-forward back-propagation networks by presenting the most important variables and introducing confusion matrices between the expected outcome read in the output neurons for the testing half of the dataset and the predicted outcome obtained from the model trained with the first half of the dataset.

Factor determination per type of accident

Figure 14 summarizes the relative importance of the input variables for the three possible outcomes when the dataset is considered with respect to the type of accident: pedestrian, single-vehicle and multiple-vehicle collisions. The age of the victim is the most important variable, and in fact cluster analysis revealed that children, teenagers and elderly are mostly victims of pedestrian accidents, young drivers up to 35 years old are mainly victims of single-vehicle crashes and experienced drivers are generally victims of multiple-vehicle collisions. The social group of both victims and drivers is also important, and in fact almost clear-cut separation between Jewish and non-Jewish population characterizes most of the clusters, regardless of the dataset partition considered. The width of the road, the location and the period of the day also contribute significantly to the discernment of the type of accident.

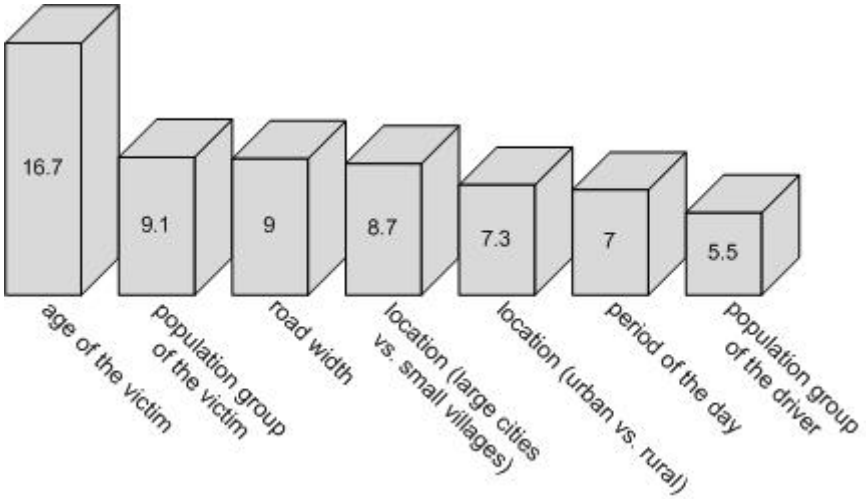

Figure 14 - Relative importance of the input variables for type of accidents

Table 2 introduces the confusion matrix between predicted and expected outcomes of the neural network. The correctness of the prediction of the neural network with three output neurons corresponding to the type of accident is 68.6%, which means that over two thirds of the outputs of the remaining half of the dataset are correctly predicted by the model trained with the initial half of the dataset. Multiple-vehicle accidents are the typology most accurately predicted by the feed-forward back-propagation neural network, while most of the error in in the confusion between pedestrian and single-vehicle, as well as between single-vehicle and multiple-vehicle collisions.

Table 2 - Confusion matrix for types of accidents

	Pedestrian	Single-vehicle	Multiple-vehicle
Pedestrian	416	113	74
Single-vehicle	69	271	76
Multiple-vehicle	56	175	543

Factor determination per location of the accident

Figure 15 represents the relative importance of the input variables for the two possible outcomes when the dataset is considered with respect to the location of the accident being either in a section or an intersection. The type of accident is the most important variable, and in fact cluster analysis revealed that, for example, single-vehicle accidents generally occur in sections and front-to-side collisions generally take place in intersections. The age of the victim is also important, as young drivers tend to have mostly accidents in sections while experienced drivers were involved often in intersections. The crossing modality of pedestrians, the zone of the accident, the road width and the use of seatbelts also help discerning accidents occurring in either sections or intersections.

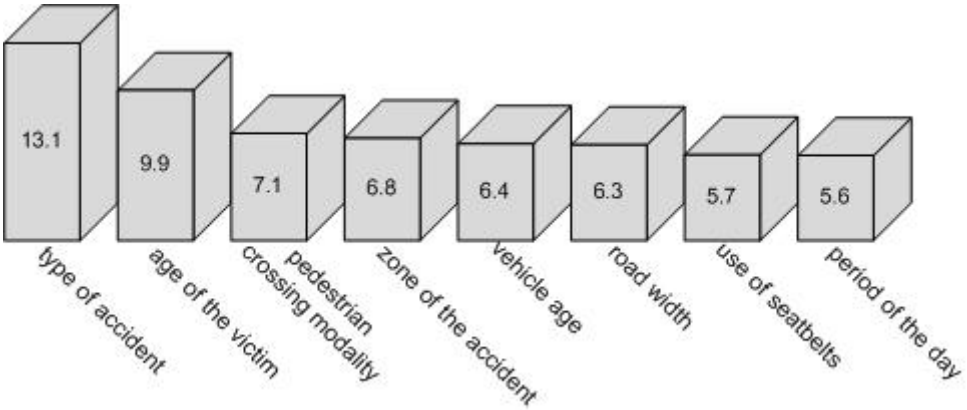

Figure 15 - Relative importance of input variables for sections/intersections accidents

Table 3 illustrates the confusion matrix between the predicted and expected outcome of the neural network. The correctness of the prediction of the neural network with two output neurons corresponding to either section or intersection location is 78.5%, which indicates that almost four out of five of the outputs for the tested half of the dataset are correctly predicted by the model trained with the other half of the dataset.

Table 3 - Confusion matrix for sections/intersections accidents

	Sections	Intersections
Sections	1070	272
Intersections	114	337

Figure 16 represents the relative importance of the input variables for the two possible outcomes when the dataset is considered with respect to the location of the accident being either in urban or rural areas. Unlike for the previous neural networks, the most important variables have similar values of relevance: type of accident, age of the victim, road width, zone of the accident and median have importance varying between 9.5 and 8.9%, which indicates that the distinction between urban and rural accidents is related to a combination of factors. For example, cluster analysis suggests that typical rural accidents are single-vehicle collisions where young drivers lost control of the car in narrow roads without separation in the north of the country.

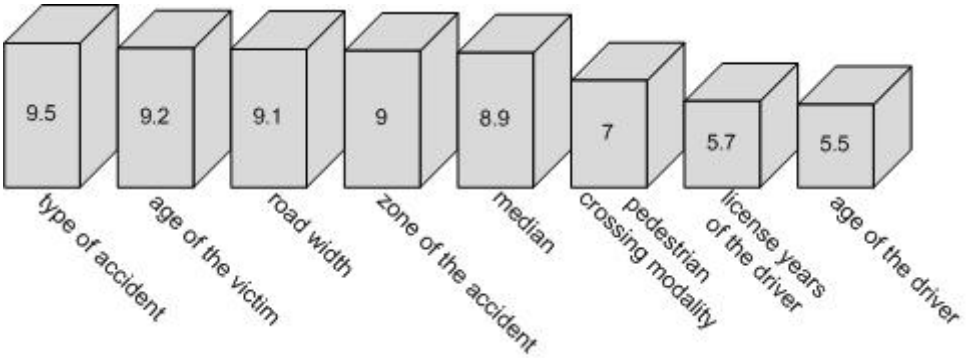

Figure 16 - Relative importance of input variables for urban/rural accidents

Table 4 shows the confusion matrix between the predicted and expected outcome of the neural network. The correctness of the prediction of the neural network with two output neurons corresponding to either urban or rural accidents is 79.8%, which suggests that around four out of five of the outputs for tested half of the dataset are correctly predicted by the model trained with the other half of the dataset.

Table 4 - Confusion matrix for urban/rural accidents

	Urban	Rural
Urban	585	164
Rural	199	845

Factor determination per type of victim

Figure 17 represents the relative importance of the input variables for the four possible outcomes when the dataset is considered with respect to the type of victims: pedestrians, car drivers, car passengers and two-wheel vehicle users. Four variables appear extremely important, and are all related to the age and social of drivers and victims. In fact, cluster analysis suggests that pedestrian victims are either teenagers or elderly, car drivers and car passengers are young but sometimes also middle-aged, two-wheel vehicle users are young but sometimes cyclists are older. Moreover, Jewish and non-Jewish drivers and victims are typically distinguished when describing which fatalities result from the accidents, and this explains the extremely high importance of the social groups.

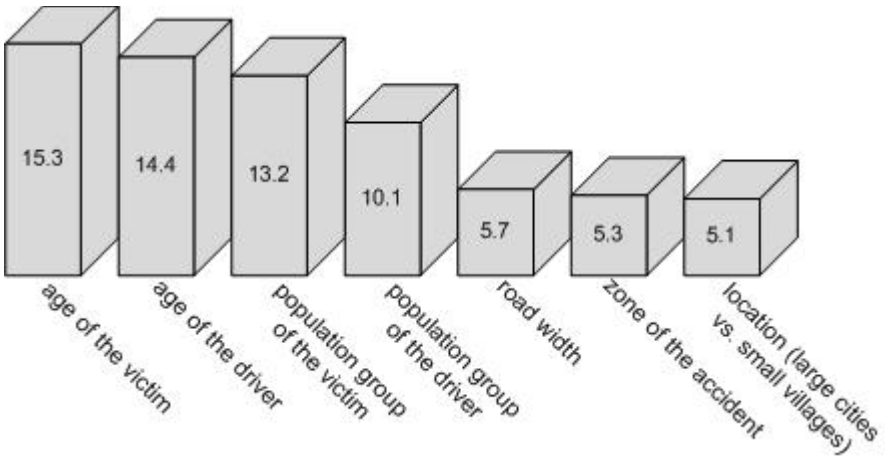

Figure 17 - Relative importance of input variables for type of victims

Table 5 presents the confusion matrix between the predicted and expected outcome of the neural network. The correctness of the prediction of the neural network with four output neurons corresponding to the four types of victims is 71.4%, which reveals that more than two thirds of the outputs for the tested half of the dataset are correctly predicted by the model trained with the other half of the dataset. Pedestrian and two-wheel vehicle users are the most correctly predicted, while the error is between car driver and car passenger fatalities.

Table 5 - Confusion matrix for type of victims

	Pedestrian	Car driver	Car passenger	Two-wheel vehicle user
Pedestrian	443	53	42	73
Car driver	26	377	104	21
Car passenger	23	93	298	19
Two-wheel vehicle user	21	18	19	163

SUMMARY AND CONCLUSIONS

This study focuses on the recognition of accident patterns and the determination of major accident factors to answer an increasing need of designing preventive measures, addressing specific situations and targeting specific social groups with the ultimate objective of reducing the annual number of traffic fatalities and accidents. This study analyzes fatal accidents in Israel and summarizes relevant elements that contribute to define specific accident patterns and major factors. The implementation of data mining techniques is successful in the recognition of accident patterns and determinants. The partition of the datasets helps obtaining less predictable solutions than the traditional patterns uncovered in similar studies. Note that the accident patterns for fatal accidents do not intend to predict that the combination of the factors obtained in the cluster analysis will result in a fatal accident, but rather to provide a classification of the fatal accidents and suggest recurrent aspects.

After the implementation of Kohonen neural networks, five accident patterns are recognized among the large number of clusters resulting from the analysis of the different partitions. After the implementation of Kohonen and feed-forward back-propagation neural networks, major crash determinants are validated and described. Demographic characteristics of both victims and drivers are the most relevant factors. With respect to the type of accident, the age of the victim is the most important variable, as Kohonen networks uncovered that children, teenagers and elderly are mostly victims of pedestrian accidents, young drivers up to 35 years old are mainly victims of single-vehicle crashes and experienced drivers are generally victims of multiple-vehicle collisions. With respect to the location, the age is again important, as cluster analysis revealed that young drivers are mostly involved in accidents in road sections while experienced drivers are involved often in intersections. Further, the social group of both drivers and victims always plays a significant role in the distinction of the clusters, as Jewish and Non-Jewish population are generally positioned at the opposite sides of the Kohonen maps obtained at convergence of the unsupervised learning process. Other relevant characteristics are the conditions of the road, the location in either urban or rural areas and in either sections or intersections, and the period of the day. Moreover, some combinations of variables appeared deadly mixtures explaining several accidents: mostly the young age of drivers, the low seatbelt compliance, the new vehicles and the narrow rural roads characterize large amount of crashes.

This study helps uncovering interesting accident patterns that safety campaigns should be differentiated according to age and social group of the target population, and should emphasize the differences between day and night driving conditions. Countermeasures should address with different interventions urban and rural areas, and most likely rethink traffic control systems or repair them where failing in intersections. This study also helps individuating some data issues that should be addressed in the future. The dataset of fatal accidents from police records contained several variables that do not play a role in the recognition of accident patterns because of the dominance of one category over the others. And if reporting good weather conditions is understandable given the climate conditions of the country throughout the year, always reporting impeccable conditions of the infrastructure is less comprehensible. Even a not scientific look at the conditions of surfaces, shoulders and median barriers of Israeli roads suggests that stating that the infrastructure is always perfect appears unlikely to be correct. Moreover, relevant information about the use of alcohol or drugs, or about estimated speed at the moment of the impact, could further improve the quality of the data and significantly add to the discussion about accident patterns. Clearly, data collection regarding traffic accidents requires more attention and precision by the police as well as integration of additional information that could help providing further insight into the major factors of accidents in the country.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the helpful comments of Ayala Cohen and Tsippy Lotan. This study was conducted when the first author was affiliated with the Ran Naor Road Safety Research Center at the Technion – Israel Institute of Technology.

REFERENCES

Berg, H. Y., N. P. Gregersen and L. Laflamme (2004). Typical patterns in road-traffic accidents during driver training: an explorative Swedish national study. Accident Analysis and Prevention, 36 (4), 603-608.

Chang, L. Y. and W. C., Chen (2005). Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research, 36 (4), 365-375.

Factor, R., Mahalel, D., Yair, G. (2008). Inter-group differences in road-traffic crash involvement. Accident Analysis and Prevention, 40 (6), 2000-2007.

Fleury, D. and T. Brenac (2001). Accident prototypical scenarios, a tool for road safety research and diagnostic studies. Accident Analysis and Prevention, 33 (2), 267-276.

Fontaine, H. and Y. Gourlet (1997). Fatal pedestrian accidents in France: a typological analysis. Accident Analysis and Prevention, 29 (3), 303-312.

Geurts, K., G. Wets, T. Brijs and K. Vanhoof (2003). Profiling of high-frequency accident locations by use of association rules. Transportation Research Record, 1840, 123-130.

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

Hasselberg, M., M. Vaez and L. Laflamme (2005). Socioeconomic aspects of the circumstances and consequences of car crashes among young adults. Social Science Medicine, 60 (2), 287-295.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59-69.

Kohonen, T. (2001). Self-Organizing Maps (3rd extended edition). Springer & Verlag, Berlin Heldelberg, Germany.

Laapotti, S. and E. Keskinen (2004). Has the difference in accident patterns between male and female drivers changed between 1984 and 2000? Accident Analysis and Prevention, 36 (4), 577-584.

Prato, C. G., S. Bekhor, A. Galtzur, D. Mahalel and J. N. Prashker (2010). Exploring the potential of data mining techniques for the analysis of accident patterns. Proceedings of the 12th World Conference on Transport Research, Lisbon, Portugal.

Preusser, D. F., A. F. Williams and R. G. Ulmer (1995). Analysis of fatal motorcycle crashes: crash typing. Accident Analysis and Prevention, 27 (6), 845-851.

Retting, R. A., A. F. Williams, D. F. Preusser and H.B. Weinstein (1995). Classifying urban crashes for countermeasure development. Accident Analysis and Prevention, 27 (3), 283-294.

Retting, R. A., J. Williams and S. I. Schwartz (2000). Motor vehicle crashes on bridges and countermeasure opportunities. Journal of Safety Research, 31 (4), 203-210.

Reed, R. D. and R. J. Marks (1999). Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks. The M.I.T. Press, Cambridge, MA.

Skyving, M., H. Y. Berg and L. Laflamme (2009). A pattern analysis of traffic crashes fatal to older drivers. Accident Analysis and Prevention, 41 (2), 253-258.

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

Wang, Y., M. Hasselberg, Z. Wu and L. Laflamme (2008). Distribution and characteristics of road traffic crashes in the Chaoyang District of Beijing. Accident Analysis and Prevention, 40 (1), 334-340.

Wong, J. T. and Y. S. Chung (2007). Rough set approach for accident chains exploration. Accident Analysis and Prevention, 39 (3), 629-637.

Wong, J. T. and Y. S. Chung (2008). Analyzing heterogeneous accident data from the perspective of accident occurrence. Accident Analysis and Prevention, 40 (1), 357-367.