TRANSPORT
ISSN 1648-4142 / eISSN 1648-3480
2017 Volume 32(3): 272–281
doi:10.3846/16484142.2014.915581

# IDENTIFYING THE KEY RISK FACTORS OF TRAFFIC ACCIDENT INJURY SEVERITY ON SLOVENIAN ROADS USING A NON-PARAMETRIC CLASSIFICATION TREE

Vesna Rovšek[1], Milan Batista[1], Branko Bogunović[2]

[1]*Faculty of Maritime Studies and Transport, University of Ljubljana, Slovenia*
[2]*Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Slovenia*

Submitted 8 October 2013; accepted 6 March 2014;
first published online 17 June 2014

**Abstract.** From both a practical and economic point of view, road transport meets almost all the requirements of modern life, but it is also a source of numerous negative effects, including traffic accidents. In order to design a safe transport system and achieve the 'zero vision' goal – no serious injuries or fatalities in traffic accidents – there is a growing need for a systematic approach to this problem. Prior to the assessment of any accident prevention measure it is necessary to identify the most important factors and significant patterns which affect the severity of accidents and injuries. In this study, the crash data from Slovenia pertaining to the period 2005–2009 were analysed with a Classification and Regression Tree (CART) algorithm, one of the most widely applied data mining technique when analysing a large amount of data with several independent quantitative or qualitative variables. Before building a non-parametric classification tree, the data were split into three totally separate subsets, the training set, the testing set, and the evaluation set. Moreover, using the Variable Importance Measure (VIM) the factor of influence of nine independent variables on the target variables were calculated. The results confirm that traffic accidents and injuries on Slovenian roads are caused by a combination of factors, the most important of them being human error, or more precisely, speeding and driving in the wrong lane.

**Keywords:** road traffic accident; injury severity; variable importance; classification error; classification tree.

## Introduction

Mobility is an important component of the socio-economic development of individuals and society. Ease of movement is essential for the transport of goods and people. From a practical and economic point of view roads allow almost all the requirements of modern life. However, road transport is the source of many negative effects, such as environmental pollution, traffic jams and, most significant, traffic accidents. These have direct or indirect impact on the economy through losses of time, money and, of course, human costs. The analysis of the existing situation and the defining of the problem are the primary steps in forming road safety strategy (Vujanić *et al.* 2013).

Despite the European objectives, 'zero road traffic victims' and the apparent decline of victims on the Slovenian roads, the numbers of accidents remain high. Among the 27 member countries of the European Union, Slovenia was ranked third in road fatalities per one million inhabitants (146) in 2007. During the same period (2007), the Netherlands and the United Kingdom recorded significantly fewer deaths: 48 and 55 per million inhabitants (ETSC 2008). Moreover, Slovenia's average reduction in road fatalities did not exceed 1.6% from 2001–2007 per annum while 'France, Portugal and Luxemburg have reduced road deaths by an average of more than 8% per year, and are well on their way to hitting the EU target at national level' (ETSC 2008). Individual participants in traffic and the general Slovenian society share a poor traffic culture, accepting of non-observance of the traffic code. Statistical results for road accidents in Slovenia show that speeding and driving in the wrong lane are the most important factors for traffic accidents, especially those with fatal consequences. In 2007, speed-

Corresponding author: Vesna Rovšek
E-mail: *vesna.rovsek@fpp.uni-lj.si*



Taylor & Francis
Taylor & Francis Group

ing contributed to almost one half (48%) and driving in the wrong lane to 34% of all fatalities on Slovenian roads (Ministrstvo za notranje zadeve 2010).

Over the last two decades a major effort was put into a systematic approach for defining road safety strategy (Dell'Acqua *et al.* 2011; Vujanić *et al.* 2013) and the research of traffic accidents – understanding and determining circumstances of accidents and the key factors that influence the severity of injuries caused by traffic accidents (Chong *et al.* 2005). From a methodological point of view, a variety of studies applied, simple (Holubowycz *et al.* 1994; Malliaris *et al.* 1996; Öström, Eriksson 2001) or multivariate (Zhang *et al.* 2000; Bédard *et al.* 2002; Valent *et al.* 2002; Hijar *et al.* 2004; Yau 2004; Çela *et al.* 2013), statistical methods. For example, Wondwossen (1999) studied the correlation of car accidents in Addis Ababa. He used the chi-square test and logistics regression analysis. The survey showed that in addition to other variables such as light conditions of the road, the main cause of physical damage is due to ignoring non-priority pedestrians. Dissanayake and Lu (2002) used regression analysis to identify factors that influence the severity of injuries in the case of accidents of older drivers in a fixed object (vehicle–pedestrian accidents). Al-Ghamdi (2002) studied with the help of logistic regression the impact of individual variables on the severity in Saudi Arabia. Results showed that the location and cause of the accident are important factors in increasing the severity of the accident. Singleton *et al.* (2004) carried out a logistic regression of severity of injury in order to examine the factors affecting the high degree of severity of injury in traffic accidents with a damaged vehicle. After analysing data on traffic accidents in Kentucky (from 2000 to 2001) they concluded the risk factors for a high degree of severity of injury: age of driver, sex, safety belt use and impaired state of the driver (due to alcohol consumption). Hanrahan *et al.* (2009) used logistic regression to determine the relationship between age and severity of injury for drivers in road transport. They used data on accidents in Wisconsin (from 2002 to 2004) to examine the 602964 drivers involved in vehicle accidents. It was observed that the highest risk for serious injury or fatality was among older drivers, particularly those over 85 years of age.

The applications of non-parametric modelling techniques to analyse traffic safety problems have been relatively few. Nevertheless, in recent years the use of data mining has increased considerably. It has been successfully used in determining circumstances of accidents and the key factors that influence the formation and the degree of severity of accidents and injuries. Artificial Neural Networks (ANN) and support vector machine, non-linear regression and classification methods, decision trees and decision rules, are the most popular data mining techniques. In the field of safety analysis, some studies introduced ANN to analyse traffic safety problems. Mussone *et al.* (1999) used artificial neural networks to analyze vehicle accidents that occurred at intersections in Milan. The model had 10 input nodes for eight variables. The output node ('accident index') was calculated as the ratio between the number of accidents at a given intersection and at the most dangerous intersection. Results showed that the highest accident index for the running over of pedestrians occurred at non-signalized intersections at night time. Abdelwahab and Abdel-Aty (2001) have used ANN to model the relationship between the severity of injuries to the driver and the number of accident factors. A similar study was conducted by Delen *et al.* (2006), who used a series of ANN for nonlinear modelling of potential links between the seriousness of injuries and factors related to the accident. Tree-based models are another non-parametric method frequently applied to analyse injury severity problems (Sohn, Shin 2001; Sohn, Lee 2003; Getnet 2009). For example, Chang and Wang (2006) have developed a Classification and Regression Tree (CART) model to demonstrate the relationships between injury severity and driver/vehicle characteristics, highway/environmental variables and accident variables. Using data on traffic accidents in Taipei (Taiwan) for the year 2001 they concluded that the most important variable associated with the severity of accidents was the type of vehicle. Yan and Radwan (2006) used a decision tree model in combination with the method quasi-exposure risk assessment (Briefings Quasi-Induced Exposure) to make an analysis on the relationship between rear-end crashes that have occurred at signalised intersection and a set of potential traffic risk factors. By analysing a database of accidents which have occurred in Florida in 2001, they concluded that such accidents usually occur due to higher speed limits (45÷55 mph) and are over-presented during day time, on wet and slippery roads, and the propensity is higher for males and drivers younger than 21 years old. Pande and Abdel-Aty (2006) identified the important parameters that lead to accidents on freeways related to changing lanes in one of the studies dealing with a classification tree based variable selection procedure. They used the traffic control data collected by loop detectors. Kunt *et al.* (2011) focused on predicting the severity of freeway traffic accidents by employing twelve accident related parameters in a Genetic Algorithm (GA), pattern search and ANN modelling methods.

In several studies, authors concentrated on one particular or only a few risk factors. Among them, most studies have focused on a specific group of road users (Zhang *et al.* 2000; Valent *et al.* 2002; Zajac, Ivan 2003) or certain types of vehicles (Quddus *et al.* 2002; Ulfarsson, Mannering 2004) or particular accident characteristics and injury severity (Malliaris *et al.* 1996; Lee, Mannering 2002; Chang, Wang 2006; Kunt *et al.* 2011; Castro *et al.* 2013; Jiang *et al.* 2013). In addition, numerous studies have attempted to identify the effect of a restraint device (Bédard *et al.* 2002; Valent *et al.* 2002) or explore the impact of drinking and driving (Zajac, Ivan 2003; Keall *et al.* 2004) on the injury severity levels.

However, in Slovenia such studies are still lacking. In order to ensure the long-term safety goal: zero serious injuries and zero fatalities caused by road accidents, it is necessary to systematically identify the key risk factors that affect the severity of accidents and injuries. There-

fore, a CART technique was implemented for modelling Slovenian traffic accident data. In line with this, the study raised two hypotheses: H1: 'A key risk factor for traffic accidents with minor and severe and fatal injuries on Slovenian roads in the period 2005–2009 is human error'; H2: 'The most important indicators that cause fatal accidents are inappropriate speed and the driving on the wrong side of the road'. Furthermore, the design of the model was developed for identifying as well as predicting the most important factors which affect injury severity due to road accidents. Identifications of the factors of accidents with an emphasis on those that cause the most serious consequences would enable the ultimate elimination of fatalities and severe injuries.

## 1. Methodology

Various decision tree algorithms such as CART (Breiman *et al.* 1984), ID3 (Quinlan 1986), C4.5/C5.0 (Quinlan 1992), CHAID (Kass 1980), MARS (Friedman 1991) produce trees that differ in the number of splits allowed at each level of the tree, how these splits are selected when the tree is built, and how the tree growth is limited in order to prevent over-fitting.

CART methodology was developed in the 1980's by Breiman and his colleagues, and has become one of the most popular and widely applied data mining algorithms (Breiman *et al.* 1984; Berry, Linoff 1999; Rokach, Maimon 2008), particularly when it comes to analysing large amount of data with several independent quantitative or qualitative variables (Breiman *et al.* 1984). Furthermore, CART does not require any pre-defined underlying relationship between dependent and independent variables and has been shown to be a powerful tool for dealing with prediction and classification problems (Chang, Wang 2006). It is known as a binary recursive partitioning, whereas the root node (also known as parent node) is always split into exactly two internal nodes (also known as child node) and recursive because the process can be repeated by addressing each of the internal nodes as the root node until it can find no more useful splits (Breiman *et al.* 1984; Tesema *et al.* 2005).

Classification trees are used where for each instance in the training set (also known as learning set) the class is already known in advance (Breiman *et al.* 1984; Rokach, Maimon 2008). Classes in the training set can be user-defined or calculated in accordance with some splitting rule – rule for division training set into smaller parts.

The maximum homogeneity of the internal nodes is determined with an impurity function $i(t)$. Since the impurity of root node $t_r$ is constant for any of the splits and the possible division $x_j \le x_j^R$, $j = 1, ..., M$, the maximum homogeneity of the left and right internal nodes will be equivalent to the maximization of change of impurity function $\Delta i(t)$ (Breiman *et al.* 1984; Hastie *et al.* 2011; Timofeev 2004):

$$\Delta i(t) = i(t_r) - E\left[i(t_i)\right], \tag{1}$$

where: $t_i$ denotes the left and right internal node of the root node $t_r$.

Considering that $P_l$ and $P_r$ are the probabilities of the left and right nodes, we get:

$$\Delta i(t) = i(t_r) - P_l i(t_l) - P_r i(t_r), \tag{2}$$

where CART at each node solves the following maximization problem:

$$\underset{x_j \le x_j^R, j=1,...M}{\arg\max} \left[ i(t_p) - P_l i(t_l) - P_r i(t_r) \right]. \tag{3}$$

Equation (3) indicates that the CART method is searching through all possible values of all variables in a matrix $X$ for the best split question $x_j \le x_j^R$ that will maximize the change of impurity criteria $\Delta i(t)$ (Breiman *et al.* 1984; Hastie *et al.* 2011).

The Gini index (also known as the Gini splitting rule) is the most commonly used splitting rule which finds in the training set the largest set of class and separates it from the remaining data (Gelfand *et al.* 1991; Berry, Linoff 1999). Moreover this algorithm is suitable for data that contains errors in measurement (noisy data). Thus, the impurity function $i(t)$ will be defined with the Gini index (Breiman *et al.* 1984; Chong *et al.* 2005; Rokach, Maimon 2008):

$$i(t) = \sum_{k \ne l} p(k|t)\, p(l|t), \tag{4}$$

where: $k$, $l$ – class index; $p(k|t)$ – conditional probability of the class $k$, assuming that we are in the node $t$.

An even more important influence on the final tree has its optimization (Hancock *et al.* 1996; Zantema, Bodlaender 2000). The maximum trees, in other words, trees with very high complexity (when terminal nodes contain observations of only one class) have to be pruned before being used for classification of independent data (Breiman *et al.* 1984; Berry, Linoff 1999). Thus, at a certain depth the model does not adapt (overfitting) over the training set (Gelfand *et al.* 1991; Cios *et al.* 1998). To find the optimal tree size, one can use a cross-validation procedure, which is also most often used method for tree pruning. The cross-validation process is based on finding the optimal ratio between the complexity of the tree and the misclassification error (Berry, Linoff 1999). This can be achieved by using the cost-complexity function (Breiman *et al.* 1984; Timofeev 2004; Rokach, Maimon 2008):

$$R_\alpha(T) = R(T) + \alpha(\tilde{T}) \to \min_T, \tag{5}$$

where: $R(T)$ – misclassification error of the tree $T$; $\alpha(\tilde{T})$ – measure of complexity, which depends on the $\tilde{T}$; $\tilde{T}$ – total number of terminal nodes in the tree; $\alpha$ – the parameter is found through sequential in-sample testing when a part of the training set is used to build the tree, while the rest of the data is taken as a testing set; the procedure is repeated several times on randomly selected training and testing sets.

Breiman *et al.* (1984) devised a Variable Importance Measure (VIM) for trees, which may be applied as a criterion to select a subset of variables that have sig-

nificant importance in predicting the target variable. The variable importance for predictor variable $x_j$ in relation to the final tree $T$ is the weighted average of the reduction in the Gini impurity measure achieved by all splits using variable $x_j$ across all internal nodes of the tree. The formula for the importance of variable $x_j$ is given by the following (Mussone *et al.* 1999):

$$VIM\left(x_j\right)=\sum_{t=1}^{\tilde{T}}\frac{n_t}{N}\Delta\text{Gini}\left(S\left(x_j,t\right)\right), \quad (6)$$

where: $\frac{n_t}{N}$ – proportion of the observation in the data set that belong to node $t$; $N$ – total number of observation in training set; $\Delta\text{Gini}\left(S\left(x_j,t\right)\right)$ – reduction of Gini index on the basis of variable $x_j$.

## 2. Data for Analysis

A study was conducted based on daily reports of the accident data from the databases of the police. Identifying key factors affecting the severity of injuries in road accidents was based on Slovenian roads for the period from 2005 to 2009. However, data have some limitations regarding alcohol as a factor that influences the severity of injury effected by a traffic accident, since in the database alcohol was regarded only as whether an alcohol test was carried out or not.

Based on extensive data research and observations some variables were omitted from the database due to irrelevance (e.g. administrative unit area where the accident occurred, the date of accident, etc.). Furthermore, some additional variables were removed: the role of people in a car accident (who caused the traffic accident or participant), the person's age, gender and driving licence period. These variables are very important in terms of identifying key risk factors, but more appropriate when considering only the drivers that caused the accident. In order to determine the key risk factors particularly for those accidents that cause injuries and fatalities, we analysed the causers as participants of accidents. Otherwise, a significant amount of the data would be lost on the consequences or weight of injuries for each variable.

The data set included the period from January 2005 to December 2009, during which 220578 accidents occurred. The total number of people with minor, severe or fatal injuries was 72518 (Table 1). For each record of individual persons (both drivers and fellow passengers) we assigned 9 properties or independent variables. For the dependent variable, we determined severity of injury of persons involved in traffic accidents. This means that every record in the database casualty corresponded to the categorical output variable – the severity of injuries. This method of data analysis has allowed us to search for the causes of many accidents and at the same time injuries and fatalities.

For calculation we developed a special purpose program implemented in MATLAB. However, statistical analysis of the model set was performed with *Statistics Toolbox*.

Firstly, the data were split into three totally separated subsets. 80% of the original data were used in the training phase. Testing and evaluation data sets each contained 10% of the original data. Before construction of the model, part of the model data was separated from the rest and held back. This set is called the evaluation set. The model was trained using pre-classified data in the training set. The testing set was used in order to prevent the model memorizing the training set (to avoiding over-fitting) and for ensuring a more general model, which will also work better with unseen data. Because both sets (training and testing) were used to build the model, they cannot be used to evaluate its effectiveness; it was therefore necessary to use the third subset, the holdout evaluation set, distinct from the previous two.

## 3. Results

The Gini splitting criterion is used in this study. The process begins with a training set and the building of an entire classification tree that correctly classified every single record. In order to verify accuracy of the classification the confusion matrix was used. Results are shown in Table 2. It can be calculated that around 83% of the labels were correctly classified (diagonal matrix). Thus, classification error, as measured by the learning data on an un-pruned tree, is 17%. Furthermore, it is evident that 92% of fatalities and 82% of non-fatal injuries were correctly classified.

Since we wanted to find key risk factors for injury and death, we used, on an un-pruned tree, one of the products of CART algorithm, namely the Variable Importance Measure (VIM) – see Table 3. The impact factor indicates which variable was of more importance for an accident with injury and fatality. The Slovenian case study recognised 'Contributing circumstance' as the most important variable for accidents with injury and fatality, such that it is more important by approximately 2 times than the second variable 'Collision type'. Next were 'Road category' and 'Weather condition'.

However, in the case of the un-pruned tree certain constraints arose. Although such a tree provides a good description of the training data, it is too large and complex for detailed analysis. Furthermore, it is unlikely to generalise unseen data. That is why the testing set was used to prune the tree and to search for the optimal size of the tree. The size of training set $(n-1)/n$ and testing set $1/n$ (where: $n$ – number of randomly selected subsets for testing) was determined in accordance with the procedure for $k$-fold cross-validation (Kohavi 1995; Duin, Tax 2004) since it was subsequently used to determine the optimal relationship between the complexity of the tree and the misclassification error. Thus, in the process of determining the optimal size of the model, calculation of the training error for different subsets of the original (full) tree was performed, and then through a process of 10-fold cross-validation the testing error was calculated for 10 different randomly selected and equal sized subsets of the original tree. Fig. 1 shows how the training error is very optimistic – the error decreases with the growing number of the terminal nodes.

Table 1. Description of the variables included in the analysis

| Independent variables | Label | Severity of injury | |
|---|---|---|---|
| | | Fatality | Injury (minor, severe) |
| | | 1198 | 71320 |
| ***Seat Belt/Helmet*** | | | |
| Used | 3 | 1014 | 64259 |
| Not used | 4 | 184 | 7061 |
| ***Contributing Circumstance*** | | | |
| Inappropriate speed | 5 | 516 | 19338 |
| Wrong side of the road | 6 | 317 | 10702 |
| Ignoring right of way | 7 | 172 | 16673 |
| Inappropriate movement of vehicles | 8 | 31 | 4248 |
| Improper overtaking | 9 | 65 | 2414 |
| Inadequate safety distance | 10 | 4 | 11942 |
| Other (irregularities in the road or cargo, irregularities on the vehicle, violation of pedestrian, other) | 11 | 93 | 6003 |
| ***Collision Type*** | | | |
| Head-on | 12 | 342 | 13993 |
| Sideway | 13 | 225 | 16673 |
| Rear-end | 14 | 65 | 14312 |
| Fixed object | 15 | 183 | 5427 |
| Sideslip | 16 | 22 | 3139 |
| Overturn | 17 | 151 | 7024 |
| Other (collision with pedestrian, collision with animal, impact with unmovable vehicle, other) | 18 | 210 | 10752 |
| ***Road Category*** | | | |
| Freeway (freeway, motorway) | 19 | 159 | 4804 |
| Regional (highway I, II category and regional I, II, III category) | 20 | 586 | 23787 |
| Local (local, tourist) | 21 | 102 | 4277 |
| Settlement (settlement with/without street system) | 22 | 351 | 38452 |
| ***Lighting Condition*** | | | |
| Day light | 23 | 746 | 52200 |
| Dark | 24 | 452 | 19120 |
| ***Weather Condition*** | | | |
| Clear | 25 | 700 | 38554 |
| Rain (rain, hail) | 26 | 83 | 7851 |
| Fog | 27 | 11 | 720 |
| Snow | 28 | 13 | 1552 |
| Other (cloudy, wind, unknown) | 29 | 391 | 22643 |
| ***Road Surface Condition*** | | | |
| Dry | 30 | 927 | 50121 |
| Wet (wet, muddy) | 31 | 211 | 15966 |
| Slippery | 32 | 40 | 2964 |
| Icy | 33 | 11 | 1610 |
| Other | 34 | 9 | 659 |
| ***Road Type*** | | | |
| Rough asphalt | 35 | 836 | 44589 |
| Uneven asphalt | 36 | 11 | 340 |
| Polished asphalt | 37 | 318 | 25065 |
| Other (macadam, other) | 38 | 33 | 1326 |
| ***Traffic Condition*** | | | |
| Normal | 39 | 704 | 42730 |
| Dense | 40 | 77 | 10102 |
| Sparse | 41 | 403 | 16518 |
| Traffic jam | 42 | 1 | 195 |
| Unknown | 43 | 13 | 1775 |

*Note*: under the 'other' are combined indicators, which have, according to our findings, a negligible share within each variable and the category that police identified as other or unknown. This allows the continued search for risk factors in accordance with the target.

Table 2. Classification matrix for the fatality and injury victims of road accidents in the case of an un-pruned tree and training data

| | | Predicted class | |
|---|---|---|---|
| | | Fatality | Injury |
| Actual class | Fatality | 981 | 90 |
| | Injury | 11199 | 52996 |

Table 3. Importance of variables in the case of an un-pruned tree

| VIM | Independent variable |
|---|---|
| 0.2791 | Contributing circumstance |
| 0.1354 | Collision type |
| 0.1199 | Road category |
| 0.1046 | Weather condition |
| 0.0994 | Traffic condition |
| 0.0846 | Road type |
| 0.0789 | Road surface condition |
| 0.0560 | Lighting condition |
| 0.0421 | Seat belt/Helmet |
| 1.0000 | Sum |

On the other hand, the testing error after a certain point increases with an expanding tree (number of terminal nodes). The point 'best decision' (shown in the figure with green) represents the optimal relationship between complexity and misclassification error tree, the number of end nodes (19), where the least error is achieved on the testing set. For the maximum tree, the misclassification error will be minimum (equal to 0) and the number of terminal nodes will be maximum, but complex decision trees poorly perform on independent data. The small tree gets a much lower penalty for its size, but their predicting abilities are naturally limited.

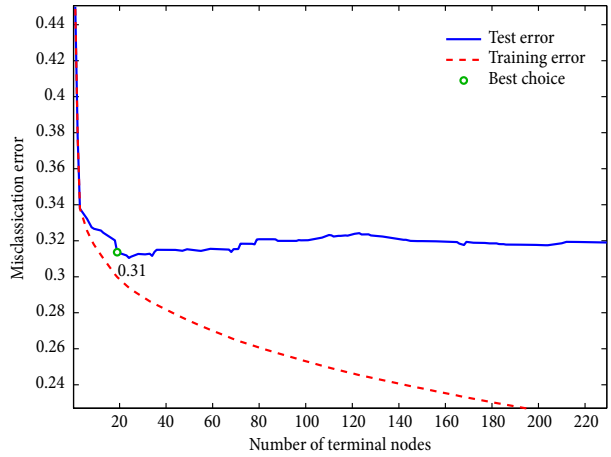


Fig. 1. Relationship between tree complexity (number of terminal nodes) and error rate on training set and testing set

Fig. 2 shows the pruned tree which has a relatively small testing error and is also transparent. The best compromise between size (complexity) of the tree and its ability to classify new data was found in the case of a pruned tree with 19 terminal nodes and 70% accuracy of classification of first seen data. It follows that the pruned tree can be used as a model to identify key patterns that cause injuries and fatality in traffic accidents on Slovenian roads, as well as for classifying new data.

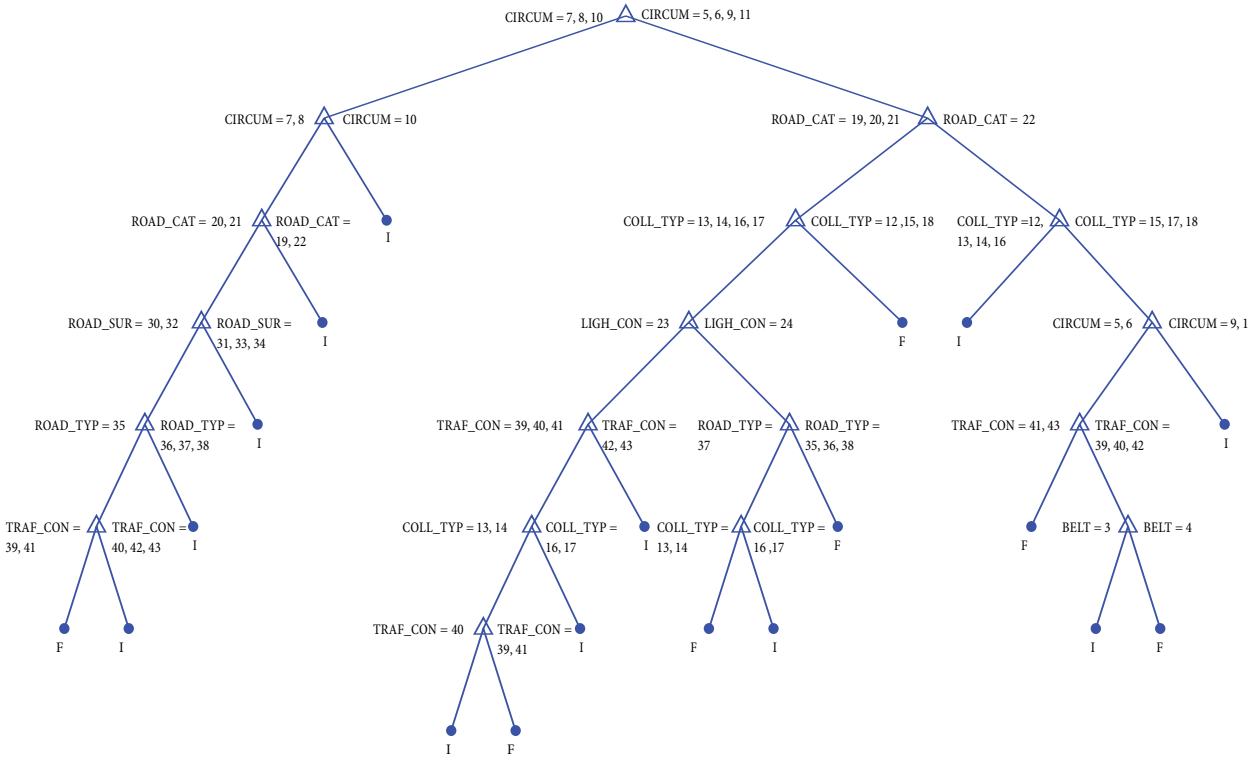


Fig. 2. Optimal size of the classification model to identify key patterns that cause injuries and fatalities in traffic accidents on Slovenian roads: I – injury; F – fatality

Even in the case of a pruned tree, the significance of individual independent variables was calculated (Table 4) and sorted by importance factor to the target variable. Likewise in this case, the 'Contributing circumstance' scores high as a major factor in accidents with injuries and fatalities. It is more important by approximately 3 times than the second variable 'Road category' and 6 times than the third one 'Collision type'.

The error rate on the hold-out evaluation set was already observed during the search for the best pruned tree. To ensure the credibility of the assessment and quality of model, the confusion matrix was also calculated. In order to estimate a more credible performance, we used an evaluation set, which was during the construction of the tree 'put aside'. This data set was independent from the data set that was used for building the tree and represents for the model data not yet seen. Thus, with the matrix the actual (not seen) data with the predicted were compared and it could be determined how well the model works on the first seen data. From the classification matrix, shown in Table 5, it is evident that the model correctly classified 71% fatalities and 70% injured in traffic accidents. Furthermore, the pruned tree correctly classified around 70% of new cases; this confirms the findings regarding the classification of new data.

Table 4. Importance of variables in the case of a pruned tree

| VIM | Independent variable |
|---|---|
| 0.5974 | Contributing circumstance |
| 0.1865 | Road category |
| 0.0988 | Collision type |
| 0.0345 | Traffic condition |
| 0.0297 | Road type |
| 0.0228 | Road surface condition |
| 0.0210 | Lighting condition |
| 0.0093 | Seat belt/Helmet |
| 1.0000 | Sum |

Table 5. Classification matrix for the fatality and injury victims of road accidents in the case of a pruned tree and first seen data

| | | Predicted class | |
|---|---|---|---|
| | | Fatality | Injury |
| Actual class | Fatality | 90 | 37 |
| | Injury | 2158 | 4967 |

## 4. Discussion

Until recently, Slovenian researchers only analysed data with simple statistical manipulations, which have a limited ability to research, analyse and display new or unexpected patterns and relationships that are hidden in conventional databases. In order to ensure better road safety systematic analyses of the causes and consequences of accidents are needed. By identifying key risk factors the impact of individual objective factors as well as their interaction on the weight of traffic accidents can be determined. On the basis of these insights further deliberate and appropriate action can be taken. Data mining techniques provide a greater potential for identifying key risk factors of traffic accident injury severity and various interesting patterns that allow organizations strategic planning and decision making in their domains.

From both analyses of VIM, for the full and pruned tree, it was confirmed (Tables 3 and 4) that the most influential factor for injuries and fatalities in traffic accidents on Slovenian roads was 'Contributing circumstance'. In the case of the full tree the most important factor followed by variables 'Collision type' and 'Road category' were the same as in the case of the pruned tree, only reversed. The least impact had variables 'Light condition' and use of seat belt or helmet. For further research there is the interestingly position of the variable 'Weather' which in the case of the pruned tree had no impact on the target variable but in the case of the full tree is ranked in the top four. In our opinion, to determine the impact of independent variables on the target variable the order of precedence is more important, as shown in Table 3 (for un-pruned tree), since it involves an analysis of all the data and distribution of the importance of all indicators. VIM for pruned tree (Table 4) was primarily calculated to figure out how it behaves in the pruned tree and if it moves in line with our assumption.

Moreover, from Fig. 2 it can easily be distinguished that the CART algorithm picks the variable 'Contributing circumstance' as the most important for traffic accident injuries because it is splitting at the root node. Also, 'Road category' and 'Collision type' are the next critical variables in classifying injury severity in traffic accidents as these variables are chosen for splitting high up in the classification tree.

The impact of individual branches or key patterns of injury and fatality in traffic accidents on Slovenian roads was discerned by looking at the different levels of pruning. On this basis, the right side of the tree was proved to be the most influential. If the participants in road accidents are involved in circumstances where there is inappropriate speed, a car on the wrong side of the road, improper overtaking or 'other' and the road category is freeway, regional or local, the injury severity is most likely to be fatal. Driving through a settlement, if the collision type is collision with a fixed object, overturn or 'other' (e.g. collision with pedestrian), in combination with inappropriate speed or wrong side of the road and on the assumption sparse or unknown traffic condition develops in another important fateful pattern. The rightmost branch also confirmed that using a seat belt or helmet saves lives.

The previous statement confirmed hypothesis 2; namely, that for the participants in road accidents caused by inappropriate speed and driving on the wrong side of the road there is more probability to succumb to fatal injuries, as in the case of the other five indicators from the set of 'Contributing circumstance'. This followed the partial verification of hypothesis 1, that the key risk factors for traffic accidents with fatal injuries on Slovenian roads are drivers who drive arrogantly, carelessly and/

or ignore traffic rules, and thereby threaten other road users. In the case of the observation of the injuries such a conclusion was not made too quickly, as under item 11 ('other'), which is located on the most influential tree branch, joined the indicators that are not linked to inappropriate road user behaviour.

## Conclusions

1. The purpose of this study was to present and introduce the use of the CART algorithm for the formulation of better Slovenian road safety strategy; it helps to understand the characteristics of driver behaviour, road conditions, and traffic conditions, etc., which are causally associated with various degrees of injuries.
2. The data (from 2005 to 2009) were split into three subsets: the training set, the testing set and the evaluation set, which represented 80%, 10% and 10% of the original data, respectively. In this study 9 independent variables were assigned and two levels of injury severity: injury and fatality.
3. The Gini splitting criterion was applied in the CART method. The training set was used to build a classification tree with an accuracy of 83%. Furthermore, 92% of fatalities and 82% of non-fatal injuries were correctly classified. The most important risk factor for injuries and fatalities was 'Contributing circumstance' (0.28).
4. Due to the size and complexity of the classification tree, the testing set was applied to prune the tree and to search for the optimal size of the tree. Therefore, 10-fold cross-validation was used to determine the optimal relationship between the complexity of the tree and the misclassification error. The optimal relationship was determined with the size of 19 terminal nodes of the tree.
5. The pruned tree has a relatively small testing error with 70% accuracy of classification of the first time seen data. The model correctly classified 71% fatalities and 70% non-fatal injuries. The significance of individual independent variables was calculated. Likewise in the case of the pruned tree the 'Contributing circumstance' scores high as a major risk factor in accidents with injuries and fatalities. With the factor 0.60 it is more important by 3 times than the second variable 'Road category' (0.19). Moreover, with the help of pruning levels the most influential branch brought together 'inappropriate speed' and 'wrong side of the road' as the most important indicators for fatal injury.
6. In conclusion, we give directions and suggestions for further research to find solutions that would improve road traffic safety. In this area we should not ignore the sets of risk factors 'Collision type' and 'Road category' because the algorithm distributed them high up in the tree. It would be interesting to analyse the indicators that are grouped under item 11 ('other') and investigate which of these has the strongest impact on the dependent variable. This would clearly define the role of human error in those accidents with minor and severe injuries.

## References


Abdelwahab, H. T.; Abdel-Aty, M. A. 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections, *Transportation Research Record* 1746: 6–13. http://dx.doi.org/10.3141/1746-02

Al-Ghamdi, A. S. 2002. Using logistic regression to estimate the influence of accident factors on accident severity, *Accident Analysis & Prevention* 34(6): 729–741. http://dx.doi.org/10.1016/S0001-4575(01)00073-2

Bédard, M.; Guyatt, G. H.; Stones, M. J.; Hirdes, J. P. 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities, *Accident Analysis & Prevention* 34(6): 717–727. http://dx.doi.org/10.1016/S0001-4575(01)00072-0

Berry, M. J. A.; Linoff, G. S. 1999. *Mastering Data Mining: The Art and Science of Customer Relationship Management*. 1st edition, Wiley. 512 p.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. 1984. *Classification and Regression Trees*. 1st edition. Chapman and Hall/CRC. 368 p.

Castro, M.; Paleti, R.; Bhat, C. R. 2013. A spatial generalized ordered response model to examine highway crash injury severity, *Accident Analysis & Prevention* 52: 188–203. http://dx.doi.org/10.1016/j.aap.2012.12.009

Çela, L.; Shiode, S.; Lipovac, K. 2013. Integrating GIS and spatial analytical techniques in an analysis of road traffic accidents in Serbia, *International Journal for Traffic and Transport Engineering* 3(1): 1–15. http://dx.doi.org/10.7708/ijtte.2013.3(1).01

Chang, L.-Y.; Wang, H.-W. 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques, *Accident Analysis & Prevention* 38(5): 1019–1027. http://dx.doi.org/10.1016/j.aap.2006.04.009

Chong, M.; Abraham, A.; Paprzycki, M. 2005. Traffic accident analysis using machine learning paradigms, *Informatica* 29: 89–98.

Cios, K. J.; Pedrycz, W.; Swiniarski, R. W. 1998. *Data Mining Methods for Knowledge Discovery*. Springer. 495 p.

Delen, D.; Sharda, R.; Bessonov, M. 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks, *Accident Analysis & Prevention* 38(3): 434–444. http://dx.doi.org/10.1016/j.aap.2005.06.024

Dell'Acqua, G.; De Luca, M.; Mauro, R., 2011. Road safety knowledge-based decision support system, *Procedia: Social and Behavioral Sciences* 20: 973–983. http://dx.doi.org/10.1016/j.sbspro.2011.08.106

Dissanayake, S.; Lu, J. J. 2002. Factors influential in making an injury severity difference to older drivers involved in fixed object–passenger car crashes, *Accident Analysis & Prevention* 34(5): 609–618. http://dx.doi.org/10.1016/S0001-4575(01)00060-4

Duin, R. P. W.; Tax, D. M. J. 2004. Statistical pattern recognition, in C. H. Chen; L. F. Pau; P. S. P. Wang (Eds.). *Handbook of Pattern Recognition and Computer Vision*. World Scientific, Singapore.

ETSC. 2008. *Road Safety as a Right and Responsibility for All: A Blueprint for the EU's 4th Road Safety Action Programme 2010–2020*. European Transport Safety Council (ETSC), Brussels. 51 p. Available from Internet: http://www.etsc.eu/documents/Blueprint_for_a_4th%20Road_Safety_Action_Programme_ETSC_Sept%2008.pdf

Friedman, J. H. 1991. Multivariate adaptive regression splines, *The Annals of Statistics* 19(1): 1–141. http://dx.doi.org/10.1214/aos/1176347963

Gelfand, S. B.; Ravishankar, C. S.; Delp, E. J. 1991. An iterative growing and pruning algorithm for classification tree design, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(2): 163–174. http://dx.doi.org/10.1109/34.67645

Getnet, M. 2009. *Applying Data Mining with Decision Tree and Rule Induction Techniques to Identify Determinant Factors of Drivers and Vehicles in Support of Reducing and Controlling Road Traffic Accidents: the Case of Addis Ababa City*: MSc Thesis. Addis Ababa University, Addis Ababa, Ethiopia.

Hancock, T.; Jiang, T.; Li, M.; Tromp, J. 1996. Lower bounds on learning decision lists and trees, *Information and Computation* 126(2): 114–122. http://dx.doi.org/10.1006/inco.1996.0040

Hanrahan, R. B.; Layde, P. M.; Zhu, S.; Guse, C. E.; Hargarten, S. W. 2009. The association of driver age with traffic injury severity in Wisconsin, *Traffic Injury Prevention* 10(4): 361–367. http://dx.doi.org/10.1080/15389580902973635

Hastie, T.; Tibshirani, R.; Friedman, J. 2011. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. 745 p.

Hijar, M.; Arredondo, A.; Carrillo, C.; Solórzano, L. 2004. Road traffic injuries in an urban area in Mexico: an epidemiological and cost analysis, *Accident Analysis & Prevention* 36(1): 37–42. http://dx.doi.org/10.1016/S0001-4575(02)00112-4

Holubowycz, O. T.; Kloeden, C. N.; McLean, A. J. 1994. Age, sex, and blood alcohol concentration of killed and injured drivers, riders, and passengers, *Accident Analysis & Prevention* 26(4): 483–492. http://dx.doi.org/10.1016/0001-4575(94)90039-6

Jiang, X.; Huang, B.; Zaretzki, R. L.; Richards, S.; Yan, X.; Zhang, H. 2013. Investigating the influence of curbs on single-vehicle crash injury severity utilizing zero-inflated ordered probit models, *Accident Analysis & Prevention* 57: 55–66. http://dx.doi.org/10.1016/j.aap.2013.03.018

Kass, G. V. 1980. An exploratory technique for investigating large quantities of categorical data, *Applied Statistics* 29(2): 119–127. http://dx.doi.org/10.2307/2986296

Keall, M. D.; Frith, W. J.; Patterson, T. L. 2004. The influence of alcohol, age and number of passengers on the night-time risk of driver fatal injury in New Zealand, *Accident Analysis & Prevention* 36(1): 49–61. http://dx.doi.org/10.1016/S0001-4575(02)00114-8

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, in *IJCAI-1995: Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 20–25 August 1995, Montreal, Canada, 2: 1137–1143.

Kunt, M. M.; Aghayan, I.; Noii, N. 2011. Prediction for traffic accident severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods, *Transport* 26(4): 353–366. http://dx.doi.org/10.3846/16484142.2011.635465

Lee, J.; Mannering, F. 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis, *Accident Analysis & Prevention* 34(2): 149–161. http://dx.doi.org/10.1016/S0001-4575(01)00009-4

Malliaris, A. C.; DeBlois, J. H.; Digges, K. H. 1996. Light vehicle occupant ejections – a comprehensive investigation, *Accident Analysis & Prevention* 28(1): 1–14. http://dx.doi.org/10.1016/0001-4575(95)00023-2

Ministrstvo za notranje zadeve. 2010. *Prometna varnost – statistika 2005–2009*. Republika Slovenija, Ministrstvo za notranje zadeve, Policija. Available from Internet: http://www.policija.si/index.php/statistika/prometna-varnost (in Slovenian).

Mussone, L.; Ferrari, A.; Oneta, M. 1999. An analysis of urban collisions using an artificial intelligence model, *Accident Analysis & Prevention* 31(6): 705–718. http://dx.doi.org/10.1016/S0001-4575(99)00031-7

Öström, M.; Eriksson, A. 2001. Pedestrian fatalities and alcohol, *Accident Analysis & Prevention* 33(2): 173–180. http://dx.doi.org/10.1016/S0001-4575(00)00028-2

Pande, A.; Abdel-Aty, M. 2006. Assessment of freeway traffic parameters leading to lane-change related collisions, *Accident Analysis & Prevention* 38(5): 936–948. http://dx.doi.org/10.1016/j.aap.2006.03.004

Quddus, M. A.; Noland, R. B.; Chin, H. C. 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models, *Journal of Safety Research* 33(4): 445–462. http://dx.doi.org/10.1016/S0022-4375(02)00051-8

Quinlan, J. R. 1986. Induction of decision trees, *Machine Learning* 1(1): 81–106. http://dx.doi.org/10.1007/BF00116251

Quinlan, J. R. 1992. *C4.5: Programs for Machine Learning*. 1st edition. Morgan Kaufmann. 302 p.

Rokach, L.; Maimon, O. 2008. *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing Company. 244 p.

Singleton, M.; Qin, H.; Luan, J. 2004. Factors associated with higher levels of injury severity in occupants of motor vehicles that were severely damaged in traffic crashes in Kentucky, 2000–2001, *Traffic Injury Prevention* 5(2): 144–150. http://dx.doi.org/10.1080/15389580490435169

Sohn, S. Y.; Lee, S. H. 2003. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea, *Safety Science* 41(1): 1–14. http://dx.doi.org/10.1016/S0925-7535(01)00032-7

Sohn, S. Y.; Shin, H. 2001. Pattern recognition for road traffic accident severity in Korea, *Ergonomics* 44(1): 107–117. http://dx.doi.org/10.1080/00140130120928

Tesema, T. B.; Abraham, A.; Grosan, C. 2005. Rule mining and classification of road traffic accidents using adaptive regression trees, *International Journal of Simulation* 6(10–11): 80–94.

Timofeev, R. 2004. *Classification and Regression Trees (CART) Theory and Applications*. MSc Thesis. Humboldt University of Berlin, Germany. Available from Internet: http://edoc.hu-berlin.de/master/timofeev-roman-2004-12-20/PDF/timofeev.pdf

Ulfarsson, G. F.; Mannering, F. L. 2004. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents, *Accident Analysis & Prevention* 36(2): 135–147. http://dx.doi.org/10.1016/S0001-4575(02)00135-5

Valent, F.; Schiava, F.; Savonitto, C.; Gallo, T.; Brusaferro, S.; Barbone, F. 2002. Risk factors for fatal road traffic accidents in Udine, Italy, *Accident Analysis & Prevention* 34(1): 71–84. http://dx.doi.org/10.1016/S0001-4575(00)00104-4

Vujanić, M.; Lipovac, K.; Jovanović, D.; Pešić, D.; Antić, B. 2013. "Bottom-up" and "top-down" approach for defining road safety strategy – case study: city of Belgrade, *International Journal for Traffic and Transport Engineering* 3(2): 185–203. http://dx.doi.org/10.7708/ijtte.2013.3(2).07

Wondwossen, M. 1999. *Correlates of Car Traffic Accident: the Case of Addis Ababa in 1990*: MSc Thesis. Addis Ababa University, Addis Ababa, Ethiopia.

Yan, X.; Radwan, E. 2006. Analyses of rear-end crashes based on classification tree models, *Traffic Injury Prevention* 7(3): 276–282. http://dx.doi.org/10.1080/15389580600660062

Yau, K. K. W. 2004. Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong, *Accident Analysis & Prevention* 36(3): 333–340. http://dx.doi.org/10.1016/S0001-4575(03)00012-5

Zajac, S. S.; Ivan, J. N. 2003. Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural Connecticut, *Accident Analysis & Prevention* 35(3): 369–379. http://dx.doi.org/10.1016/S0001-4575(02)00013-1

Zantema, H.; Bodlaender, H. L. 2000. Finding small equivalent decision trees is hard, *International Journal of Foundations of Computer Science* 11(2): 343–354. http://dx.doi.org/10.1142/S0129054100000193

Zhang, J.; Lindsay, J.; Clarke, K.; Robbins, G.; Mao, Y. 2000. Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario, *Accident Analysis & Prevention* 32(1): 117–125. http://dx.doi.org/10.1016/S0001-4575(99)00039-1