Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks

By: Juan de Oña, Randa O. Mujalli and Francisco J. Calvo

Juan de Oña, Randa O. Mujalli and Francisco J. Calvo (2011) *Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks*. **Accident Analysis and Prevention, 43, 402–411.**

Direct access to the published version: http://dx.doi.org/10.1016/j.aap.2010.09.010

Analysis of Traffic Accident Injury Severity on Spanish Rural Highways Using Bayesian Networks

Juan de Oña[1,#], Randa Oqab Mujalli[1] and Francisco J. Calvo

TRYSE Research Group. Department of Civil Engineering, University of Granada

[#] Corresponding author, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada (Spain), Phone: +34 958 24 99 79, email: jdona@ugr.es

[1] These authors have contributed equally to this work.

Abstract

Several different factors contribute to injury severity in traffic accidents, such as driver characteristics, highway characteristics, vehicle characteristics, accidents characteristics, and atmospheric factors. This paper shows the possibility of using Bayesian Networks (BNs) to classify traffic accidents according to their injury severity. BNs are capable of making predictions without the need for pre assumptions and are used to make graphic representations of complex systems with interrelated components. This paper presents an analysis of 1,536 accidents on rural highways in Spain, where 18 variables representing the aforementioned contributing factors were used to build 3 different BNs that classified the severity of accidents into slightly injured and killed or severely injured. The variables that best identify the factors that are associated with a killed or seriously injured accident (accident type, driver age, lighting and number of injuries) were identified by inference.

Keywords: Bayesian networks; injury severity; traffic accidents; classification.

1. Introduction

The number of traffic accidents and their effects, mainly human fatalities and injuries, justify the importance of analyzing the factors that contribute to their occurrence. Identifying the factors that significantly influence the injury severity of traffic accidents was the main objective of many previous studies. Factors affecting injury severity of a traffic accident are usually caused by one or more of the following factors: driver characteristics, highway characteristics, vehicle characteristics, accidents characteristics and atmospheric factors (Kopelias et al., 2007; Chang and Wang, 2006).

Regression analysis has been widely used to determine the contributing factors that cause a specific injury severity. The most commonly used regression models in traffic injury analysis are the logistic regression model and the ordered Probit model (Al-Ghamdi, 2002; Milton et al. 2008; Bédard et al. 2002; Yau et al., 2006; Yamamoto and Shankar, 2004; Kockelman and Kweon, 2002). However, most of the regression models that are used to model traffic injury severity have their own model assumptions and pre-defined underlying relationships between dependent and independent variables (i.e. linear relations between the variables) (Chang and Wang, 2006). If these assumptions are violated, the model could lead to erroneous estimations of the likelihood of severe injury.

Gregoriades (2007) highlighted the interest of using Bayesian Networks (BNs) to model traffic accidents and discussed the need to not consider traffic accidents as a deterministic assessment problem. Instead, researchers should model the uncertainties involved in the factors that can lead to road accidents. He listed a number of candidate approaches for modeling uncertainty, such as, Bayesian probability.

BNs make it easy to describe accidents that involve many interdependent variables. The relationship and structure of the variables can be studied and trained from accident data. They do not need to know any pre-defined relationships between dependent and independent variables.

The three main advantages of BNs are bi-directional induction, incorporation of missing variables and probabilistic inference. By using BNs, it is relatively easy to discover the underlying patterns of data, to investigate the relationships between variables and to make predictions using these relationships. Incident data used in a study by Ozbay and Noyan (2006) were collected from incident clearance survey forms to understand incident clearance characteristics and then used to develop incident duration prediction models. The researchers modeled the incidents' clearance durations using BNs and were able to represent the stochastic nature of incidents.

Using BNs to analyze traffic accident injury severity is scarce. A two car accident injury severity model was constructed using BNs (Simoncic, 2004). A BN was built using several variables, and the Most Probable Explanation (MPE) was calculated for the most probable configuration of values for all the variables in the BN, in order to serve as an indication of the quality of the estimated BN. The results pointed out that BNs could be applied in road accident modeling, and some improvements, such as using more variables and larger datasets, were recommended. Although this study highlighted the possibility of using BNs to model traffic accidents, the results were based on building only one possible network, without measuring the performance of the Bayesian classifier.

The scope of this paper is to validate the possibility of using BNs to classify traffic accidents according to their injury severity, and to find out the best BN classification performance along with the best graphical representation, in order to be capable of identifying the relevant variables that affect the injury severity of traffic accidents.

This paper is organized as follows. Section 2 presents the data used and briefly reviews the concept of BNs and Bayesian learning. The methods used for preprocessing and evaluating the data are also presented; finally a brief description of inference is presented. In section 3, the results and their discussion are presented. In section 4, summary and conclusions are given.

2. Materials and methods

2.1. Accident Data

Accident data were obtained from the Spanish General Traffic Directorate (DGT) for rural highways for the province of Granada (South of Spain) for three years (2003-2005). The total number of accidents obtained for this period was 3,302. The data was first checked out for questionable data, and those which were found to be unrealistic were screened out. Only rural highways were considered in this study; data related to intersections were not included, since intersections have their own specific characteristics and need to be analyzed separately. Finally, the database used to conduct the study contained 1,536 records. Table 1 provides information on the data used for this study.

(insert Table 1 here)

Eighteen variables were used with the class variable of injury severity (SEV) in an attempt to identify the important patterns of an accident that usually require an explanation.

The data included variables describing the conditions that contributed to the accident and injury severity.

- Injury severity variables: number of injuries (e.g., passengers, drivers and pedestrians), severity level of injuries (e.g., fatal, severe, slight). Following previous studies (Chang and Wang, 2006; Milton et al., 2008) the injury severity of an accident is determined according to the level of injury to the worst injured occupant.
- Roadway information: characteristics of the roadway on which the accidents occurred (e.g., grade, pavement width, lane width, shoulder type, pavement markings, sight distance, if the shoulder was paved or not, etc.).
- Weather information: weather conditions when the accident occurred (e.g., good weather, rain, fog, snow and windy).
- Accident information: contributing circumstances (e.g., type of accident, time of accident (hour, day, month and year), and vehicles involved in the accident).
- Driver data: characteristics of the driver, such as age or gender.

2.2. BN Definition

Over the last decade, BNs have become a popular representation for encoding uncertain expert knowledge in expert systems. The field of BNs has grown enormously, with theoretical and computational developments in many areas (Mittal et al., 2007) such as: modeling knowledge in bioinformatics, medicine, document classification, information retrieval, image processing, data fusion, decision support systems, engineering, gaming, and law.

Let $U=\{x_1, \ldots, x_n\}$, $n \geq 1$ be a set of variables. A BN over a set of variables U is a network structure, which is a Directed Acyclic Graph (DAG) over U and a set of probability tables $B_p = \{p(x_i|pa(x_i)),\ x_i \in U)\}$ where $pa(x_i)$ is the set of parents or antecedents of x_i in BN and $i=(1,2,3,\ldots,n)$. A BN represents joint probability distributions $P(U) = \prod_{x_i \in U} p\left(x_i|pa(x_i)\right)$.

The classification task consists in classifying a variable $y = x_0$ called the class variable, given a set of variables $U = x_1 \ldots x_n$, called attribute variables. A classifier $h : U \rightarrow y$ is a function that maps an instance of U to a value of y. The classifier is learned from a dataset D consisting of samples over (U, y). The learning task consists of finding an appropriate BN given a data set D over U.

BNs are graphical models of interactions among a set of variables, where the variables are represented as nodes (also known as vertices) of a graph and the interactions (direct dependences) as directed links (also known as arcs and edges) between the nodes. Any pair of unconnected/nonadjacent nodes of such a graph indicates (conditional) independence between the variables represented by these nodes under particular circumstances that can easily be read from the graph. Each node contains the states of the random variable and it represents a conditional probability table. The conditional probability table of a node contains the probabilities of the node being in a specific state, given the states of its parents.

Figure 1 shows that the dependencies and independencies among the factors that affect the time of journey (the class variable) are represented in the form of direct edges (arrows) between

factors that are represented as nodes. For example, the variable (vehicle type) is a parent (antecedent) of the two variables (cost and velocity) called children or descendents. Any knowledge (evidence) about the parent variable affects the probabilities of occurrence of the children or descendent variables.

(insert Figure 1 here)

It should be noticed that the edges in a BN are not necessarily causal. That is, a BN can satisfy the probability distribution of the variables in the BN without the edges being causal (Neapolitan, 2009). Thus, the edges between variables in a non causal BN could imply a sort of interrelationship(s) among these variables.

2.3. BN learning and the scoring metrics used

When there are masses of data available and it is necessary to interpret them and to provide a model for predicting the behavior of unobserved cases, the learning of both structure and parameters is used (Cooper and Herskovits, 1992). There are two main approaches to structure learning in BNs:

- **Constraint based:** Perform tests of conditional independence on the data, and search for a network that is consistent with the observed dependencies and independencies.
- **Score based:** Define a score that evaluates how well the dependencies or independencies in a structure match the data and search for a structure that maximizes the score.

The advantage of score-based methods over the constraint-based methods is that they are less sensitive to errors in individual tests; compromises can be made between the extent to which variables are dependent in the data and the cost of adding the edge. Because of the aforementioned advantages, the score based method is followed in this study.

Weka software (Witten and Frank, 2005) was used in this study to build the BN. This software is freely available, it is implemented in Java language, it contains a collection of data processing and modeling techniques and it contains a graphical user interface. The BNs built here used all the nineteen variables of the 1,536 records.

In order to build BN structures; BDe Score metric, Minimum Description Length (MDL) and the Akaike Information Criterion (AIC) score functions were run, based on the hill climbing algorithm.

Let r_i $(1 \le i \le n)$ be the cardinality of x_i, q_i is used to denote the cardinality of the parent set of x_i in BN, that is, the number of different values to which the parents of x_i can be instantiated. So, q_i can be calculated as the product of cardinalities of nodes in $pa(x_i)$, $q_i = \prod_{x_j \in pa(x_i)} r_j$. Note $pa(x_i) = \emptyset$ implies $q_i = 1$.

N_{ij} $(1 \le i \le n, 1 \le j \le q_i)$ denotes the number of records in D for which $pa(x_i)$ takes its j^{th} value. N_{ijk} $(1 \le i \le n, 1 \le j \le q_i, 1 \le k \le r_i)$ denotes the number of records in D for which $pa(x_i)$ takes its j^{th} value and for which x_i takes its k^{th} value. So, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. N denotes the number of records in D.

Let the *entropy metric H (BN,D)* of a network structure and database be defined as:

$$H(BN, D) = -N \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N} \log \frac{N_{ijk}}{N_{ij}} \quad (1)$$

and the number of parameters K as:

$$K = \sum_{i=1}^{n}(r_i - 1) \cdot q_i \quad (2)$$

The AIC metric $Q_{AIC}(BN, D)$ of a Bayesian network structure for a database D is:

$$Q_{AIC}(BN, D) = H(BN, D) + K \quad (3)$$

A term *P(BN)* can be added representing prior information over network structures, but will be ignored for simplicity in the Weka implementation (Bouckaert, 1995).

The MDL metric $Q_{MDL}(BN, D)$ of a Bayesian network structure BN for a database D is defined as:

$$Q_{MDL}(BN, D) = H(BN, D) + \frac{K}{2}\log N \quad (4)$$

The BDe metric $Q_{BDe}(BN, D)$ of a BN structure for a database D is:

$$Q_{BDe}(BN, D) = P(BN)\prod_{i=0}^{n}\prod_{J=1}^{q_i}\frac{\Gamma\left(\frac{1}{q_i}\right)}{\Gamma\left(\left(\frac{1}{q_i}\right)+N_{ij}\right)}\prod_{k=1}^{ri}\frac{\Gamma\left(\left(\frac{1}{r_i}\right).q_i+N_{ijk}\right)}{\Gamma\left(\left(\frac{1}{r_i}\right).q_i\right)} \quad (5)$$

where *P(BN)* is the prior on the network structure (taken to be constant hence ignored in the Weka implementation) (Bouckaert, 1995) and $\Gamma(\cdot)$ the gamma-function.

Using hill climbing algorithm, the states of search space are possible models. Operations are the insertion, deletion and reverse of an edge in the network to modify a model. The hill climbing search algorithm was applied in this study mainly because it is fast and widely used, and also produces good results in terms of network complexity and accuracy (Madden, 2009).

2.4. BN data preprocessing

The variables obtained from the DGT were further refined and categorized into distinct values in order to be able to work with them. Other variables were merged or abstracted on the basis of procedures followed in previous studies (Simoncic, 2004; Helai et al., 2008), where the class variable was injury severity (slight injured –SI– and killed or seriously injured –KSI), and the severity was considered for the most severe case in the accident (Chang and Wang, 2006; Simoncic, 2004).

The only preprocessing filter used on this dataset was the unsupervised variable filter for replacing missing values. This filter replaces the missing values with the modes and means from the training data. The cross validation method was used to split the data into ten equal folds (or subsets), the BN was built on the fold (called training set) and the analysis was validated on the other subset (called the validation set or testing set). Multiple repetitions or trials (10 times) of cross validation are used to reduce variability, and the validation results are averaged over the trials.

2.5. BN evaluation indicators

Five indicators are used in this study to compare the BNs built (see Eqs. 6-9): accuracy, sensitivity, specificity, HMSS, and ROC area were calculated for each BN.

$$\text{Accuracy} = \frac{\text{tSI+tKSI}}{\text{tSI+tKSI+fSI+fKSI}}100\% \quad (6)$$

$$\text{Sensitivity} = \frac{\text{tSI}}{\text{tSI+fKSI}}100\% \quad (7)$$

$$\text{Specificity} = \frac{\text{tKSI}}{\text{tKSI+fSI}}100\% \quad (8)$$

$$\mathrm{HMSS} = \frac{2\times\text{sensitivity}\times\text{specificity}}{\text{sensitivity}+\text{specificity}} \quad (9)$$

Where *tSI* is true slight injured cases, *tKSI* true killed or seriously injured cases, *fSI* false slight injured cases, and *fKSI* false killed or seriously injured cases.

Accuracy (see Eq. 6) is proportion of instances that were correctly classified by the classifier. Accuracy only gives information on the classifier's general performance.

Sensitivity represents the proportion of correctly predicted slight injured among all the observed slight injured. Specificity represents the proportion of correctly predicted killed or seriously injured among all the observed killed or seriously injured (see Eqs. 7-8). Another measure used to assess the performance of the BN built was the Harmonic Mean of Sensitivity and Specificity (HMSS), which gives an equal weight of both sensitivity and specificity (see Eq. 9).

Another indicator is the Receiver Operating Characteristic Curve (ROC) Area. What ROC curves represent is the true positive rate (sensitivity) vs. the false positive rate (1-specificity). ROC curves are more useful as descriptors of overall test performance, reflected by the area under the curve, with a maximum of 1.00 describing a perfect test and an ROC area of 0.50 describing a valueless test.

Other measures used in the literature to evaluate the performance of BNs specifically include both the Most Probable Explanation (MPE) (Simoncic, 2004) and the complexity or the total number of BN arcs (Cruz-Ramírez et al., 2007). MPE is a technique that is developed for generating explanation in BNs, in which the configuration with the maximum posterior probability is calculated (Pearl, 1988).

For the analysis of traffic accident injury severity and to determine the optimal BN, the measures described above will be calculated first: accuracy, sensitivity, specificity, ROC area, the MPE and the complexity of the built BNs. Later, the best BN found in terms of these measures will be used for inference.

2.6. BN Inference

Inference in BNs consists of computing the conditional probability of some variables, given that other variables are set to evidence. Inference may be done for a specific state or value of a variable, given evidence on the state of other variable(s). Thus, using the conditional probability table for the BN built, their values can be easily inferred. Figure 1 shows an example of a conditional probability table, where it could be seen that given evidence for the distance to be "short" and the velocity to be "high", the probability that the time of journey will be less than 1 hour is 0.75. Thus, other inferences could be extracted using this figure, where the example presented here is used to explain how inference in BNs works.

In this paper, inference is used to determine the most significant variables that are associated with KSI in traffic accidents.

3. Results and discussion

Table 2 shows the results obtained from building BNs using the hill climbing search method and three different score metrics (BDe, MDL and AIC) using both the training and the test set to validate the results. From the original dataset, 2/3 of the data was held for training the BNs and the other 1/3 was used for testing them.

Ten different schemes of training/testing datasets were used to analyze the effect of swapping training and test datasets. Table 2 shows the average and the standard deviation of each one of the indicators for the score metrics used.

(insert Table 2 here)

It can be seen that both the training and the test results are very similar. The accuracies performed in this study did not vary significantly; the highest accuracy was for the BDe score (61%). Abdel Wahab and Abdel-Aty (2001) used some data mining techniques to model injury severity in traffic accidents. They obtained accuracies of 60.4% and 65.6% for training and testing sets respectively when using an MLP neural network, 56.2% when using fuzzy ARTMAP neural network and 58.1% when using O-ARTMAP. Thus, the results obtained in this paper were within the range of accuracies found by Abdel Wahab and Abdel-Aty (2001).

Also, the highest sensitivity was for BDe score; where 74% of the cases observed to be slight were also predicted to be slight. Although the BDe was capable of classifying 74% of the slight injured correctly, its specificity results indicated that its ability to classify killed or seriously injured were relatively poor. None of the score metrics achieved good results regarding the classification of killed or seriously injured (specificity); the best was for MDL and AIC scores, and test dataset with 53% of correctly classified killed or seriously injured.

The results of sensitivity for all the score metrics were relatively better than those of specificity, thus indicating that the models were able to classify slight injured rather than killed or seriously injured. This, however, was expected, since the original dataset contained more slight injuries.

HMSS could be used as a single measure of performance of the BN instead of using sensitivity and specificity separately. The results indicated that the best HMSS was achieved by using MDL and AIC scores (58%).

Figure 2 shows the ROC curves for the BNs built using the three score methods, where the X-axis represents (1-specificity) and the Y-axis represents the sensitivity.

(insert Figure 2 here)

The best ROC area obtained by BDe and MDL scores was 62%.

Table 2 suggests that the three score metrics were valid and equally effective on average.

Following Simoncic (2004), , the most convenient way to analyze the graphical performance of the three metrics is to calculate the Most Probable Explanation (MPE) for the training dataset and compare it with the results obtained from the test dataset. The training/testing dataset that showed the best results for the previous indicators was used for this purpose.

MPE is given by the most probable configuration of values for all variables in the BN. For the three estimated structures, the MPE is given by the following values for variables (see Table 1):

ACT=AS; AGE=(25-64]; ATF=GW; CAU=DC; DAY=WD; GEN=M; LAW=WID; LIG=DL; MON=SUM; NOI=1; OI=2; PAS=Y; PAW=WID; ROM=SLD; SEV=SI; SHT=THI; SID=WR; TIM=(12-18]; VI=2

Given the estimated BN structures (BDe, MDL and AIC) and the conditional probabilities for each node (see Figure 3), the probability of the MPE can be computed as shown in Table 3.

(insert Figure 3 here)

(insert Table 3 here)

For the network built by the BDe score metric, the MPE is given by the probability values shown in Table 3, column 2, row 2. Using these values, MPE for the BDe score equals 0.00088. The same calculations for the test dataset produced MPE_{test}=0.00081. This comparison of MPE and MPE_{test} can provide an indication of the quality of the estimated BN using BDe score metric; where it can be seen that there is a difference (8.2%) between the MPE produced by the training dataset and the test dataset.

The MPE for the MDL BN is given by the probability values shown in Table 3, column 2, row 3. Using these values, MPE equals 0.00076. The test dataset produced MPE_{test}=0.00073. So, the MPE as explained by the MDL is closer to the test dataset estimation (4.4% of difference), thus representing a network that is more capable of explaining different data.

The MPE for the AIC BN is given by the probability values shown in Table 3, column 2, row 4. Using these values, MPE is 0.00100. The test dataset produced MPE_{test}=0.00092. The most probable explanation has a higher probability than that produced by the test subset (8.7% of difference).

The conclusion from the above calculations of the MPE for the three score metrics as compared to the MPEs calculated for the test subset is that, in relative terms, the MDL score metric MPE gives the best explanation with regard to the MPE_{test}, whereas the difference between MPE of the built network and that computed for the test subset is the least among all the other MPEs produced by BDe, and AIC score metrics.

The last step in comparing the various score metrics and evaluating their performance was to compare the graphs' complexity, measured by the total number of arcs produced by the three score metrics studied.

Figure 3 shows the number of arcs obtained by using the three score metrics. The most complicated BN (having the highest number of arcs) is the BN built using the AIC score; this BN has 35 arcs, while the least complicated BN was the BN built by the BDe score, with 28 arcs; followed by the BN built by the MDL score, with 29 arcs.

The results of building the BNs showed that the three different score metrics did not vary significantly in terms of their accuracy, specificity, sensitivity, HMSS and ROC area. This however, indicates that BNs are valid for analyzing traffic accident injury severities and builds on the results presented by Simoncic (2004), who indicated that BNs could effectively be used to analyze this specific problem.

On the other hand, the results for the complexity of the BN graphs, the number of arcs and the MPE show some differences between the three score metrics. MDL shows the best results in terms of MPE (smaller differences between training and test sets). BDe and MDL show the best results in terms of complexity of BN graphs and number of arcs.

A closer look at the results obtained by MDL score shows that it produced a network that was relatively successful in terms of classification and prediction, where it had the second best total accuracy (59-60%). Also, HMSS showed a relatively good result for both training and testing sets respectively (56-58%,) and the ROC area results were good as well (61-62%). The BN built by the MDL score is shown in Figure 4.

(insert Figure 4 here)

Setting evidences for the variables used to build the BN using the MDL score could give indications of the values of variables that contribute to the occurrence of a killed or seriously injured (KSI) individual in a traffic accident.

Table 4 assists in the identification of the variables and values that contribute the most to the occurrence of a KSI individual in a traffic accident. For each variable, the probability of a value was set to be 1.0 (setting evidence) and the other values of the same variable were set to be 0.0. Thus, the associated probability of severity was calculated. Underlined values in Table 4 show the values of variables in which the probability of a KSI was found to be higher than that of SI.

For example, this table shows that assigning a probability of 1.0 to the value AS (angle or side impact) of the variable ACT, the probability of SI becomes 0.6219 and the probability of KSI becomes 0.3780. These probabilities are calculated from the conditional probability table of the BN built using the MDL score. Since it is intended to determine which values of variables contribute the most to the occurrence of a KSI individual in a traffic accident, Table 4 does not include the variables in which the values of probabilities of SI are always higher than those of KSI.

Setting evidences for the values of variables used to build the BN indicated that ACT, AGE, LIG and NOI were found to be significant.

(insert Table 4 here)

A detailed discussion of the most significant variables that were found to contribute to the occurrence of a killed or seriously injured (KSI) individual in a traffic accident is given below.

3.1. Accident type (ACT)

As shown in Table 4, when setting the probabilities of both HO (head on collisions) and R (rollover) values to be equal to 1.0, the probability of having KSI accidents increased, which means that these types of accidents are more significant in accidents with killed or seriously injured. Kockelman et al. (2002) found that head on crashes were more dangerous than angle crashes, left-side, and right-side crashes; they also found that they were significant in accidents that involved killed or seriously injured, but rollover crashes were more dangerous than all of the preceding crash types.

3.2. Age (AGE)

The results shown in Table 4 indicate that drivers in the age group [18-25] years were found to be more involved in accidents that resulted in KSI. Tavris et al. (2001) found that male drivers in the age group (16–24) years were much more likely to be involved in killed or seriously injured accidents than those involving older drivers.

3.3. Lighting (LIG)

Gray et al. (2008) found that among the factors that lead to a slight injury is driving in the daylight, and that more severe injuries are predicted during darkness. Helay et al. (2007) and Abdel-Aty (2003) found the same results. This coincides with the results found in this study, which indicate that roadways Without lighting (W) are associated with accidents that had KSI individuals.

3.4. Number of injuries (NOI)

The results obtained in this study indicate that when an accident results in one injury, it is more likely to be a serious injury or even fatal. Scheetz et al. (2009) used classification and regression trees to model the injury severity of traffic accidents. They also found that the number of injured occupants was a significant factor in classifying injury severity.

4. Limitations of the study

Before conclusions, some limitations should be pointed out:

- The need for large datasets when working with Bayesian networks, and the effect that imbalanced dataset (slight injured versus killed or seriously injured) has on both sensitivity and specificity.
- The data collection is based on the standard traffic police report used in Spain. So, the variable cause of the accident (CAU) was determined and judged based on the experience of the traffic police. However, a different person might have determined the same cause differently, since different time and person might lead to a different judgment.

5. Summary and conclusions

This paper uses BNs to analyze traffic accident data in order to validate the ability of this data-mining technique to classify traffic accidents according to their injury severity, and to identify the significant factors that are associated with KSI in traffic accidents.

Traffic accident data was obtained from the DGT for a period of three years (2003-2005) for Granada (Spain). Three BNs were built using three different score metrics: BDe, MDL and AIC.

Several indicators have been used in order to evaluate the performance of the built BNs: accuracy, sensitivity, specificity, HMSS, ROC Area, MPE and graph complexity (or number of arcs). The results obtained for these indicators do not vary significantly between the different score metrics used and they are within the range of previous studies (Abdel Wahab and Abdel-Aty, 2001; Simoncic, 2004). So, it could be concluded that BNs might be a useful tool for classifying traffic accidents according to their injury severity.

Inference was used to identify the values of the variables that are associated with KSI in traffic accidents on Spanish rural highways. Based on the results, it would be possible to identify the type of accident that would most probably be classified as KSI on Spanish rural highways. It would be a head-on or rollover traffic accident in a roadway without lighting with only one injury within the age of 18 and 25 years. These factors (head-on or rollover, unlit roadway, only one injury and within the age of 18 and 25 years) do not have to exist all at once in order to have a KSI accident. Any of these or a combination of them might increase the probability of a KSI accident. In general, these results are consistent with the literature (Tavris et al., 2001; Kockelman et al., 2002; Abdel-Aty, 2003; Helay et al., 2007; Gray et al., 2008; Scheetz et al., 2009). However, this finding may vary for other countries and datasets.

BNs, which have proved their effectiveness in different research areas, could be usefully applied in the domain of traffic accident modeling. Their effectiveness has been found to be similar to other data-mining techniques used to model severity in traffic accidents. Compared with other well-known statistical methods, the main advantage of the BNs seems to be their complex approach where system variables are interdependent and where no dependent and independent variables are needed (Simoncic, 2004).

Acknowledgements

The authors are grateful to the Spanish General Directorate of Traffic (DGT) for supporting this research and offering all the resources that are available to them. The authors appreciate the reviewers' comments and effort in order to improve the paper.

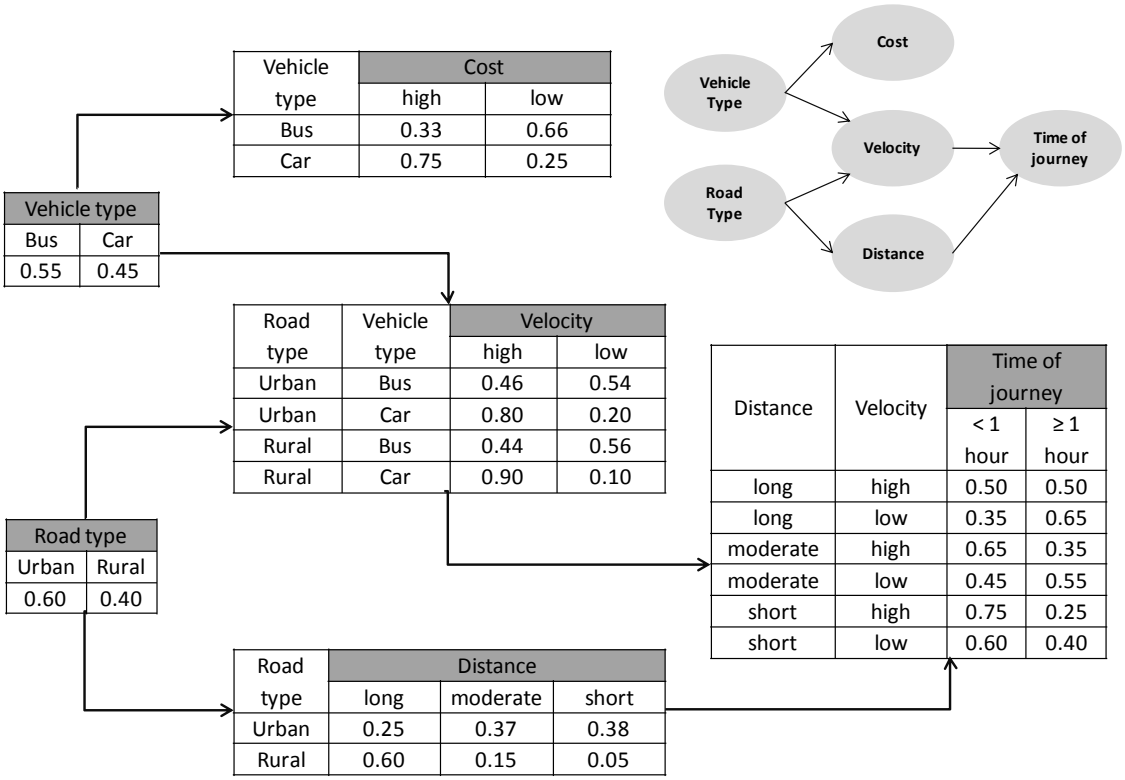
References

1. Abdel-Aty M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34, 597-603.
2. 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.
3. Al-Ghamdi A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis and Prevention 34, 729–741.
4. 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 and Prevention 34, 717–727.
5. Bouckaert. R.R., 1995. Bayesian Belief Networks: from Construction to Inference. Ph.D. thesis, University of Utrecht.
6. Chang L.Y, Wang H.W., 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019-1027.
7. Cooper, G. F. and Herskovits, E. 1992. A Bayesian method for the induction of probabilistic networks from data. Mach. Learn 9, 309-347.
8. Cruz-Ramírez N, Acosta-Mesa H.G, Carrillo-Calvet H., Nava-Fernández L.A., 2007. Barrientos-Martínez R.E. Diagnosis of breast cancer using Bayesian networks: A case study. Computers in Biology and Medicine 37, 1553 – 1564.
9. Dirección General de Tráfico – DGT [online]. Available from World Wide Web: (http://www.dgt.es/portal/es/seguridad_vial/estadistica/accidentes_30dias/anuario_estadistico/).
10. Gray R.C., Quddus M.A., Evans A., 2008. Injury severity analysis of accidents involving young male drivers in Great Britain. Journal of Safety Research 39, 483–495.
11. Gregoriades A. 2007. Towards A User-Centred Road Safety Management Method Based on Road Traffic Simulation. In: Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come, Washington D.C., pp. 1905-1914.
12. Helai H, Chor C.H., Haque M. M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: A Bayesian hierarchical analysis. Accident Analysis and Prevention 40, 45–54.
13. Kockelman K.M., Kweon Y.J., 2002. Driver injury severity: an application of ordered probit models, Accident Analysis and Prevention 34, 313–321.
14. Kopelias P., Papadimitriou F, Papandreou K, Prevedouros P., 2007. Urban Freeway Crash Analysis. Transportation Research Record 2015, 123-131.
15. Madden M. G. 2009. On the classification performance of TAN and general Bayesian networks. Journal of Knowledge-Based Systems 22, 489-495.
16. Milton J.C., Shankar V.N., Mannering F.L., 2008. Highway accident severities and the mixed logit model: An exploratory empirical analysis. Accident Analysis and Prevention 40, 260–266.
17. Mittal A., Kassim A. Tan T., 2007. Bayesian network technologies: Applications and graphical models. IGI Publishing, New York.
18. Neapolitan R.E., 2009. Probabilistic Methods for Bioinformatics. Morgan Kaufmann Publishers, San Francisco, California.
19. Ozbay K, Noyan N., 2006. Estimation of incident clearance times using Bayesian Networks approach. Accident Analysis and Prevention 38, 542–555.
20. Pearl J., 2004. Probabilistic reasoning in intelligent systems. Morgan Kaufmann, San Francisco, California.

21. Scheetz L.J., Zhang J., Kolassa J., 2009. Classification tree to identify severe and moderate injuries in young and middle aged adults. Artificial Intelligence in Medicine 45, 1-10.
22. Simoncic M., 2004. A Bayesian network model of two-car accidents. Journal of transportation and Statistics 7, No.2/3, 13-25.
23. Tavris D.R., Kuhn E.M. Layde P.M., 2001. Age and gender patterns in motor vehicle crash injuries: importance of type of crash and occupant role. Accident Analysis and Prevention 33, 167–172.
24. Witten I. H., Frank E., 2005. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann, San Francisco, California.
25. Yamamoto T., Shankar V.N., 2004. Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. Accident Analysis and Prevention 36, 869–876.
26. Yau, K.K.W., Lo H.P., Fung S.H.H., 2006. Multiple-vehicle traffic accidents in Hong Kong. Accident Analysis and Prevention 38, 1157–1161.

List of Tables:

List of Figures:



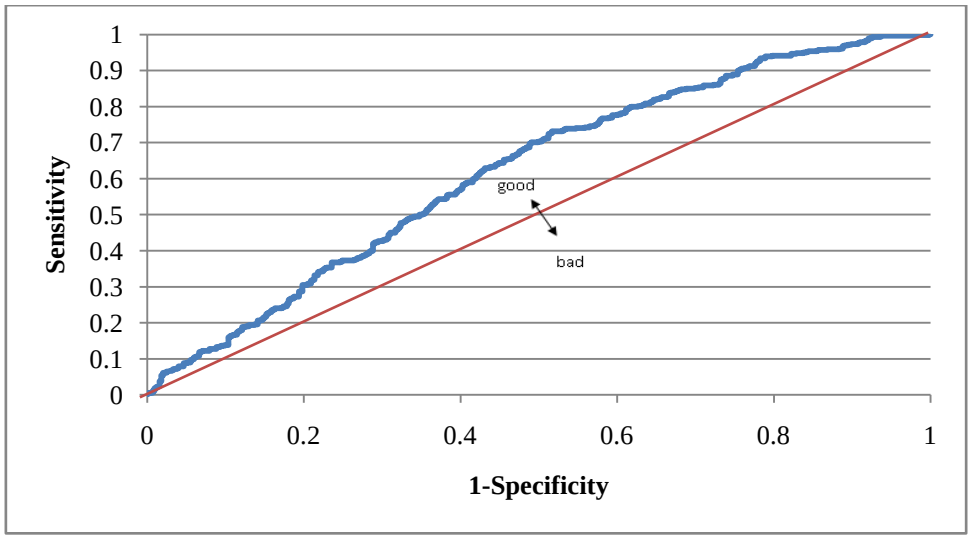
Vehicle type	Cost	
	high	low
Bus	0.33	0.66
Car	0.75	0.25

Road type	Vehicle type	Velocity	
		high	low
Urban	Bus	0.46	0.54
Urban	Car	0.80	0.20
Rural	Bus	0.44	0.56
Rural	Car	0.90	0.10

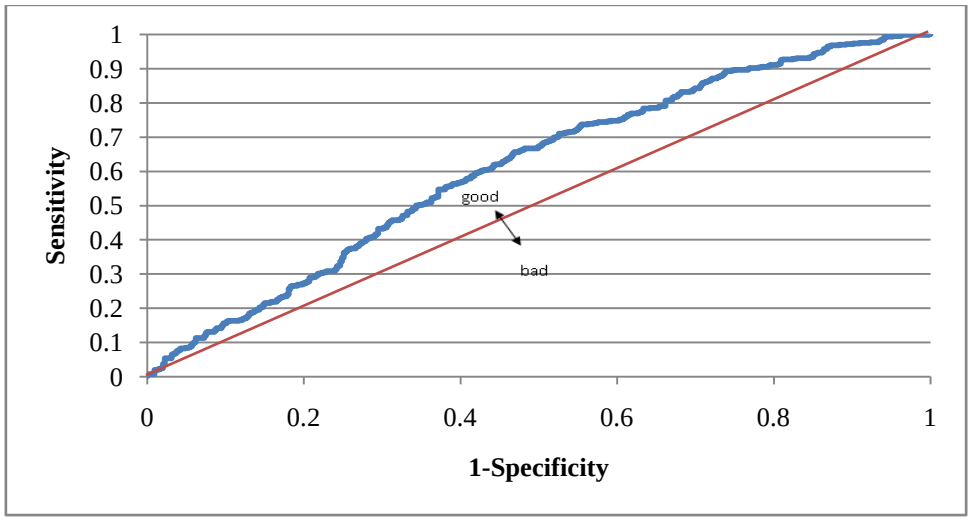
Distance	Velocity	Time of journey	
		< 1 hour	≥ 1 hour
long	high	0.50	0.50
long	low	0.35	0.65
moderate	high	0.65	0.35
moderate	low	0.45	0.55
short	high	0.75	0.25
short	low	0.60	0.40

Road type	Distance		
	long	moderate	short
Urban	0.25	0.37	0.38
Rural	0.60	0.15	0.05

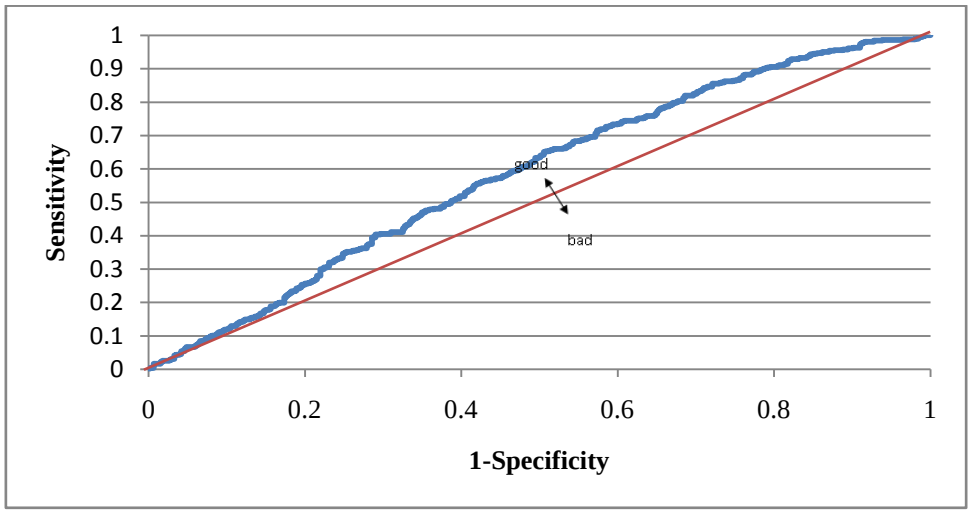
Figure 1: An example of a BN with the corresponding CPTs for each node.



(a) The ROC curve for the BDe score, ROC area is 0.62



(b) The ROC curve for the MDL score, ROC area is 0.61



(c) The ROC curve for the AIC score, the ROC area is 0.59

Figure 2. The ROC curves for the three score methods and one dataset.

	BDe			MDL			AIC		
1	SEV	→	AGE	PAS	→	ACT	VI	→	ACT
2	SEV	→	ATF	SEV	→	AGE	LIG	→	AGE
3	SID	→	ATF	SEV	→	ATF	SID	→	ATF
4	SEV	→	CAU	SID	→	ATF	SEV	→	CAU
5	SEV	→	DAY	SEV	→	CAU	GEN	→	CAU
6	SEV	→	GEN	SEV	→	DAY	SEV	→	DAY
7	SEV	→	LAW	SEV	→	GEN	VI	→	DAY
8	SEV	→	LIG	SEV	→	LAW	DAY	→	GEN
9	SEV	→	MON	PAW	→	LAW	SEV	→	LAW
10	ATF	→	MON	SEV	→	LIG	ROM	→	LAW
11	VI	→	NOI	TIM	→	LIG	PAW	→	LAW
12	SEV	→	OI	SEV	→	MON	MON	→	LIG
13	NOI	→	OI	ATF	→	MON	TIM	→	LIG
14	VI	→	OI	VI	→	NOI	PAS	→	MON
15	SEV	→	PAS	SEV	→	OI	ATF	→	MON
16	SHT	→	PAS	NOI	→	OI	AGE	→	NOI
17	SEV	→	PAW	VI	→	OI	VI	→	NOI
18	LAW	→	PAW	SHT	→	PAS	SEV	→	OI
19	SEV	→	ROM	SHT	→	PAW	NOI	→	OI
20	PAS	→	ROM	SEV	→	ROM	VI	→	OI
21	PAW	→	ROM	PAS	→	ROM	PAW	→	PAS
22	PAW	→	SHT	PAW	→	ROM	SHT	→	PAS
23	PAS	→	SID	PAS	→	SID	SHT	→	PAW
24	SEV	→	TIM	VI	→	TIM	PAS	→	ROM
25	LIG	→	TIM	ACT	→	VI	PAW	→	ROM
26	ACT	→	VI	SHT	→	SEV	ACT	→	SHT
27	ACT	→	SEV	PAS	→	SEV	PAS	→	SID
28	NOI	→	SEV	ACT	→	SEV	ROM	→	SID
29				NOI	→	SEV	TIM	→	VI
30							MON	→	SEV
31							LIG	→	SEV
32							ATF	→	SEV
33							AGE	→	SEV
34							NOI	→	SEV
35							ACT	→	SEV

Figure 3: The arcs as obtained by applying the three score metrics.

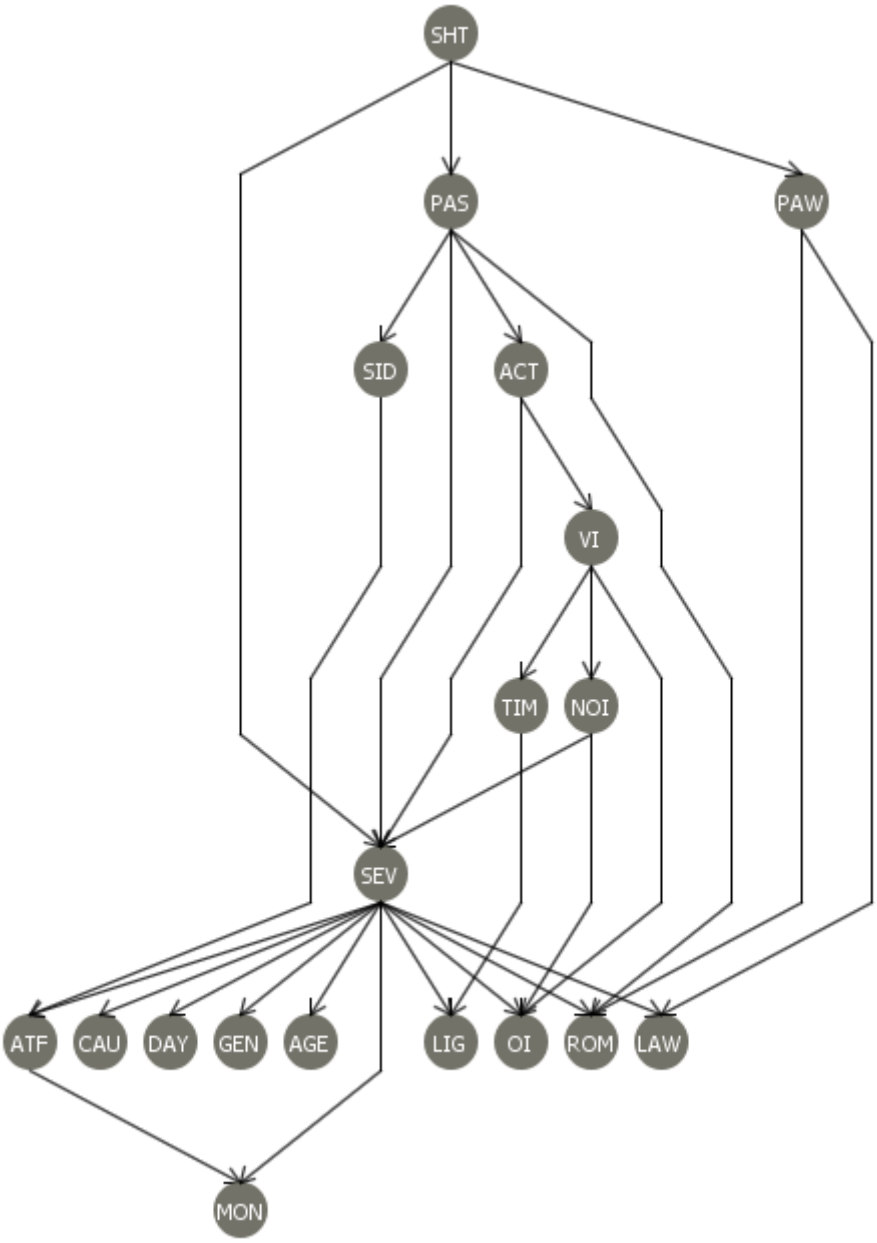


Figure 4: BN structure for the MDL score.

Table 1: Variables, values and actual classification by severity.

Variables	Values	SEV*				Total
		SI		KSI		
ACT: accident type	AS: angle or side collision	381	61.45%	239	38.55%	620
	CF: fixed objects	99	52.94%	88	47.06%	187
	HO: head on	84	40.58%	123	59.42%	207
	O: other	75	59.06%	52	40.94%	127
	PU: pile up	33	78.57%	9	21.43%	42
	R: rollover	163	49.39%	167	50.61%	330
	SP: straight path	17	73.91%	6	26.09%	23
AGE: age	[18-25]	225	50.34%	222	49.66%	447
	(25-64]	586	57.73%	429	42.27%	1015
	>64	41	55.41%	33	44.59%	74
ATF: atmospheric factors	GW: good weather	730	54.23%	616	45.77%	1346
	HR: heavy rain	23	71.88%	9	28.13%	32
	LR: light rain	84	61.76%	52	38.24%	136
	O: other	15	68.18%	7	31.81%	22
CAU: cause	DC: driver characteristics	791	54.93%	649	45.07%	1440
	OF: other factors	50	66.67%	25	33.33%	75
	RC: road characteristics	3	75.00%	1	25.00%	4
	VC: vehicle charactersitics	8	47.06%	9	52.94%	17
DAY: day	BW: beginning of week	123	60.29%	81	39.71%	204
	EW: end of week	132	57.14%	99	42.86%	231
	F: festive	29	61.70%	18	38.30%	47
	WD: week day	325	55.65%	259	44.35%	584
	WE: week end	243	51.70%	227	48.30%	470
GEN : gender	F: female	148	63.79%	84	36.21%	232
	M: male	704	53.99%	600	46.01%	1304
LAW: lane width	THI: thin: <3.25m	19	67.86%	9	32.14%	28
	MED: medium: 3.25m<=L<=3.75m	176	51.16%	168	48.84%	344
	WID: wide: >3.75m	657	56.44%	507	43.56%	1164
LIG: lighting	D: dusk	52	61.18%	33	38.82%	85
	DL: daylight	573	58.65%	404	41.35%	977
	I: insufficient	27	54.00%	23	46.00%	50
	S: sufficient	36	59.02%	25	40.98%	61
	W: without lighting	164	45.18%	199	54.82%	363
MON: month	AUT: autumn	218	54.23%	184	45.77%	402
	SPR: spring	206	59.03%	143	40.97%	349
	SUM: summer	246	56.55%	189	43.45%	435
	WIN: winter	182	52.00%	168	48.00%	350
NOI: number of injuries	1	539	49.95%	540	50.05%	1079
	>1	313	68.49%	144	31.51%	457

OI: occupants involved	1	229	51.58%	215	48.42%	444
	2	374	55.99%	294	44.01%	668
	>2	249	58.73%	175	41.27%	424
PAS: paved shoulder	missing values	66	51.56%	62	48.44%	128
	N: no	253	57.11%	190	42.89%	443
	Y: yes	533	55.23%	432	44.77%	965
PAW: pavement width	THI: thin: <6m	95	53.98%	81	46.02%	176
	MED: medium: 6 m<=law<=7m	209	54.29%	176	45.71%	385
	WID: wide: >7m	548	56.21%	427	43.79%	975
ROM: pavement markings	DME: does not exist or was deleted	60	58.25%	43	41.75%	103
	DMR: define margins of roadway	60	57.69%	44	42.31%	104
	SLD: separate lanes and defined road margins	714	55.26%	578	44.74%	1292
	SLO: separate lanes only	18	48.65%	19	51.35%	37
SHT: Shoulder type	NOS: does not exist	311	55.24%	252	44.76%	563
	THI: thin:<1.5m	402	54.47%	336	45.53%	738
	MED: medium: 1.5m<=sht<2.50m	133	58.85%	93	41.15%	226
	WID: wide >= 2.50 m	6	66.67%	3	33.33%	9
SID: sight distance	A: atmospheric	26	81.25%	6	18.75%	32
	B: building	10	55.56%	8	44.44%	18
	O: other	6	66.67%	3	33.34%	9
	T: topological	187	55.49%	150	44.51%	337
	V: vegetation	6	54.55%	5	45.45%	11
	WR: without restriction	617	54.65%	512	45.35%	1129
TIM: time	[0-6]	99	46.26%	115	53.74%	214
	(6-12]	236	57.99%	171	42.01%	407
	(12-18]	314	57.72%	230	42.28%	544
	(18-24)	203	54.72%	168	45.28%	371
VI: vehicles involved	1	316	52.06%	291	47.94%	607
	2	468	56.73%	357	43.27%	825
	>2	68	65.38%	36	34.62%	104
Total		852	55.47%	684	44.53%	1536

Table 2: Accuracy, sensitivity, specificity, HMSS and ROC Area for BDe, MDL and AIC score metrics (training and test sets).

Score Metric	BDe		MDL		AIC	
dataset	training	test	training	test	training	test
Indicator	average± s.d.*	average ± s.d.*	average ± s.d.*	average ± s.d.*	average ± s.d.*	average ± s.d.*
Accuracy	0.61±0.01	0.57±0.02	0.60±0.01	0.59±0.02	0.58±0.01	0.58±0.03
Sensitivity	0.74±0.02	0.65±0.04	0.73±0.02	0.65±0.03	0.66±0.02	0.63±0.04
Specificity	0.44±0.03	0.49±0.05	0.45±0.03	0.53±0.05	0.47±0.03	0.53±0.04
HMSS	0.55±0.02	0.56±0.03	0.56±0.02	0.58±0.03	0.55±0.02	0.58±0.02
ROC Area	0.62±0.04	0.58±0.02	0.61±0.02	0.62±0.02	0.58±0.02	0.61±0.03

*s.d.: standard deviation

Table 3: MPE for the three score metrics.

Score metric	MPE Formulas	MPE	MPE_{test}
BDe	P(ACT=AS)•P(AGE=(25-64]|SEV=SI)•P(ATF=GW|SEV=SI,SID=WR)• P(CAU=DC|SEV=SI)•P(DAY=WD|SEV=SI)•P(GEN=M|SEV=SI)• P(LAW=WID|SEV=SI)•P(LIG=DL|SEV=SI)•P(MON=SUM|SEV=SI,ATF=GW)• P(NOI=1|VI=2)•P(OI=2)|SEV=SI,NOI=1,VI=2)•P(PAS=Y|SEV=SI,SHT=THI)• P(PAW=WID|SEV=SI,LAW=WID)•P(ROM=SLD|SEV=SI,PAS=Y,PAW=WID)• P(SEV=SI|ACT=AS,NOI=1)•P(SHT=THI|PAW=WID)•P(SID=WR|PAS=Y)• P(TIM=(12-18]|SEV=SI,LIG=DL)•P(VI=2|ACT=AS)	0.00088	0.00081
MDL	P(ACT=AS|PAS=Y)•P(AGE=(25-64]|SEV=SI)•P(ATF=GW|SEV=SI,SID=WR)• P(CAU=DC|SEV=SI)•P(DAY=WD|SEV=SI)•P(GEN=M|SEV=SI)• P(LAW=WID|SEV=SI,PAW=WID)•P(LIG=DL|SEV=SI,TIM=(12-18])• P(MON=SUM|SEV=SI,ATF=GW)•P(NOI=1|VI=2)•P(OI=2|SEV=SI,NOI=1,VI=2)• P(PAS=Y|SHT=THI)•P(PAW=WID|SHT=THI)• P(ROM=SLD|SEV=SI,PAS=Y,PAW=WID)• P(SEV=SI|SHT=THI,PAS=Y,ACT=AS,NOI=1)•P(SHT=THI)•P(SID=WR|PAS=Y)• P(TIM=(12-18]|VI=2)•P(VI=2|ACT=AS)	0.00076	0.00073
AIC	P(ACT=AS|VI=2)•P(AGE=(25-64]|LIG=DL)•P(ATF=GW|SID=WR)• P(CAU=DC|SEV=SI,GEN=M)•P(DAY=WD|SEV=SI,VI=2)•P(GEN=M|DAY=WD)• P(LAW=WID|SEV=SI,ROM=SLD,PAW=WID)• P(LIG=DL|MON=SUM,TIM=(12-18])•P(MON=SUM|PAS=Y,ATF=GW)• P(NOI=1|AGE=(25-64],VI=2)•P(OI=2|SEV=SI,NOI=1,VI=2)• P(PAS=Y|PAW=WID,SHT=THI)•P(PAW=WID|SHT=THI)• P(ROM=SLD|PAS=Y,PAW=WID)• P(SEV=SI|MON=SUM,LIG=DL,ATF=GW,AGE=(25-64],NOI=1,ACT=AS)• P(SHT=THI|ACT=AS)•P(SID=WR|PAS=Y,ROM=SLD)•P(TIM=(12-18])• P(VI=2|TIM=(12-18])	0.00100	0.00092

Table 4: Inference results for variables that are associated with KSI in traffic accidents.

Variables	Values	Probabilities when setting evidences	
		SI	KSI
ACT	AS	0.6219	0.3780
	CF	0.5226	0.4773
	HO	0.3412	0.6587
	O	0.5808	0.4191
	PU	0.6683	0.3316
	R	0.4944	0.5055
	SP	0.6066	0.3933
AGE	[18-25]	0.4999	0.5000
	(25-64]	0.5567	0.4432
	≥64	0.5937	0.4062
LIG	D	0.5486	0.4513
	DL	0.5615	0.4384
	I	0.6239	0.3760
	S	0.6254	0.3745
	W	0.4527	0.5472
NOI	1	0.4957	0.5042
	>1	0.6545	0.3454

SI: slight injured; KSI: killed or seriously injured