Limitation of classification tree models in investigating road accident severity

Limitações do uso dos modelos de árvore de classificação na investigação da severidade de acidentes rodoviários

Maria Lígia Chuerubim(1); Alan Valejo(2); Barbara Stolte Bezerra(3); Irineu da Silva(4)

1 Universidade Federal de Uberlândia (UFU), Uberlândia, MG, Brasil.
  E-mail: marialigia@ufu.br | ORCID: https://orcid.org/0000-0002-2019-9198
2 Universidade de São Paulo (USP), São Paulo, SP, Brasil.
  E-mail: alanvalejo@gmail.com | ORCID: https://orcid.org/0000-0002-9046-9499
3 Universidade Estadual Paulista (Unesp), São Paulo, SP, Brasil.
  E-mail: barbara.bezerra@unesp.br | ORCID: https://orcid.org/0000-0002-8459-4664
4 Universidade de São Paulo (USP), São Paulo, SP, Brasil.
  E-mail: irineu@sc.usp.br | ORCID: https://orcid.org/0000-0001-5775-6683

Endereço correspondente / Correspondence address
Maria Lígia Chuerubim
Faculdade de Engenharia Civil – FECIV
Campus Santa Mônica - Bloco 1Y
Av. João Naves de Ávila, 2121, Santa Mônica, Uberlândia,
MG, Brasil. CEP: 38400-902

Como citar este artigo / How to cite item: clique aqui!/click here!

Abstract

The objective of this study is to discuss the main limitations identified in the classification process of traffic accident severity, as based on Classification and Regression Trees models (CART). With this purpose, CART was used in the collection of an unbalanced database of road accidents, considering injury severity, categorized as accidents without victims and with victims (fatal and non-fatal), as the dependent variable. The variables associated with accident characteristics, road infrastructure and environmental conditions were used to identify the influence of these factors on accident severity. Although the overall classification by CART resulted in a high accuracy, it also indicated a low rate of accuracy in the classification of accidents with victims, which in turn corresponds to the rarest observations in the database. In addition, it was obtained a high number of decision rules, considering the number of categories of independent variables in the prediction process of the target variable. The results indicated that CART is not efficient in the study of multiple-effect phenomena, such as road accidents, since it does not have the potential to associate a large number of parameters, which restricts the analysis and interpretation of the results to the binary structure of the tree. Thus, an exploratory analysis of the database is suggested, when the influence of a database variable was analyzed in the occurrence of traffic accidents.

Keywords: Road accidents. Severity. Data Mining. Classification. Decision Tree.

Resumo

O objetivo deste estudo foi discutir as principais limitações encontradas no processo de classificação da severidade dos acidentes de tráfego, com base em modelos de árvore de decisão (CART). Para atingir este objetivo, a CART foi utilizada na mineração de um banco de dados desbalanceado de acidentes rodoviários, considerando a variável dependente severidade da lesão, a qual foi categorizada em acidentes sem vítimas e com vítimas (fatais e não fatais). Para tanto, foram utilizadas as variáveis associadas às características dos acidentes, à infraestrutura viária e às condições ambientais, com a finalidade de se identificar a influência desses fatores na variação da severidade dos acidentes. Embora a classificação pela CART tenha resultado em uma alta acurácia, a mesma forneceu baixa taxa de acerto na classificação dos acidentes com vítimas, que correspondem às observações mais raras do banco de dados. Além disso, resultou na extração de um elevado número de regras de decisão, considerando o número de categorias das variáveis independentes no processo de predição da variável alvo. Os resultados indicaram que a CART não é eficiente no estudo de efeitos multicausais como os acidentes rodoviários, pois não tem a potencialidade de associação de um vasto número de parâmetros, o que restringe a análise e interpretação dos resultados quanto à estrutura binária da árvore. Ela é indicada, no entanto, para a análise exploratória de bancos de dados, quando se deseja analisar a influência de uma categoria específica de uma variável do banco de dados na ocorrência dos acidentes de tráfego.

Palavras-chave: Acidentes rodoviários. Severidade da lesão. Mineração de Dados. Classificação. Árvore de decisão.

1 Introduction

Road accident prevention researches aim at investigating the main factors associated with a road accident. These researches are fundamental to develop a proactive safety management in the road environment. Although the number of studies covering this theme has increased in recent years, there are still aspects to be investigated, especially in a developing country, where the majority of road accidents resulting in fatalities occur.

At present, researches on this issue have explored traditional methods that investigate the relationship between the severity of accidents and the characteristics of traffic and environment. These include traditional statistical tests (CHANG; WANG, 2006), linear regression techniques (MIAOU; LUM, 1993), binomial negative regression (ABDEL-ATY; RADWAN, 2000; HAUER, 2007), and Logistics and Probit regression (KASHANI et al., 2011; MUJALLI; OÑA, 2013; SAVOLAINEN et al., 2011).

Traditional statistical tests are criticized for their inability to analyze the variables in relation to the observed phenomenon, which are traffic accidents (HAUER, 1997). The regression models require the establishment of relations between the dependent and independent variables, which can lead to misleading estimates of probability of injury severity (ABELLÁN et al., 2013; CHANG; WANG, 2006; GRISELDA et al., 2012). Although those techniques are still in use, they are inadequate in the treatment of a large number of variables, are insensitive to the detection of outliers, noises and missing data, characteristics inherent to the accident database. (CALIENDO et al., 2007; KARLAFTIS; VLAHOGIANNI, 2011).

Due to this problem, non-parametric data mining techniques have been explored, where there is no need to settle down previous relationships between the target (dependent) variables and the predictor variables (independent) in the classification and forecasting process (ABELLÁN et al., 2013; KASHANI et al., 2011; PAKGOHAR et al., 2011). These analytical techniques allow automatic detection of the best predictor variables and their respective thresholds by extracting "if-then" decision rules that can be used to discover attitudes that occur within a specific data set (ABELLÁN et al., 2013; DE OÑA et al., 2013, 2014; KASHANI et al., 2011, LÓPEZ et al., 2013). Among the data mining techniques that have provided efficient results in relation to the classic regression models, the decision tree (CART - Classification and Regression Trees) is worth to be mentioned.

Some studies on road safety use CART for assessing the severity caused by different motor vehicles (CHANG; WANG, 2006; SAVOLAINEN et al., 2011; CHANG; CHIEN, 2013), for the identification of factors associated with traffic severity and accident patterns (ABELLÁN et al., 2013; DE OÑA et al., 2013; DE OÑA et al., 2014; GRISELDA et al., 2012; MONTELLA et al., 2011, 2012; KASHANI et al., 2011);in the

analysis of the effects of road geometry in road accidents (RUSSO et al., 2016). However, most studies detected in the literature do not discuss the challenges encountered in mining data from traffic accident databases.

In addition, road accident databases are generally not balanced, that is, they are composed of classes with just a few elements and classes with large quantity of elements, since they comprise events of a random nature whose parameters are not constant over time and space. In general, this leads to very optimistic results that do not fit the reality. In most cases, classifications with high accuracy are obtained for the majority classes and low accuracy for the minority classes. The most frequent classes in road accident databases are categorized as accidents without victims, while the rarest as with victims (fatal and non-fatal).

However, the essential variables for reducing injury severity are linked to the minority classes of the database, associated with the number of accidents with fatalities. In addition, the classifications obtained with CART are restricted to binary trees, which in most cases make it difficult to interpret and represent certain classes of the database (ABELLÁN et al., 2013).

Thus, this paper has the objective to discuss the limitations of the use of CART for road accident severity prediction.

In this perspective, this paper uses four years of unbalanced road accident database of a segment of Dom Pedro I Highway (SP-065), aiming to identify the limitations of the decision rules obtained with CART in the study of the relationship between the severity of the driver's injury and the variables associated with the road environment (weather, visibility and road surface condition) and accident characteristics (type of accident, probable cause and time of day).

2 Materials and Methods

2.1 Road accident database

The data set used in this work was provided by the "Rota das Bandeiras" concessionaire and includes the database of individual accidents recorded between km 125 and km 145 + 500 m of Dom Pedro I Highway (SP-065), in the urban area of the city of Campinas, Brazil. Four years of data were used (2009 to 2012).

The SP-065 highway has 145.5 km of extension and occupies the third position within the best highways in the national ranking. The city of Campinas occupies the eighth position in the national ranking of traffic deaths with a rate of 19.4 deaths/100 thousand inhabitants, and with a Human Development Index (HDI) in 2010 of 0.805 and a population of 1,098,630 million inhabitants (ONSV, 2014).

In order to identify the main contribution factors to road accident severity in the SP-065 highway, eight variables were used as described in Table 1. In this study, it was

built a decision tree with one main node with the target variable the accident severity, which was categorized into accidents no injury victims (NI) and with victims (WI), fatal and non-fatal.

Table 1. Description of Variables

Variables	Description	Number of Occurrences	% Total	Severity	
				NI 2,150 (76.13%)	WI 674 (24.87%)
ACT	1. Rear-end collision (REC)	910	32.22	80.11	19.89
	2. Head-on collision (HEC)	9	0.32	22.22	77.78
	3. Transverse collision (TRC)	37	1.31	75.68	24.32
	4. Lateral collision (SDC)	375	13.28	78.67	21.33
	5. Pile-up (PUP)	373	13.21	85.79	14.21
	6. Rollover (ROL)	106	3.75	53.77	46.23
	7. Run over (PEC)	63	2.23	36.51	63.49
	8. Overturning (OVE)	93	3.29	51.61	48.39
	9. Crash – fixed or mobile object (CRA)	684	24.22	82.16	17.84
	10. Fall - motorbikes and motorcycles (FAL)	174	6.16	49.43	50.57
WTC	1.Good (GO)	2358	83.50	75.06	24.94
	2. Rain (RA)	334	11.83	81.74	18.26
	3. Cloudy (CL)	71	2.51	84.51	15.49
	4. Haze (HA)	6	0.21	83.33	16.67
	5. Drizzle (DR)	55	1.95	76.36	23.64
SGC	1. Normal (NO)	1682	59.56	76.10	23.90
	2. Partial (PA)	1120	39.66	76.43	23.57
	3. Adverse (AD)	22	0.78	63.64	36.36
PFR	1. Descending (DE)	808	28.61	74.75	25.25
	2. Level (LE)	1322	46.81	77.31	22.69
	3. Ascending (AS)	694	24.58	75.50	24.50
GER	1. Straight (ST)	2514	89.02	76.49	23.51
	2. Smooth Curve (SC)	143	5.06	72.73	27.27
	3. Sharp Curve (SH)	167	5.91	73.65	26.35
PAV	1. Dry (DR)	2387	84.53	75.49	24.51
	2. Wet (WE)	428	15.16	79.67	20.33
	3. Oily (OI)	9	0.32	77.78	22.22
PER	1. Morning (MO)	1163	41.18	75.75	24.25

Variables	Description	Number of Occurrences	% Total	Severity	
				NI 2,150 (76.13%)	WI 674 (24.87%)
	2. Afternoon (AF)	1108	39.24	75.00	25.00
	3. Night (NI)	553	19.58	79.20	20.80
ACC	1. Driver (DR)	1951	69.09	76.68	23.32
	2. Vehicle (VH)	82	2.90	65.85	34.15
	3. Road and environment (RE)	122	4.32	84.43	15.57
	4. Other factors (OF)	669	23.69	74.29	25.71

Legend: Accident Type (ACT); Weather Condition (WTC); Sight Condition (SGC); Road Profile (PFR); Road Geometry (GER); Pavement Condition (PAV); Period (PER) and Accident Cause (ACC).

The variables selected for the analysis were: type of accident (rear-end collision, frontal collision, transversal collision, lateral collision, pile-up, rollover, run over, overturning, crash fixed or mobile object, and fall of motorbike or bicycle); weather conditions (good, rain, cloudy, haze and drizzle); road profile (level, ascending and descending slope); road geometry (straight, smooth curve and sharp curve); pavement condition (dry, wet and oily); visibility condition (good, partial and poor); period of day (morning, afternoon and night); and probable cause (driver, vehicle, road/environment, others). Those variables are present in several traffic accidents analysis (SAVOLAINEN et al., 2011; DE OÑA et al., 2013, 2014; GRISELDA et al., 2012; MONTELLA et al., 2011, 2012).

Initially, the quality of the data records was analyzed. The observations with inconsistent, questionable or missing information were excluded from the analysis (total of 86 observations). Thus, the most prevalent conditions of the occurrences were maintained, yielding 2,824 accidents, corresponding to 97.04% of the original data set. Subsequently, the data were divided into two subsets, which comprised the test sample (10% of the data) and training (90% of the data), so that the CART algorithm was trained and validated by the classification process of cross-correlation (KOHAVI, 1995; LÜ; ZHOU, 2010).

It can be observed (Table 1) that accidents with victims represent a small portion of the database (23.87%) in relation to accidents without victims (76.13%), meaning that this database is unbalanced.

2.2 CART construction principles

The structure of the CART tree is constructed recursively with each node representing a variable and the branches representing their respective attributes, according to a threshold or decision rule. Each terminal node or leaf specifies the expected value for each variable. To do so, metrics were used to maximize the purity

score of each node among possible input variables. In this work, the Gini metric, which seeks to maximize the homogeneity of the nodes in relation to the dependent variable, was used (PANDE; ABDEL-ATY, 2006). This metric reaches minimum values (zero) when all cases in a node fall into a single category (LÓPEZ et al., 2016; WEI et al., *2017*). The routine was implemented in SPSS software (Statistical Package for the Social Sciences), (MONTELLA et al., 2011; SINGH et al., 2016).

Thus, based on a binary structure or binary decision, the variables are grouped according to their importance to describe the target variable. In this case, a single rule is fired when an attribute is classified. In practice, a decision rule is an implication such as: if "A" then "B", where A represents a set of conditions. Each condition is defined by a relation of type attribute equal to value, attribute greater than or equal to value, and attribute less than or equal to value, where the value belongs to the domain of the attribute under analysis.

In this study, the grouped variables correlate to the dependent variable, road accident severity, classified in accidents without victims (No Injury - NI) and with victims (With Injury - WI). Through the CART structure makes it is possible to detect the nodes that contain the largest possible number of occurrences related to each category of the independent variables.

Due to the cross-validation between the test and training samples, it is possible to identify the structure that best fits the data set, cutting the tree when necessary and excluding the nodes or branches that add little to the classification process of the dependent variable, that is, have little predictive value.

3 Results and discussion

Table 2 shows that the accuracy obtained in the classification process using CART for the database used in this study was 78.6%, which is in agreement with the literature (KASHANI et al., 2011; DE OÑA et al., 2013; PAKGOHAR et al., 2011; CHANG; CHIEN, 2013).

Table 2. Accuracy rate classification using CART

Severity level	Prediction		
	Accidents without victims	Accidents with victims	Accuracy rate (%)
Accidents without victims	2,039	111	94.8%
Accidents with victims	492	182	27.0%
Percent of correct	89.6%	10.4%	78.6%

However, CART presented a low accuracy rate in the classification of accidents involving fatal and non-fatal victims (27%) and high accuracy rate for accidents without victims (94.8%), as shown in Table 2.

It can be seen from the results presented in Table 2 that of the total of 2,150 NI accidents, 2,039 were classified correctly (94.80%) and 111 were erroneously classified as WI accidents (5.20%). While of the total of 674 WI accidents, 73% were classified as NI accidents and only 27% were properly classified as accidents involving fatal and non-fatal victims.

The decision tree resulted in 40 nodes, from which 21 are terminal nodes. The importance of each predictive or independent variable for the classification of traffic accident severity: ACT (100.0%), PER (70.4%), ACC (58.6%), GER (40.4%), SGC (30.8%), WTC (10.0%), PFR (3.8%) and PAV (2.9%).

In this study, the most important independent variable for characterization of road accident severity is associated with ACT, with 100% importance (Table 3). Subsequently, were the variables PER (70.4%), ACC (58.6%) and GER (40.4%). The other variables had less importance, and the variable PAV (2.9%) was the one that had least influence in the prediction process and, therefore, was eliminated in the CART adjustment process.

The results obtained with CART are shown in Figure 1 and the decision rules in Table 3. The binary tree structure represents the number of induced decision rules, which correspond to the number of terminal node. The larger the CART structure, the greater the number of both terminal nodes and decision rules. This characteristic of CART limits and restricts its application in big data searches, which have a large number of variables and multiple possible combinations.

In this experiment, 21 decision rules were extracted. Even though 10 nodes could be identified as associate to classify the WI (fatal and non-fatal victims) accidents (terminal nodes 7, 14, 16, 26, 33, 34, 36, 38, 39, and 40), only six of them (nodes 14, 33, 34, 36, 38, and 39) showed the highest probability to be linked with WI accidents.

Node 33 indicates that collision type accident (frontal, transversal, lateral), pill-up, rollover, overturning, run over, crash with fixed object and fall were those that culminated in WI accidents with a probability of 61.10%, in straight or with smooth curve road segments, under good visibility. The probable cause of the occurrence of these accidents was associated with driver and vehicle factors.

Node 34 considers the same types of WI accidents of node 33, but only those that occurred exclusively in the morning, in straight or with smooth curve road segments, partial and poor visibility. These accidents had a probability of 44.70% and the probable cause was, again, associated with the driver and the vehicle.

In node 36 the same types of WI accidents of nodes 33 and 34 are verified, but they occurred exclusively due to probable causes road/environment and others

(presence of cyclist, pedestrian and animal on the road, congestion, previous accident and suicide), in the morning and in straight or with smooth curve road segment, with probability of 51.60%.

Node 38 shows the WI accidents of the rollover type, run over, pile-up, crash with fixed object and fall, which occurred in the afternoon, with good, rain, or drizzle weather conditions, with probability of 59.70%.

Node 39 presents the same types of WI accidents as node 38, but those occurred at night and with a straight segment, with a probability of 30.60%. Node 14 concentrates on the WI accidents caused by frontal collision and run over accidents in the afternoon and night, with probability of 86.80%.

Figure 1. Decision Tree results

Legend: No Injury (NI); With Injury (WI); Accident Type (ACT); Weather Condition (WTC); Sight Condition (SGC); Road Profile (PFR); Road Geometry (GER); Pavement Condition (PAV); Period (PER) and Accident Cause (ACC).

Table 3. Decision Rules

Node	Classification Rules	NI		WI		
	IF	N	Total (%)	N	Total (%)	Then
7	If [(ACT = 5) and (ACC = 2)]	2	50.00	2	50.00	NI and WI
25	If [(ACT = 5) and (ACC = 4 or 1 or 3) and (WTC = 1 or 2 or 3 or 4) and (SGC = 1 or 2)]	312	86.90	47	13.10	NI
26	If [(ACT = 5) and (ACC = 4 or 1 or 3) and (WTC = 1 or 2 or 3 or4) and (SGC = 3)]	3	60.00	2	40.00	NI and WI
16	If [(ACT = 5) and (ACC = 4 or 1 or 3) and (WTC = 5)]	3	60.00	2	40.00	NI and WI
27	If [(ACT = 1 or 4 or 9 or 3) and (ACC = 4 or 1) and (PER = 1) and (SGC = 2)]	167	73.90	59	26.10	NI
28	If [(ACT = 1 or 4 or 9 or 3) and (ACC = 4 or 1) and (PER = 1) and (SGC = 1 or 3)]	437	79.30	114	20.70	NI
29	If [(ACT = 1 or 9) and (ACC = 4 or 1) and (PER = 2 or 3)	689	82.80	143	17.20	NI
30	If [(ACT = 4 or 3) and (ACC = 4 or 1) and (PER = 2 or 3)	201	77.30	59	22.70	NI
31	If [(ACT = 1 or 4 or 9 or 3) and (ACC = 2 or 3) and (GER = 1 or 2) and (PER = 2)]	58	90.60	6	9.40	NI
32	If [(ACT = 1 or 4 or 9 or 3) and (ACC = 2 or 3) and (GER = 1 or 2) and (PFR = 1 or 3)]	53	82.80	11	17.20	NI
20	If [(ACT = 1 or 4 or 9 or 3) and (ACC = 2 or 3) and (GER = 3)]	9	100.00	0	0.00	NI
33	If [(ACT = 10 or 6 or 8 7 or 2) and (PER = 1) and (GANDR = 1 or 2) and (ACC = 1 or 2) and (SGC = 1)]	21	38.90	33	61.10	NI and WI
35	If [(ACT = 10 or 6 or 8 or 7 or 2) and (PER = 1) and (GER = 1 or 2) and (ACC = 4 or 3) and (PFR = 2)]	23	82.1	5	17.90	NI
36	If [(ACT = 10 or 6 or 8 or 7 or 2) and (PER = 1) and (GER = 1 or 2) and (ACC = 4 or 3) and (PFR = 1 or 3)]	15	48.40	16	51.60	NI and WI
12	If [(ACT = 10 or 6 8 7 or 2) and (PER= 1) and (GER= 3)]	22	78.60	6	21.40	NI
37	If [(ACT = 10 or 6 or 8) and (PER = 2) and (WTC = 3)]	4	100.00	0	0.00	NI
38	If [(ACT = 10 or 6 or 8) and (PER = 2) and (WTC = 1 or 2 or 5)]	62	40.30	92	59.70	NI and WI
39	If [(ACT = 10 or 6 or 8) and (PER = 3) and (GER = 1)]	34	69.40	15	30.60	NI and WI

Node	Classification Rules	NI		WI		
	IF	N	Total (%)	N	Total (%)	Then
40	If [(ACT = 10 or 6 or 8) and (PER = 3) and (GER= 2 or 3)]	4	33.30	8	66.70	NI e WI
14	If [(ACT = 7 or 2) and (PER = 2 or 3)	5	13.20	33	86.80	WI
	Total 2,824 accidents	2,150		674		

*N = sample size; NI = without victims; WI= with victims.

The main factors contributing to the occurrence of WI accidents (nodes 14, 33, 34, 36, 38, and 39) are associated with road characteristics such as road profile, environmental conditions (weather and visibility condition), time of day and the characteristics of accidents (type of accident and probable cause). Note that the accidents that are most likely to generate victims were accidents due to frontal collision and run over (node 14). WI accidents in road segments with sharp curve are less recurrent, with 21.40% probability (node 12). These accidents were of the type collision (frontal, transversal, lateral), overturning, rollover, run over, crash with fixed object and fall (bicycle and motorcycle), being more frequent in the morning. These results are in accordance with previous studies (GRISELDA et al., 2012).

The pavement condition variable was pruned from the final tree structure, as it presented little relevance to the model (Table 3). However, it is directly associated with environmental conditions in the segment, with the exception of accidents that occurred in an oily pavement condition.

NI accidents were identified with a higher likelihood of occurrence and frequency of data at nodes 12, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 38 and 39. This indicates that NI accidents are recurring in all segments of the highway and can happen due to all types of accidents. They are also associated with environmental variables (weather and visibility condition), road characteristics (road geometry and road profile) and probable cause. The most recurrent NI accidents are observed at nodes 28 and 29, with probability of occurrence of 79.30% and 82.80%, respectively. These accidents were collision type (rear, transversal and lateral) and crash with fixed object, with probable cause associated with the driver and other ones (congestion and previous accident), at different periods of the day (morning, afternoon, and night) and with good or bad visibility condition. The driver variable is in accordance to the previous studies of Abdel-Aty e Radwan (2000).

4 Conclusions

The results obtained with the CART algorithm express the influence on road accident severity classification of the variables related to road infrastructure and environmental conditions. In the classification process, although an accuracy

compatible with the values found in the literature (78.6%) was obtained, it was verified that the accuracy rate for WI accidents (27%) was much lower than the NI accidents accuracy rate (94.8%). This is due to fewer observations related to WI in the road accident database, which means that the database is unbalanced. Due to this characteristic, it is recommended in road accident severity classification studies that databases be balanced in order to properly train the classifier employed by different class classification approaches.

Since the classification process with CART is based on set of attributes where each internal node corresponds to a test on the values of the attributes of a given variable, it is expected that a balanced database will produce more consistent results with good accuracy.

In the context of Road Safety, CART can be efficient when analyzing the impact of a specific category of a particular dependent variable on road accident severity, such as the driver's profile (age, gender), type of vehicle (motorcycles, trucks, vehicles), level of drunkenness (high, low, medium), among others. However, CART is not efficient for the analysis of multi-causal effects associated with the study of road accident. For these cases, it is indicated the elaboration of a CART for each dependent variable, with a smaller and significant number of decision rules.

In the spatial scope of the problem, it would also be interesting to add neighborhood relations based on measures of similarity, since the location of accidents can affect the probability of road accident severity.

For future work, more refined data mining techniques are recommended, based on network structures, in order to extract not only exploratory and visual information from the database, but also to identify, in a more efficient way, the main factors, which when combined, contribute to the occurrence of road accidents.

References

ABDEL-ATY, M. A.; RADWAN, A.E. Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, v. 32, n. 5, p. 633-642, 2000.

ABELLÁN, J et al. Analysis of traffic accident severity using Decision Rules via Decision Trees. *Expert Systems with Applications*, v. 40, n. 15, p. 6047-6054, 2013.

CHANG, L.; CHIEN, J. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. *Safety Science*, v. 51, n. 1, p. 17-22, 2013.

CHANG, L.; WANG, H. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention*, v. 38, p. 1019–1027, 2006.

DE OÑA et al. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis and Prevention*, v. 51, p. 1–10, 2013.

DE OÑA et al. Analysing the relationship among accident severity, drivers' behaviour and their socio-economic characteristics in different territorial contexts. *Procedia - Social and Behavioral Sciences*, v. 160, n. Cit, p. 74–83, 2014.

GRISELDA, L. et al. Using Decision Trees to extract Decision Rules from Police Reports on Road Accidents. *Procedia - Social and Behavioral Sciences*, v. 53, n. SIIV-5th International Congress-Sustainability of Road Infrastructures, p. 106–114, 2012.

HAUER, E. Safety Models for Urban Four-lane Undivided Road Segments. *Transportation Research Record: Journal of the Transportation Research Board*, n. 96–105, p. 1–22, 2007.

HAUER, E. *Observational Before-After Studies in Road Safety*. Pergamon Press, Elsevier Science Ltd., Oxford, England. 1997.

KASHANI, A. T. et al. A data mining approach to identify key factors of traffic injury severity. *Traffic&Transportation*, v. 23, p. 11–17, 2011.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Appears in the International Joint Conference on Articial Intelligence, v. 5, p. 1–7, 1995.

LÓPEZ, G. et al. Analysis of traffic accidents on rural highways using Latent Class clustering and Bayesian Networks. *Accident Analysis and Prevention*, v. 51, p. 1–10, 2013.

LÓPEZ, G. et al. Influence of deficiences in traficc control devices in crashes on two-lane rural roads. *Accident Analysis and Prevention*, v. 96, p. 130-139, 2016.

LÜ, L.; ZHOU, T. Link prediction in weighted networks: The role of weak ties. *Europhysics Letters*, v. 89, n. 1, p. 18001, 2010.

MIAOU, S. P. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis and Prevention*, v. 26, n. 4, p. 471–482, 1994.

MIAOU, S. P.; LUM, H. Modeling vehicle accidents and highway geometric design relationships. *Accident Analysis and Prevention*, v. 25, n. 6, p. 689–709, 1993.

MONTELLA, A. et al. Data-Mining Techniques for Exploratory Analysis of Pedestrian Crashes Data-Mining Techniques for Exploratory Analysis of Pedestrian Crashes. *Transportation Research Record*, n. 2237, p. 107–116, 2011.

MONTELLA, A. et al. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident Analysis and Prevention*, v. 49, p. 58–72, 2012.

MUJALLI, R. O.; DE OŇA, J. Injury severity models for motor vehicle accidents : a review. *Proceedings of the Institution of Civil Engineering – Transport.*, v. 166, p. 255–270, 2013. DOI: http://dx.doi.org/10.1680/tran.11.00026

ONSV. *National Road Safety Observatory*. Statistics. 2014. Available at: http://iris.onsv.org.br/iris-beta/#/stats/maps. Access: 22 fev. 2018. In Portuguese.

PAKGOHAR, A. et al. The role of human factor in incidence and severity of road crashes based on the CART and LR regression : a data mining approach. *Procedia Computer Science*, v. 3, p. 764–769, 2011.

PANDE, A.; ABDEL-ATY, M. Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention*, v. 38, p. 936–948, 2006.

RUSSO, F. et al. Safety performance functions for crash severity on undivided rural roads. *Accident Analysis and Prevention*, v. 93, p. 75–91, 2016.

SAVOLAINEN, P. T. et al. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis and Prevention*, v. 43, n. 5, p. 1666–1676, 2011.

SINGH, G. et al. M5 model tree based predictive modeling of road accidents on non-urban sections of highways in India. *Accident Analysis and Prevention*, v. 96, p. 108–117, 2016.

WEI, X. et al. Analyzing traffic crash severity in work zones under different light conditions. *Journal of Advanced Transportation*, v. 2017.