Using Classification and Regression Trees (CART) to Identify Factors Contributing to Vehicle Crash Severity in a Port City

ELAHE ABBASI[1], YUEQING LI[1]*, XING WU[2], BRIAN CRAIG[1]
[1]Department of Industrial and Systems Engineering
[2]Department of Civil and Environmental Engineering
Lamar University
4400 MLK Blvd., PO Box 10009, Beaumont, Texas 77710
USA
yueqing.li@lamar.edu

Abstract: Vehicle crash is one of the leading causes of deaths and results in around 1.35 million fatalities in the world each year. As one of the busiest ports in the United States, Beaumont (TX) has a lot of heavy vehicles in its traffic flow. Between 2010 and March 2019, there were a total of 37,984 crashes involving 103,407 persons in Beaumont. This study identifies the factors influencing crash injury severity on Beaumont. Yet, not only the conventional roadway and temporal factors but also environmental characteristics which need particular attention were considered in the this research. To identify the critical factors influencing crash severity, Classification and Regression Trees (CART) method was used. The CART model had an accuracy of 62% in prediction. The results indicated that "light condition", "crash time" and "weather condition" are three factors that affect the crash severity. The current investigation can help researchers and policymakers to achieve a better understanding of traffic crashes in humid subtropical climate port cities and assist decision-makers to make more efficient decisions.

Keywords: Crash Severity, Classification, Regression, Classification and Regression Trees (CART), Data Mining

1 Introduction

Traffic crashes have been considered as one of the main causes of death. Based on the World Health Organization report, over one million people were killed in traffic crashes each year [1]. In Texas, United States, 648,370 crashes occurred in 2019 [2]. These crashes not only have detrimental impacts on individuals but also have adverse effects on the societal level and bring lots of physical, psychological, financial and environmental losses [3]. Hence, the need for diminishing the severity and frequency of the crashes is considered more than any other day [4].

A lot of noticeable studies with different models have been implemented in traffic crashes to lessen the numbers and damages of crashes, but most investigations in this area are regression based [4–6]. In fact, regression based models need some basic assumptions and the relationship between the dependent (target) and independent variables in these models should be investigated too [4].

Lately, data mining methods and non-parametric approaches have been used in traffic crash analysis fields [7–9]. As Breiman et al. [10] introduced, Classification and Regression Tree (CART) is a non-parametric method without any presumed relationship between dependent and independent variables. It is a powerful method that can easily search and identify the best predictor to classify the dependent (target) variable. Its tree-based structure helped complex relations between the variables be easily shown. As a matter of fact, CART has a recursive partitioning structure and can be used for classification and regression trees to predict both categorical and continuous target variables [7].

The purpose of this research is to distinguish the crucial factors influencing crash severity in Beaumont, Texas. Beaumont has one of the busiest ports in the United States [11] and is one of the 10 dominant seaports along Texas' 367-mile coastline of Gulf of Mexico [12]. As the Port of Beaumont reported, cargo volume has increased 87% from 2017 to 2019. As a result, vehicles moving to transport these shipments have grown significantly in this area [13]. This causes a lot of heavy vehicles marching on the roads daily and can increase the traffic and crashes constantly. Therefore, as Brooks [14] stated, the Texas Department of Transportation (TxDOT), has different plans to mitigate the congestion of heavy vehicle from streets and reduce traffic and crashes. For example, $1.57 million grant is assigned to the port of Beaumont to enhance their port access program in 2019-2020 and build a new truck queuing area to accommodate the trucks and

removing them from city streets [14]. Also, regarding climate changes, southeast Texas area including Beaumont experience more severe weather condition recently which can cause more extreme crashes [15]. Thus, this study focused on employing the CART approach to identify the crash contributing factors in Beaumont in order to eliminate them and reduce severe injuries and fatalities.

2 Literature review

Transportation issue brings about challenges to professionals in various fields such as safety, urban and regional planning, logistics, social science and economy [16]. Hence, many investigations on transportation, especially on the severity of traffic crashes and their consequences have been conducted [5,8,17,18]. These studies vary both empirically and methodologically.

2.1 Factors contributing to crash severity

From empirical point of view, most researches tried to identify the critical factors which affect the crash severity with the purpose to control or eliminate them. These factors may include age [4,5,19], gender [5,20,21], number of lanes [8,22], alcohol involvement [20,23], light condition [4,8, 20,24], speeding [6,19,20], road type [21,24], crash type [6], time of day [4,7,18] and average traffic volume [20,25]. Among them, some investigations focused on the impact of these factors on crash severity in specific roads. For instance, crashes which happened in rural roads [6,8], freeway tunnel [18,26] and freeway [20,25]. Also, some other studies concentrated on specific target group such as pedestrian [23,27] bicycles [28,29] or young drivers [30,31]. But not just roadway characteristics, human and temporal factors are contributed to the severity of crashes. Regarding the changes in the climate patterns, environmental factors like weather conditions [4,7,21,24] play more vital roles in comparison to the past [32].

2.2 Logit Models and Classification and Regression Tree

From the methodological perspective, many different methods have been carried out to distinguish key factors influencing the severity of the crashes. While some performed classic statistical methods like χ^2-test [19,23], other utilized more complex methods such as multivariate regression analysis and data mining approaches [4,5,7,8].

One of the most favorable, useful and practical regression-based methods is logit model [33–36]. For example, multinomial logit (MNL) was used to investigate and recognize important factors of pedestrian-vehicle crash severity in North California [27]. This study showed that the crash fatality can be increased by factors such as vehicle type, driver's physical condition, pedestrian age, weekend, light condition, roadway characteristic, roadway surface and speed limit. In another research, mixed logit and latent class models were implemented and the severity of driver's injury in rural area crashes, which happened in rainy weather was studied [37]. This investigation indicated that variables like curve path, on grade road, signal control, multiple lanes, pickup cars, drug/alcohol impaired, and not using seat belt were the crucial factors which could raise the injuries and fatalities. Furthermore, the multinomial logit model was utilized to explore the important variables which caused different levels of injury severity for teenage and adult drivers in intersection-related crashes too [38].

Recently, nonparametric approaches and data mining methods are used more by researchers in crash and traffic areas to classify the crash severity. Random Forests (RF), Support Vector Machines (SVM) and Nearest Neighbor Classification (NNC) were methods applied to predict crash severity in Nebraska [39]. Some researchers applied Artificial Neural Networks (ANN) [39,40], others utilized Decision Trees in their traffic safety related studies [7,17,41,42].

Among different methods of tree-based data mining methods, the Classification and Regression Tree (CART) can be employed to assess the relationship between risk factors include driver/vehicle characteristics, highway/environmental variables and accident variables and target variable like crash severity [7]. The CART was used by Change and Wang [7] in their inquiry, because it is a useful prediction method which does not need any pre-defined underlying correlation between dependent and independent variables. They analyzed 12,604 crashes in Taiwan and found that pedestrians, bicyclists, motorcyclists and passengers were vulnerable users with more severe injuries in crashes. Furthermore, the CART was applied on the two-lane, two-way rural area to evaluate the crash injury severity [8]. This study's result illustrated that not using a seat belt and improper overtaking were

two main reasons that increased the crash severity in Iran.

2.3 Study objectives

There are claims that non-parametric methods like applied hierarchical tree-based regression are more efficient than parametric models such as multiple linear and negative binomial regression models [22]. In this regard, Kuhnert et al. [43] suggested that CART and multivariate adaptive regression splines (MARS) can be utilized to gain more detailed results. Since the relationship between variables are not an important concern when using CART, this method can be deemed as a more practical, easier and powerful technique to find the important risk variables [43]. Yet, some studies found no significant difference in the classification accuracy with different techniques, such as logistic regression, neural network and applied classification tree [44].

Therefore, regarding its simplicity and power, this study carried out with the CART to extract the main factors which causes crashes in Beaumont. The city has a high volume of cargo transportation [13] and humid subtropical climate which causes the region receives the highest amount of rain in the state and experiences short cold winters [45]. Different from most of the previous studies that concentrated on some limited risk factors, such as some particular kinds of roads, users or crashes, this study investigated all types of crashes happened in different subtropical weather conditions. Additionally, while most of the previous researches used just categorical risk factors as independent variables, this investigation also considered the numerical variables.

3 Methodology

3.1 Classification and Regression Tree (CART)

The CART is an approach used to predict either continuous target variables by regression or categorical ones by classification [10]. It is a binary recursive partitioning approach in which the root node is divided into two internal child nodes and the process will continue by considering each internal node as a root node until it cannot find any other useful splits [7,46]. In addition, the dataset should be randomly divided into two subsets of training and testing to achieve more accurate results [47]. While the training sample is used to split nodes, the testing sample is utilized to check misclassification and predict the target variable [8].

As noticed, splitting is a fundamental step in CART. There are various available criteria which are used for splitting. Gini index is one of the most famous of them and can be illustrated as below:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \qquad (1)$$

where i is the number of classes and p_i is the probability of an object classified to a particular class. In fact, the Gini index or Gini impurity, measures the probability of a particular variable being wrongly classified when it is randomly chosen. In other words, choosing the variable with the least Gini index as the root node is preferable [10]. The aim of decision tree is to create homogenous groups as much as possible. Similarly, the CART is designed to make terminal nodes which consist of subjects belonging to only one of the n categories of the target variable [7].

3.2 Overfitting in Decision Tree

Like other decision tree-based methods, the decision tree of the CART can grow bigger and bigger until it reached the same observation in each terminal node. Following this way, a tree will be generated which has the maximum size and overfit the training data. To overcome the overfitting problem and reduce the complexity of decision trees which will cause misclassification in training set, pruning [10]. can be used. Pruning is a cost-complexity based algorithm to remove branches that add less predictive value to the tree and create simpler subtrees. In the first step, to find the optimal tree, a sequence of various subtrees is built on the training data. In the next step, one of these trees is going to be chosen as the pruned tree based on its accuracy on a pruning set. Pruning set is a portion of the training data that is set aside exclusively for pruning alone. Breiman et al. [10] have presented more detailed explanations about the mathematical aspect of CART analysis and its different applications.

4 Data description

The crash data for this research has derived from the monthly reports issued by Texas Department of Transportation (TxDOT). TxDOT contains large,

general and important data concerning each crash. This study used the published data from January 2010 to March 2019 in Beaumont, Texas. A total of 37,984 crash records were collected which were grouped regarding the level of the severity of each crash. In this investigation, according to considerable previous observations and studies, some irrelevance variables namely county, street name or number were disregarded. Moreover, potential curtail factors such as traffic flow features (e.g., adjusted average daily traffic amount), temporal features (e.g., crash time), environmental features (e.g., light condition, weather condition), and roadway features (e.g., road class, surface width) were used in attempt to identify the factors affecting injury severity.

Hence, crash severity was considered as the target variable (y) and adjusted average daily traffic amount, crash time, road light condition, road class, surface width and weather condition were deemed as independent variables (x).

4.1 Data preparation

Crash severity in this dataset consisted of six categories (levels) including suspected serious injury, non-incapacitating injury, possible injury, killed, not injured and unknown. Yet, as this investigation tried to concentrate on the severity of crashes, killed and suspect serious injury were merged and considered as serious injuries or killed category. Also, possible injury and non-incapacitating injury categories were recognized as light injuries and not injuries category was deemed as no injuries. By the same token, crashes with no injuries were eliminated from the analyses. In addition, as the percentage of crashes which labeled as unknown was negligible (almost 2%), they were deleted from the dataset. Hence, the target variable was classified into two major levels: serious injuries or killed, and light injurie.

Furthermore, some crashes in the dataset were repeated because there was more than one entry for the same crash by different persons, so duplicated data were omitted. Additionally, some crashes missed the light condition and weather condition variables. All crashes with missing variables were also omitted. Finally, the data set reduced to 5,557 crashes. Besides, 4780 crashes (80% of the data) were designated to train and the rest was selected to test the model.

4.2 Data processing

In the data processing step, the database was explored more in detail. In crash time, peak time and off-peak time are more essential to be investigated. Thus, as Xu et al. [25] suggested, the time from 6 AM to 9 AM and 5 PM to 8 PM were considered as peak times.

As Iacobucci et al. [48] presented, the median split can be considered as robust, refined and revived method in data mining approach to transform a continuous variable into a categorical one. Hence, for the two continuous variables (surface width and adjusted average daily traffic amount), median split was used. As a result, for adjusted average daily traffic amount variable, 43,459 daily traffic was presumed as the median (midpoint) and the amount more than that was labeled high and less than that considered as low volume traffic. The same happened about the surface width and 68 (ft) was considered as the median of this variable. Thus, areas with surface width more than 68 (ft) were recognized as high width and less than that deemed as low width. The summary of variable levels and the percentage of distribution of injury severity by key variables are presented in table 1 and table 2, respectively.

Table 1. variables and their new levels

Variable	Type	Variable Range	Detailed level
Crash Severity (Target Variable)	Nominal	2 levels	Serious Injuries or Killed Light Injuries
Adjusted Average Daily Traffic Amount	Numeric	2180-110962	Low Volume Traffic High Volume Traffic
Crash Time	Interval	0-23:59	Off Peak Hour Peak Hour

Light Condition	Nominal	6 levels	Daylight Dark, Lighted Dusk Dark, Not Lighted Dawn Dark, Unknown Lighting
Road Class	Nominal	4 levels	Interstate US & State Highways Farm to Market City Street
Surface Width	Numeric	24-168	Low Width, High Width
Weather Condition	Nominal	7 levels	Clear Cloudy Rain Fog Sleet/Hail Snow Blowing Sand/Snow

Table 2. Distribution of injury severity by key variables

Crash conditioning variables	Severity frequency	
	Serious Injuries or Killed 2147 (39%)	Light Injuries 3410 (61%)
Adjusted Average		
Daily Traffic Amount		
Low Volume Traffic	1247 (40%)	1866 (60%)
High Volume Traffic	900 (37%)	1544 (63%)
Crash Time		
Off Peak Hour	1353 (41%)	1936 (59%)
Peak Hour	794 (35%)	1474 (65%)
Light Condition		
Daylight	1334 (34%)	2597 (66%)
Dark, Lighted	565 (49%)	580 (51%)
Dusk	27 (47%)	30 (53%)
Dark, Not Lighted	179 (55%)	147 (45%)
Dawn	21 (45%)	26 (55%)
Dark, Unknown Lighting	21 (41%)	30 (59%)
Road Class		
Interstate	663 (39%)	1053 (61%)
US & State Highways	1319 (38%)	2144 (62%)
Farm to Market	157 (43%)	204 (57%)
City Street	8 (47%)	9 (53%)
Surface Width		
Low Width	1100 (39%)	1705 (61%)
High Width	1047 (38%)	1705 (61%)
Weather Condition		
Clear	1539 (39%)	2380 (61%)
Cloudy	323 (37%)	543 (63%)
Rain	271 (37%)	466 (63%)
Fog	5 (42%)	7 (58%)
Sleet/Hail	8 (38%)	13 (62%)
Snow	1(100%)	0 (0%)
Blowing Sand/Snow	0 (0%)	1 (100%)

5 Results

Using RStudio, the CART method was employed to classify the crash severity. To recognize the crucial factors of injury severity, six independent variables were used. Additionally, The Gini index was used as the CART's default splitting criterion.

Fig. 1 shows the classification tree. As it can be easily distinguished, this tree has five terminal nodes and the Light Condition, Crash Time and Weather Condition are the basic splitters. This implies that the crucial factors in crash severity in Beaumont's crashes are these three variables. The first split in node 1 is based on the most important factor light condition, which points out the most appropriate variable to classify the crash severity base on the dataset. CART splits the light conditions into dark-lighted, dark-not lighted, dawn or dusk in the left node and dark-unknown lighting and

daylight in the right node. In fact, the tree predicts that if a crash happens in the light condition of dark with unknown lighting or daylight, 34% of the crashes will cause serious injuries or killed and 66% will cause light injuries (Terminal node 5).

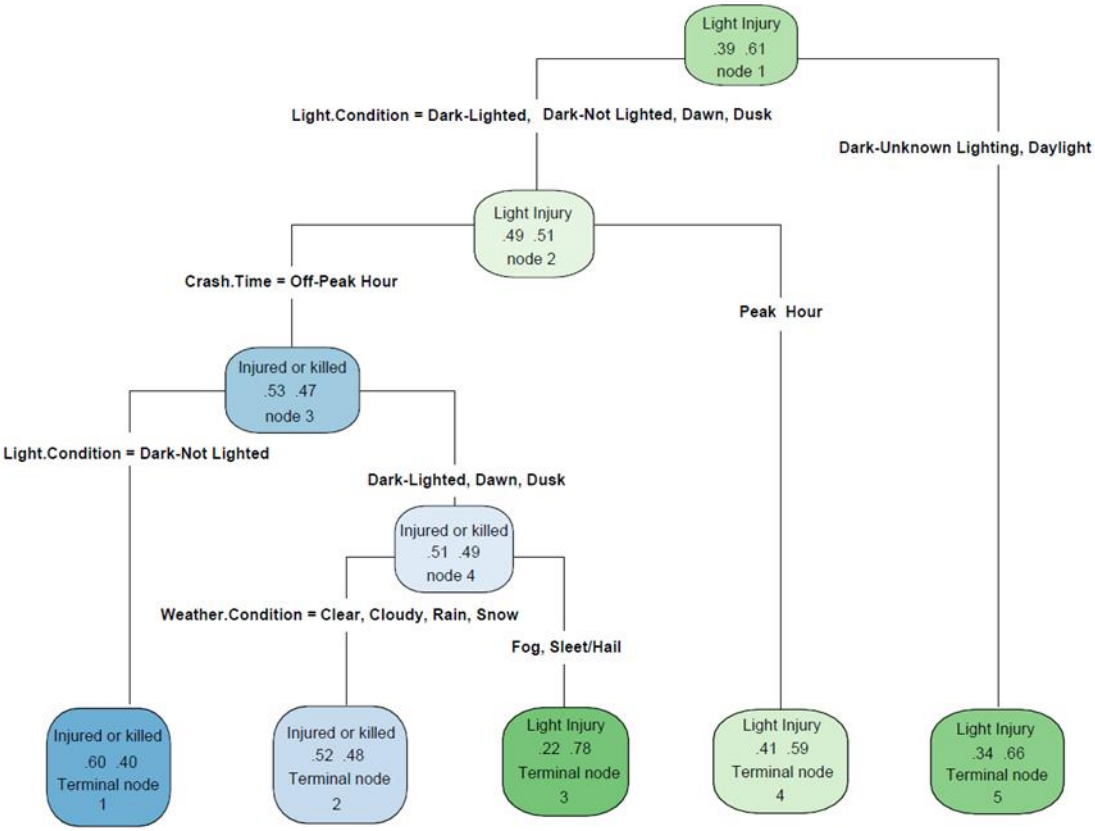

Fig. 1. The output of the CART tree

Yet, in the left branch of the tree, CART continues to split node 2 based on the variable of crash time and divide the crash time into peak hour and off-peak hour. It forecasts that in the light condition of dark-lighted, dark-not lighted, dawn or dusk and the crash time of peak hour, 41% of the crashes will cause serious injuries or killed and 59% will lead to light injuries (Terminal node 4). Regarding to the crash time of off-peak hour and light condition node 3 is formed. CART splits node 3 based on light condition and sends the dark-lighted, dawn and dusk to the right branch which forms node 4 and the rest of light conditions to the left branch which makes terminal node 1. Therefore, as indicated in terminal node 1 if the light condition is dark-not lighted and the crash time is not in peak hours (off-peak hours), the tree predicts that the injury severity is more likely to be severe and will cause severe injuries or people killed (60%). Moreover, the CART divides nod 4 based on weather condition into two parts. At terminal node 2, tree predicts that 52% of the crashes in clear, cloudy, rain or snow weather condition have severe injuries while in weather conditions of fog, sleet or hail (Terminal 3) 78% of crashes will cause light injuries. Besides, applying the CART method, table 3 represents the accuracy of the prediction which is made by this decision tree for testing models. As the table 3 indicates the model prediction accuracy for the testing data is 62% which means the CART model with the probability of 62% predicts the future data properly. As, Xu et al. [25] achieved the prediction accuracy of 55.2% for testing data and Wang et al. [21] got 62% overall model prediction accuracy in the testing phase, this model accuracy, is acceptable. Table 3 gives general information about the accuracy and performance of the model but should not be considered as the singular measure for assessing model performance.

Table 3. Prediction accuracy of the CART model for two classes of severity injuries

	Testing data	
	Correctly predicted	Observed severity
Serious Injuries or Killed	16 (64%)	25
Light Injuries	453 (61%)	733
Overall	469 (62%)	758

6 Discussion

CART results show that light condition, crash time and weather condition are the three most influential factors to crash severity. The output of terminal node 1 shows that in off-peak hours and dark-not lighted conditions the probability of a severe crash is bigger than a light one. Chen and Fan [27] implied in their research that crashes in daytime and off-peak were not severe. These findings show that low visibility in dark conditions is highly likely to increase the severity of crashes.

In terminal nodes 2 and 3 which are divided regarding weather condition, clear, cloudy, rainy and snowy weather are associated with severe injuries and other weather conditions are in association with light injuries. While some approaches confirm this study's result, others have conflicts with these outputs. Likely, regarding each investigation characteristic, many factors can play roles and could affect the results. For example, the study of Tavakoli Kashani and Shariat Mohaymany [8] confirmed that weather condition of clear, snowy and foggy is associated with serious injuries. Also, Nilsson et al. [49] analysis mentioned that adverse weather conditions enhance the risk of fatal run-off-road crashes. However, Dissanayake [30] argued that severe weather and physical disabilities do not significantly affect single vehicle crashes which is on contrary with other researches. Since the percentage of crashes in terminal node 3 is only 0.17% of the whole data, it can be concluded that clear, cloudy, rainy and snowy weather can increase the severity of injuries in this research.

According to terminal node 4, crashes in peak hours, and dark light conditions can cause less severe injuries. Chu [50] found similar results in his investigation that crashes during peak time are less severe in comparison with those during the off-peak time. This can be associated with the fact that drivers may drive at a higher speed in off-peak hours when the traffic volume is smaller than off-peak hours.

Furthermore, terminal node 5 reveals that driving in daylight time can cause more light injuries. This may be because, lighting increases drivers' ability to see the scenes properly and respond rapidly and in an appropriate manner if they detect any danger. Driving in poor light conditions may cause drivers to ignore the presence of traffic signs or pedestrians which is stated by a study of Li [51].

7 Conclusion

The purpose of this investigation is to identify the major factors of crash severity in Beaumont, Texas, and give some evidence-based recommendations to policymakers to alleviate the effects of crashes. This study showed that “light condition”, “crash time” and “weather condition” are the most crucial factors influencing the injury severity of crashes in Beaumont. The output indicates Beaumont roads have inadequate or insufficient lightning condition. Crash time is another main recognized variable. To be more precise, most probably drivers in peak hours, drive more carefully and as a result, the injuries are less severe in Beaumont. Also, the weather condition is found to be another important factor that causes severe injuries or fatality. The analysis revealed that in clear, cloudy and rainy weather, which is the dominant weather condition in Beaumont, more fatality and serious injuries will occur.

However, this research is implemented in Beaumont which is a port with large number of heavy vehicles on its roads. The result of this study can be taken into consideration to reduce the crash severity in new humid subtropical climate port and coastal urban areas. Additionally, regarding the complexity of transportation and traffic crashes, some future investigations are suggested. First, using other data mining methods, may help to extract additional risk factors and information. Combining human related with road-based factors can be the next step, as well. It can also help researchers and policy makers to achieve better understanding of traffic crashes and as a result assist decision-makers to make more efficient and cost-effective decisions.

Acknowledgments

This study was partially supported by the Natural Science Foundation (1726500) and the Center for Advances in Port Management (CAPM). The findings and conclusions of this paper are those of the authors and do not necessarily represent the official position of NSF and CAPM.

References:

[1] World Health Organization. Global Health Observatory data repository. *Road traffic deaths data by country*, 2019. Available from: https://apps.who.int/gho/data/node.main.A997

[2] Texas Department of Transportation. Crash Record Information System. *TxDOT Crash Query Tool*, 2019. Available from: https://cris.dot.state.tx.us/public/Query/app/welcome

[3] Fanny M, Norros I, Innamaa S. Accident risk of road and weather conditions on different road types. *Accident Analysis & Prevention.* 2019;122: 181–188.

[4] Taamneh M, Alkheder S, Taamneh S. Data mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *Journal of Transportation Safety & Security*. 2017; 9(2):146–166.

[5] Pakgohar A, Sigari Tabrizi R, Khalil M, Esmaeili A. The role of human factor in incidence and severity of road crashes. *Procedia Computer Science*, 2011. P. 764–769.

[6] Ossenbruggen P, Pendharkar J, Ivan J. Roadway safety in rural and small urbanized areas. *Accident Analysis and Prevention.* 2001;33: 485–498.

[7] Chang L-Y., Wang H-W. Analysis of traffic injury severity: An application of non-parametric classification tree techniques . *Accident Analysis and Prevention*. 2006;38: 1019–1027.

[8] Tavakoli Kashani A, Shariat Mohaymany A. Analysis of the traffic injury severity on two-lane, two-way rural roads based. *Safety Science*, 2011;49(10): 1314–1320.

[9] Shirali G, Valipour Noroozi M, Saki Malehi A. The outcome of occupational accidents by CART and CHAID. *Journal of Public Health Research.* 2018; 7(1361): 74–80.

[10] Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression. Montery: Wadsworth and Brooks/Cole; 1984.

[11] American Association of Port Authorities (APPA). 2013. *U.S. Port Ranking by Cargo Volume 2013.* United States of America.

[12] Hegar G. Port of entery, Port of Beaumont, *Texas Comptroller of Public Accounts*. 2018. Available from: https://comptroller.texas.gov/economy/economic-data/ports/snap-beaumont.php

[13] Dick J. Port of Beaumont plans new truck-queuing station, *Beaumont Enterprise*. 2020. Available from: https://www.beaumontenterprise.com/news/article/Port-of-Beaumont-plans-new-truck-queuing-station-15372117.php

[14] Brooks S. TxDOT awards Port of Beaumont $1.57 million grant, *Beaumont Business Journal.* 2020. Available from: https://www.beaumontbusinessjournal.com/news/txdot-awards-port beaumont-157-million-grant

[15] Freedman A, Samenow J. Flooded again: Climate change is making flooding more frequent in Southeast Texas. *The Washington Post*, 2019. Available from: https://www.washingtonpost.com/weather/2019/09/20/flooded-again-climate-change-is-making-flooding-more-frequent-southeast-texasthanks-part-climate-change/

[16] Washington S P, Karlaftis M G, Mannering F. Statistical and Econometric Methods for Transportation Data Analysis (2nd Edition ed.). Chapman and Hall/CRC; 2010.

[17] Chang L, Chen, W. Data mining of tree-based models to analyze freeway accident frequency. *Journal of Safety Research.* 2005;36: 365–375.

[18] Huang H, Peng Y, Wang J, Luo Q, Li X. Interactive risk analysis on crash injury severity at a mountainous freeway with tunnel groups in China. *Accident Analysis and Prevention.* 2018;111: 56–62.

[19] Rakotonirainy A, Steinhardt D, Delhomme P, Darvell M, Schramm A. Older drivers' crashes in Queensland, Australia. *Accident Analysis and Prevention.* 2012;48: 423–429.

[20] Mergia W Y, Eustace D, Chimba D, Qumsiyeh M. Exploring factors contributing to injury severity at freeway merging and diverging locations in Ohio. *Accident Analysis and Prevention.* 2013; 55: 202–210.

[21] Wang J, Zheng Y, Li X, Yu C, Kodaka K, Li K. Driving risk assessment using near-crash database through data mining of tree-based model. *Accident Analysis and Prevention.* 2015; 84:54–64.

[22] Karlaftis M G, Golias I. Effects of road geometry and traffic volumes on rural roadway accident rates. *Accident Analysis and Prevention.* 2002;34: 357–365.
[23] Öström M, Eriksson A. Pedestrian fatalities and alcohol. *Accident Analysis and Prevention.* 2001;33(2): 173–180.
[24] Castro Y, Kim Y J. Data mining on road safety: factor assessment on vehicle accidents using classification models. *International Journal of Crashworthiness*. 2015; 21(2): 1–7.
[25] Xu X, Šaric Ž, Kouhpanejade A. Freeway incident frequency analysis based on CART method. *Promet – Traffic & Transportation.* 2014;26: 191–199.
[26] Ma Z, Chien S I-J, Dong C, Hu D, Xu T. Exploring factors affecting injury severity of crashes in freeway tunnels. *Tunnelling and Underground Space Technology*. 2016;59: 100–104.
[27] Chen Z, Fan W D. A multinomial logit model of pedestrian-vehicle crash severity. *International Journal of Transportation.* 2019;8: 43–52.
[28] Prati G, Pietrantoni L, Fraboni F. Using data mining techniques to predict the severity of bicycle crashes. *Accident Analysis and Prevention.* 2017;101: 44–54.
[29] Carlin J B, Taylor P, Nolan T. School based bicycle safety education and bicycle injuries in children: a case-control study. *Injury Prevention.* 1998;4: 22–27.
[30] Dissanayake S. Young Drivers and Run-Off-the-Road Crashes. *Proceedings of the 2003 Mid-Continent Transportation Research Symposium;* 2003. P. 1–6.
[31] Qiong W, Guohui Z, Yusheng C, Lina, W, Rafiqul, A T, Adélamar A. Exploratory multinomial logit model–based driver injury severity analyses for teenage and adult drivers in intersection-related crashes. *Traffic Injury*. 2016;17(4):1–9.
[32] Koetse M J, Rietveld P. The impact of climate change and weather on transport: An overview of empirical findings. *Transportation Research Part D*. 2009;14: 205–221.
[33] Çelik A K, Oktay E. A multinomial logit analysis of risk factors influencing road trafficinjury severities in the Erzurum and Kars Provinces of Turkey. *Accident Analysis and Prevention.* 2014;72: 66–77.
[34] Fan W D, Kane M R, Haile E. Analyzing severity of vehicle crashes at highway-rail grade crossings: multinomial logit modeling. *Journal of the Transportation Research Forum.* 2015;54(2): 39–56.
[35] Moore D N, Schneider IV W H, Savolainen P T, Farzaneh M. Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. *Accident Analysis and Prevention.* 2011;43: 621–630.
[36] Tay R, Choi J, Kattan L, Khan A. A multinomial logit model of pedestrian–vehicle crash severity. *International Journal of Sustainable Transportation.* 2011;5: 233–249.
[37] Li Z, Ci Y, Chen C, Zhang G, Wu Q, Qian Z. Investigation of driver injury severities in rural single-vehicle crashes under conditions using mixed logit and latent class models. *Accident Analysis and Prevention.* 2019;124: 219–229.
[38] Wu Q, Zhang G, Ci Y, Wu L, Tarefder R A. Exploratory multinomial logit model–based driver injury severity analyses for adult drivers in intersection-related crashes. *Traffic Injury Prevention.* 2016;4(17): 413–422.
[39] Iranitalab A, Khattakb A. Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis and Prevention.* 2017;108: 27–36.
[40] Zeng Q, Huang H, Pei X, Wong S C. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. *Analytic Methods in Accident Research.* 2016;10: 12–25.
[41] Tang J, Liang J, Han C, Li Z, Huang H. Crash injury severity analysis using a two-layer stacking framework." *Accident Analysis and Prevention.* 2019; 122: 226–238.
[42] Abellán J, López G, de Ona J. Analysis of traffic accident severity using Decision Rules via Decision. *Expert Systems with Applications.* 2013;40: 6047–6054.
[43] Kuhnert P M, Do K-A, McClure R. Combining non-parametric models with logistic regression: an application to motor vehicle injury data. *Computational Statistics & Data Analysis.* 2000;34: 371–386.
[44] Young Sohn S, Shin H. Pattern recognition for road traffic accident severity Korea. *Ergonomics.* 2001;44(1): 107–117.
[45] Wikipedia, Cimate of Beaumont, Texas, 2021. Available from: https://en.wikipedia.org/wiki/Climate_of_Beaumont,_Texas
[46] Rovšek V, Batista M, Bogunović B. Identifying the key risk factors of traffic accident injury severity on slovenian roads using a non-

parametric classification tree. *Transport.* 2017; 32(3): 272–281.

[47] Stewart J R. Applications of classification and regression tree method in roadway safety study. *Transportation Research Record*. 1996;1542(1): 1–5.

[48] Iacobucci D, Posavac S S, Kardes F R, Schneider M J, Popovich D L. The median split: Robust, refined, and revived. *Journal of Consumer Psychology*. 2015;25(4): 690–704.

[49] Nilsson D, Lindman M, Victor T, Dozza M. Definition of run-off-road crash clusters—For safety benefit estimation and driver assistance development. *Accident Analysis and Prevention.* 2018;113: 97–105.

[50] Chu H-C. Assessing factors causing severe injuries in crashes of high-deck busesin long-distance driving on freeways. *Accident Analysis and Prevention*, 2014;92: 130–136.

[51] Li G. Big data based exploration of risk factors to traffic crashes in southeast Texas and an experimental validation. *Doctoal thesis*. Lamar University Beaumont, Texas; 2019.