Modeling Injury Severity of Vehicular Traffic Crashes

Randa Oqab Mujalli
Department of Civil Engineering
The Hashemite University
Zarqa, Jordan
randao@hu.edu.jo

Griselda López
Highway Engineering Research Group
Polytechnic University of Valencia Valencia, Spain
grilomal@tra.upv.es

Laura Garach
Department of Civil Engineering
University of Granada
Granada, Spain
lgarach@ugr.es

***Abstract*— Data mining techniques constitute an alternative approach that has received increasing attention from researchers in recent years in road safety analysis field. In this paper Bayesian networks were used to develop models in order to identify factors that affect the injury severity of a crash within urban areas based on traffic crashes data on Jordanian roads collected for three years (2009-2011). The following variables were found to have a significant effect on classifying crashes according to their injury severity: lighting of roadway, crash type, road type, crash manner, surface condition, number of lanes, number of vehicles, gradient, type of pavement, traffic control devices, and speed limit. The results of this research can be used to determine the factors that should be taken into consideration when designing a new roadway or for improving safety of existing roads.**

Keywords- Data mining; traffic crashes; injury severity.

I. INTRODUCTION

One of the major problems that affect countries worldwide are fatalities and severe injuries resulting from traffic crashes, which negatively affect economies especially in low and middle income countries. Attempts to understand the reasons behind the occurrence of fatalities and severe injuries were of utmost importance for safety analysts, and still there are many reasons or causes of these outcomes yet to be discovered [1]. According to World Health Organization (WHO), crashes resulting in fatalities are estimated to be 1.25 million persons annually worldwide. As a result of road these traffic crashes, 50 million people incur nonfatal injuries each year. Cost of traffic accidents is estimated to be 518 billion US dollars representing (1-3%) of Gross Domestic Product (GDP) worldwide [2].

In Jordan which is a middle income country with a population of 9.531 million [3], 111057 traffic crashes occurred in 2015 with 9712 crashes resulting in casualties which resulted in 608 fatalities and 2021 sever injuries with an estimated cost of 388 million US dollar [4].

Crashes occurring in urban areas are considered to be more dangerous than those occurring elsewhere, where the probability of occurrence for highly severe or fatal injuries in urban areas is eight times larger than slight or no injury; on the other hand, fatal and severe injuries are 2.5 times more probable to occur in rural areas than slight or no injuries [5].

Reference [6] compared crashes severity between inside and outside urban areas. They found that factors affecting crashes severity inside urban areas included young driver age, bicyclists, intersections, and collision with fixed objects, whereas factors affecting severity outside urban areas were weather conditions, head-on and side collisions.

Data mining techniques are of the methods currently in use by safety analysts and researchers due to their ability to deal with large data bases. Some of the most used types of data mining techniques are association rules [7], [8] and decision trees [9], [10] in which they were both used to define the circumstances under which crashes are more probable to occur. One of the methods that is trending in analyzing the severity of traffic crashes is Bayesian networks (BNs). One of the first researchers to use BNs for the analysis of traffic crashes severity was [11], where they proved that BNs is a powerful method that could be used to analyze the traffic crash records in order to find put the factors that causes a specific injury. In an attempt to simplify the process of analysis using BNs, [12] used data mining selection algorithms that enables the researchers to limit the number of variables used in their model by only modeling those variables with the most significant effect on injury severity of a traffic crash. Recently [13] used BNs to model traffic crashes on Jordanian roads.

In this work, factors affecting injury severity of urban crashes in Jordan are analyzed. For this purpose, BNs are used to develop different models. Finally, the models

developed are compared, and the results for the best model are illustrated and discussed.

II. Methodology

Traffic crashes database from Police Security Directorate (PSD) was used to develop BNs models. According to [8] and [13] a common problem that is encountered in crashes databases is that they include fewer records for fatal and severe injury crashes than for slight injury crashes in which the database is then said to be imbalanced. A database is considered to be imbalanced if one of the classes (minority class) contains a much smaller number of instances than the remaining class (majority class) [14]. This problem affects learning from data process and results in high predictive accuracy over the majority class, but poor predictive accuracy over the minority class [15]. To deal with this problem, oversampling methods could be used prior to developing models, in which these methods are aimed at balancing class populations through creating new samples from the minority class and adding them to the database [14], [16].

In this study, a balanced database was developed from the imbalanced database using oversampling methods. Using the balanced database, BNs were developed in order to analyze injury severity of crashes occurring on urban Jordanian roads. The developed models were compared to each other using 10- folds cross validation method, where the database was first divided into 10 subsets, nine were used to train the model and the remaining one subset was used to test the model. The process was repeated ten times and the average was obtained. As a result, 9 models were developed and compared.

A. Data

Historical crashes records over a period of 3 years (2009-2011) for crashes that occurred on urban roads in Jordan were obtained from the (PSD) with a total of 16,815 crashes excluding run-off-road crashes and property damage only crashes.

Fourteen variables were used in analysis (Table 1), where these variables were chosen based on previous studies in this field [6], [17]. The included variables described the prevailing conditions at the time of the occurrence of the crash.

Based on previous research in this field [10], [13], [18], [19] the injury severity was determined according to the level of injury to the worst injured occupant. Two classes of injury were used: slight injuries (SI) and fatality or severe injuries (FS).

In the imbalanced database, there were 13,725 slight injuries and 3,090 fatalities and severe injuries which indicate that the database was highly imbalanced.

TABLE I. DESCRIPTION OF THE VARIABLES AND CLASSIFICATION BY SEVERITY

Code: Variable	Description
NOV: no. of vehicles	1 vehicle 2 vehicles 3 vehicles 4 vehicles
CT: Crash type	Animal Static object Thrown out vehicle Other Motorcycle
CM: Crash manner	Roundabout Head-on Intersection maneuver rear-end single vehicle other
RT: Road type	1 way 2 way divided 2 way undivided other
NOL: Number of lanes	1lane 2lanes 3lanes 4 lanes
ALI: Alignment	straight curve
SL: Road gradient	upgrade level downgrade
PA: Pavement type	Asphalt Concrete other
SC: Surface condition	Ice or snow Dry Mud, sand or oil Wet
ATM: adverse weather	Clear Rain Snow, storm wind, for or dust
LI: lighting	Dark Daylight Night good light Night bad light Other
TCD: Traffic control	Flashing light No control Regulatory sign Marking Police Police with traffic signal Stop Traffic signal
SPL: speed limit	

	20
	30
	40
	50
	60
	70
	80
	90
	≥100

B. Re-sampling techniques

According to [20] if the number of records in each class of the target variable is not approximately equal, then the database is said to be imbalanced.

In order to balance the class variable (injury severity) a re-sampling techniques was applied as a preprocessing step. Weka's preprocess supervised filter [21] was used to perform the re-sampling on the database. Synthetic minority oversampling technique (SMOTE) was used which is a heuristic method that creates a subset of the original dataset by creating synthetic minority examples [22].

C. Bayesian networks

Let X={X1, . . . ,Xn}, n≥1 be a set of variables. BN over a set of variables X is a network structure, which is a Directed Acyclic Graph over X and a set of probability tables Bp = {p(Xi|pa(Xi), Xi X)} where pa(Xi) is the set of parents or antecedents of Xi in BN and i=(1,2,3,….,n). A BN represents joint probability distributions $\prod_{X_i \in X} p(X_i \mid pa(X_i))$ [23].

In this study, we used BNs in order to develop different models and to compare their results in terms of their ability to correctly classify crashes according to their injury severity into either FS or SI. When building the models using BNs, three search methods were used: hill climber, hill climber algorithm restricted by an order on the variables (K2) and simulated annealing search algorithm. Also, three different score metrics functions were used: BDe score metric (BDeu); Minimum Description Length (MDL); and the Akaike Information Criterion (AIC).

D. Performance evaluation

In order to evaluate the performance of the different developed models, using a confusion matrix the results of correctly and incorrectly predicted records for each class are illustrated. The performance measures used in this study were accuracy, sensitivity, specificity. Their equations are as follows:

$$Accuracy = \frac{TSI + TFS}{TSI + FSI + TFS + FFS} \quad (1)$$

$$Sensitivity = \frac{TSI}{TSI + FFS} \quad (2)$$

$$Specificity = \frac{TFS}{TFS + FSI} \quad (3)$$

Where,

TSI: the number of SI instances correctly classified
TFS: the number of FS instances correctly classified
FSI: the number of SI instances incorrectly classified
FFS: the number of FS instances incorrectly classified

Area under a Receiver Operating Characteristic (ROC) curve is also used as a performance measure. ROC curve represents the sensitivity against (1-specificity). ROC curves are more useful as descriptors of overall performance, reflected by the area under the curve, with a maximum of one describing a perfect test and a ROC area of 0.50 describing a valueless test.

I. Discussion and conclusion

The original database included was highly imbalanced and to solve this problem, a balanced database was developed using oversampling methods. Table 2 describes both balanced and imbalanced databases and shows that after balancing the database the size of the FS class was increased to be equal to that of the SI class.

TABLE II. NUMBER OF ACCIDENTS AND SEVERITY DISTRIBUTION IN THE DIFFERENT DATASETS.

Database	Total	SI	FS
Original	16815	13725	3090
Oversampled	27450	13725	13752

As shown in Table 3, 9 models were developed and comparing these models it is shown that the best results were obtained by using simulated annealing search algorithm with AIC score. This model had the largest values obtained in terms of Accuracy, sensitivity and ROC Area and hence will be used to analyze traffic crashes.

TABLE III. PERFORMANCE MEASURES RESULTS AS OBTAINED USING DIFFERENT SEARCH ALGORITHMS WITH SCORES.

Performance measure	Measure value	Standard deviation ±
	Accuracy	
Hillclimber + Bdeu	0.619	0.084
Hillclimber + MDL	0.616	0.089
Hillclimber + AIC	0.623	0.091
k2 + Bdeu	0.618	0.077
k2 + MDL	0.619	0.080
k2 + AIC	0.623	0.082
Simulated + Bdeu	0.619	0.079
Simulated + MDL	0.614	0.079
Simulated + AIC	0.628	0.088
Performance measure	Sensitivity	
Hillclimber + Bdeu	0.620	0.030
Hillclimber + MDL	0.620	0.030
Hillclimber + AIC	0.610	0.040
k2 + Bdeu	0.640	0.010
k2 + MDL	0.640	0.010
k2 + AIC	0.590	0.030
Simulated + Bdeu	0.640	0.010
Simulated + MDL	0.640	0.010
Simulated + AIC	0.650	0.010
Performance measure	Specificity	
Hillclimber + Bdeu	0.620	0.030
Hillclimber + MDL	0.610	0.030
Hillclimber + AIC	0.640	0.030
k2 + Bdeu	0.600	0.020
k2 + MDL	0.600	0.010
k2 + AIC	0.650	0.030
Simulated + Bdeu	0.600	0.020
Simulated + MDL	0.590	0.010
	ROC Area	
Simulated + AIC	0.610	0.020
Hillclimber + Bdeu	0.660	0.010
Hillclimber + MDL	0.660	0.010
Hillclimber + AIC	0.680	0.010
k2 + Bdeu	0.670	0.010
k2 + MDL	0.660	0.010
k2 + AIC	0.680	0.010
Simulated + Bdeu	0.670	0.010
Simulated + MDL	0.660	0.010
Simulated + AIC	0.680	0.010

In order to identify the variables that significantly affect resulting injury severity outcome in a crash; BNs were developed using simulated annealing search with AIC score. Fig. 1 shows the relationships between injury severity (INJ) and the rest of the variables as well as interdependences amongst the different variables.

As illustrated in Fig.1, eleven variables had a direct arc with INJ: no. of vehicles involved (NOV), crash type (CT), crash manner (CM), road type (RT), number of lanes (NOL), road gradient (SL), pavement type (PA), surface condition (SC), speed limit (SPL), traffic control (TCD) and lighting (LI).

These variables were found to affect injury severity by many researchers; where [13], [24], [25] found that road condition and speed limit is directly related to injury severity. Type of crash was found to affect injury severity by [6], [13], [11], [25] and most recently by [13], and [24].

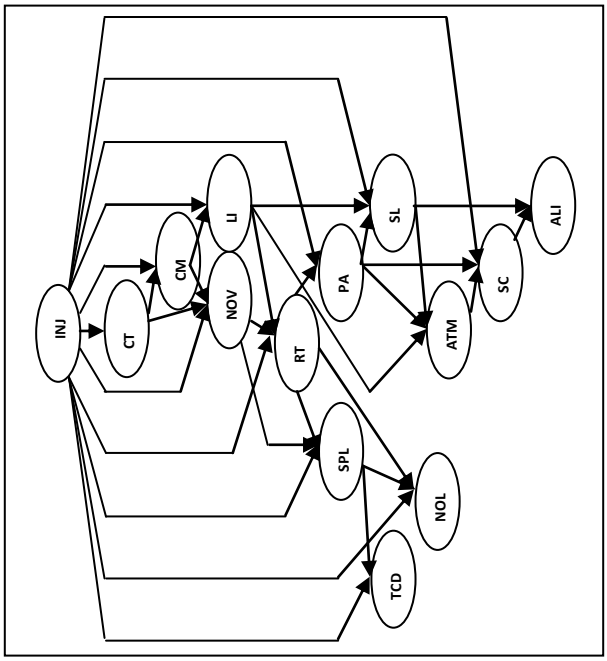

Figure 1. BN model developed using simulated annealing search algorithm with AIC score.

According to [25] and [26] the number of vehicles was found to be significantly associated with the resulting injury severity, while lighting was found to be significant by both [11] and [27]. Ma et al. [27] found that road alignment was one of the contributing variables that affect severity while [6] showed that time at which the crash occurred had also a significant effect.

The conclusion of this study is that number of vehicles, crash type, crash manner, road type, number of lanes, gradient, pavement type, surface condition, adverse weather, speed limit, traffic control and lighting were all found to have a significant effect on a crash occurring within urban areas and hence affects its resulting injury severity.

It is noticed that most of the variables found to affect injury severity of traffic crashes within urban areas are mostly related to either design of roads, such as road type (RT), number of lanes (NOL), gradient (SL), pavement type (PA) or operation and management of roads such as: surface condition (SC), speed (SPL), traffic control (TCD) and lighting (LI). Thus, indicating the need to enhance the current condition of existing roads and to take into consideration these factors when designing new roads prior to construction in order to enhance road safety and to reduce the occurrence of fatal and severe injuries crashes.

ACKNOWLEDGMENT

The authors are grateful to the Police Security Directorate in Jordan for providing the data necessary for this research.

REFERENCES

[1] Kwon, O.H., Rhee, W., Yoon, Y. 2015. Application of classification algorithms for analysis of road safety risk factor dependencies, Accident Analysis and Prevention, 75: 1-15.

[2] World Health Organization (WHO). 2015. Global status report on road safety. Available from Internet: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/

[3] Department of Statistics (DOS). 2013. Available from internet: <http://web.dos.gov.jo/>.

[4] Police Security Directorate (PSD). 2013. Available from internet: https://www.psd.gov.jo/images/docs/studyAcc2015.pdf.

[5] Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F. 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis, Accident Analysis and Prevention, 37: 910–921.

[6] Theofilatos, A., Graham, D., Yannis, G. 2012. Factors affecting accident severity inside and outside urban areas in Greece, Traffic Injury Prevention, 13: 458-467.

[7] Pande, A., Abdel-Aty, M. 2009. Market basket analysis of crash data from large jurisdictions and its potential as a decision supporting tool, Safety Science, 47: 145-154.

[8] Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F. 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery, Accident Analysis and Prevention, 49: 58-72.

[9] López, G., de Oña, J., Abellán, J. 2012. Using Decision Trees to extract Decision Rules from Police Reports on Road Accidents, Procedia-Social and Behavioral Sciences, 53: 106-114.

[10] De Oña, J., López, G., Abellán, J. 2013. Extracting decision rules from police accident reports through decision trees, Accident Analysis and Prevention, 50: 1151-1160.

[11] De Oña, J., Mujalli, R.O., Calvo, F.J. 2011. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks, Accident Analysis and Prevention, 43: 402-411.

[12] Mujalli, R.O., De Oña, J. 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks, Journal of Safety Research, 42: 317-326.

[13] Mujalli, R.O.; López, G.; Garach, L. 2016. Bayes classifiers for imbalanced traffic accidents datasets, *Accident Analysis and Prevention*, 88: 37–51.

[14] Stefanowski, J., Wilk, S. 2008. Selective pre-processing of imbalanced data for improving classification performance. In Proceedings of 10th International conference on Data Warehousing and Knowledge Discovery, 283-292.

[15] Thammasiri, D., Delen, D., Meesad, P., Kasap, N. 2014. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition, Expert Systems with Applications, 41: 321-330.

[16] Błaszczyński, J., Stefanowski, J. 2015. Neighbourhood sampling in bagging for imbalanced data," Neurocomputing, 150: 529-542.

[17] Pahukula, J., Hernandez, S., Unnikrishnan, A. 2015. A time of day analysis of crashes involving large trucks in urban areas," Accident Analysis and Prevention, 75: 155-163.

[18] Abellán, J., De Oña, J., López, G. 2013. Analysis of traffic accident severity using decision rules via decision trees, Expert Systems with Application, 40: 6047-6054.

[19] 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.

[20] Crone, S., Finlay, S. 2012. Instance sampling in credit scoring: An empirical study of sample size and balancing, International Journal of Forecasting, 28: 224-238.

[21] Witten, I.H., Frank, E.2005. Data mining: Practical machine learning tools and techniques, 2nd ed. Morgan Kaufmann Inc. San Francisco, CA, USA. 560 p.

[22] Chawla, N., Hall, L., Bowyer, K., Kegelmeyer, W. 2002. SMOTE: synthetic minority oversampling technique, Journal of Artificial Intelligence Research, 16: 321-357.

[23] Mittal, A., Kassim, A. Tan, T. 2007. Bayesian network technologies: Applications and graphical models, IGI Publishing. New York. 368 p.

[24] Manner, H., Wünsch-Ziegler, L. 2013. Analyzing the severity of accidents on the German Autobahn," Accident Analysis and Prevention, 57: 40-48.

[25] Kadilar, G.O. 2016. Effect of driver, roadway, collision, and vehicle characteristics on crash severity: a conditional logistic regression approach, International Journal of Injury Control and Safety Promotion, 23:135-144.

[26] Kockelman, K.M. Kweon, Y.J. 2002. Driver injury severity: An application of ordered probit models, Accident Analysis and Prevention, 34: 313-321.

[27] Ma, Z., Shao, C., Yue, H., Ma, S. 2009. Analysis of the logistic model for accident severity on urban road environment. In Proceedings of (IV) IEEE Intelligent Vehicles Symposium, 983-987.