EXAMINING SEVERE TRAFFIC CRASHES IN RIYADH USING DIFFERENT DATA MINING METHODS

Hany M. Hassan, King Saud University, Prince Mohamed Bin Naif Chair for Traffic Safety Research, P.O. Box 800, Riyadh 11421, Saudi Arabia

Loukas Dimitriou, King Saud University, Prince Mohamed Bin Naif Chair for Traffic Safety Research, P.O. Box 800, Riyadh 11421, Saudi Arabia

Mohamed Abdel-Aty, University of Central Florida, Department of Civil, Environmental and Construction Engineering, Orlando, FL 32816-2450, United States

Hesham Al-Faleh, King Saud University, Prince Mohamed Bin Naif Chair for Traffic Safety Research, P.O. Box 800, Riyadh 11421, Saudi Arabia

ABSTRACT

Recently, growing concern has been shifted towards traffic safety in the kingdom of Saudi Arabia (KSA). KSA has a unique state regarding traffic safety problems. KSA can be classified as one of the developed countries in terms of the magnitude and quality of the road network available and compatible with international standards. However, it can be also considered as one of the developing countries as the rate of increase in the number of road accidents in the KSA is substantial compared with the relevant figures of other developing countries and other countries of the Gulf region and hence, more research efforts are still needed. This paper aims at better examining the nature and causes of fatal and serious traffic crashes in KSA so that remedies and/or future studies could be suggested.

For this purpose, data from 7470 reported fatal and serious traffic crashes occurred in Riyadh (the capital of KSA), during the period 2006-2011, were used in the analysis. Two data mining techniques (decision trees and random forests) were estimated to better identify the significant factors affecting the severity of traffic crashes. In addition, logistic regression model was developed to quantify the effects of significant variables affecting the dependent variable (crash severity).

The results revealed that crash type (i.e., hitting pedestrian), reason of crash (i.e., sudden lane change, speeding, distraction), point of collision (i.e., head on, rear, right angle, sideswipe), and weekday (i.e., day of the week where a crash occurred) were the significant factors affecting the binary target variable (fatal or serious crashes).

A comparison between these different techniques developed in this study in addition to practical suggestions on how to improve traffic safety in KSA are also discussed.

Keywords: *Crash severity, Traffic safety in developing countries, data mining techniques, logistic regression*

1. INTRODUCTION

In the Kingdom of Saudi Arabia (KSA), traffic crashes are among the greatest sources of external costs (in terms of fatalities and injuries, property damage, congestion effects and travel time). In the metropolitan complex of Riyadh in particular, an unusual number of car crashes occur. According to the Saudi official statistics, a total of 498,203 traffic crashes occurred in Saudi Arabia in 2010, with an increase of 8% compared with 2009. The number of injuries and fatalities reported in 2010 was 38,595 and 6,596, with an increase compared with 2009 of 11.5% and 7%, respectively. Additionally, approximately 8,947,955 traffic violations were reported in 2010, which represents a decrease of only 0.4% compared to 2009. These statistics indicate increases in the number of traffic crashes, injuries and fatalities and a slight decrease in the number of traffic violations.

The number of registered vehicles in KSA in 2010 were 1,264,327. This number increased to 1,434,152 in 2011 (11.84% increase).

KSA is composed of 13 administrative regions. The Riyadh region was first with respect to traffic safety problems, followed by Makkah and the Eastern Region. Approximately 26% of traffic crashes occurred in Riyadh between 2004-2009, followed by Makkah and the Eastern Region, with approximately 25% and 24%, respectively. These statistics are consistent with the distribution of the population (and most likely, the number of registered drivers) in Saudi Arabia.

Several studies have been conducted to examine traffic safety problems in Saudi Arabia. For instance, Koushki and Al-Ghadeer (1992) examined driver compliance with traffic regulations in Riyadh. The results indicated that drivers do not comply with traffic regulations everywhere (e.g., in urban or rural areas).

Al-Ghamdi (2002a) investigated the factors affecting pedestrian related crashes in Riyadh. Data reported by the police, from 638 pedestrian related crashes reported during the period 1997–1999 were used. The findings showed that about 77% of pedestrians involved in these crashes were struck while crossing a road where no crosswalk existed. In addition, it was found that about 34% of the fatal injuries were located on the head and chest.

In addition, Al-Ghamdi (2002b) evaluated ambulance response time in Riyadh and compared it with the corresponding times in other countries. The results revealed that the mean response time was 10.2 min, which is below the acceptable standards in developed nations like the UK and the US. Also, the time to serve one call takes, on average, 61.2 min with an 85th-percentile time of 66 min.

Moreover, Al-Ghamdi and AlGadhi (2004) assessed the effectiveness of using warning signs as countermeasures to camel–vehicle collisions in Saudi Arabia. The mean speed reduction of motorists passing these warning signs was the measure of effectiveness used in that study. The results indicated that after using such signs, the speed reduction ranged from 3 to 7 km/h.

Furthermore, Bendak (2005) investigated drivers' behaviour, personal characteristics and their relationship with respect to using seat belts using a questionnaire survey. The results showed that the rate of using seat belt in two Riyadh suburbs for drivers were 33% and 87%, respectively. However, for the front-seat passengers, the rates of using seat belt were only 4% and 41%.

Additionally, Al-Ghamdi (2007) evaluated the effectiveness of fog detection and warning system on speed and headway under reduced visibility conditions resulted from heavy fog. The findings of this study revealed that the mean speed throughout the experimental sections was reduced by about 6.5 km/h. However, the warning system was ineffective in reducing speed variability.

Considering these prior studies, the present study aims at contributing to the literature by identifying significant variables affecting the severity of crashes in Riyadh. To achieve this goal, logistic regression model and two data mining techniques (decision trees and random forests) were developed to identify and quantify the significant factors affecting the binary response variable (fatal vs. non-fatal crash).

2. DATA AND PRELIMINARY ANALYSIS

Generally in developing countries, crash data are difficult to obtain and rarely complete. The research team has made substantial effort to obtain and guarantee the completeness of the data.

According to the Saudi crash report, crash severity is classified into three levels: fatal crash, injury crash (no injury classification is available) and property damage only (PDO) crash. In KSA, crash reports are filled by the traffic departments if they resulted in fatal or injury crashes. However, if the crash resulted in PDO, the crash report is filled by a private insurance company (Najm) and in this case a short form of the Saudi crash report is filled. It was not possible for this study to obtain data for the PDO crashes or about the details on the degree of severity of crashes. Thus, only fatal and injury (non-fatal) crash records were considered for the purpose of this study.

The dataset used in this study consists of 7470 crash reports of fatal and injury crashes that have occurred in Riyadh, the capital city of Saudi Arabia, for the period from March 2006 to February 2011.

As shown in Table 1, only 14 variables were available. These variables are (1) crash severity, (2) number of vehicles involved in the crash, (3) number of injuries resulting from the crash, (4) day of the week, (5) crash time (day or night), (6) damages in private property, (7) damages in public property, (8) crash Location (not at intersections, at intersections, etc.), (9) road surface condition (dry or wet), (10) road lighting condition (with or without lighting), (11) weather, (12) reason of crash (distraction, Speeding, sudden lane change, etc.), (13) point of collision (head-on, rear end, etc.), (14) crash type (single vehicle, multiple vehicles, pedestrians).The distributions of all variables used in this study are provided in Table 1.

AS shown in Table 1, about 24% of the dataset were fatal crashes and about 76% were non-fatal crashes. Three or more vehicles were involved in approximately 12% of these crashes. Additionally, about 12% of these crashes resulted in four or more injuries. Regarding day of the week, It was found that the larger percentage of severe crashes in Riyadh occur during the day just before the weekend (in Saudi Arabia, Thursday and Friday are the weekend) with about 20%. Moreover, the results shown in Table 1 indicate that the larger percentage of the Saudi crashes (59%) occurs during the day time and 41% occurs during the night time.

Regarding damages in private and public property, the results indicate that about 91% of these crashes resulted in damages in private property (mainly damages in vehicles) and

about 7% of these crashes resulted in damages in public property (mainly damages in lighting poles).

Table 1: Distributions of dependent and explanatory variables

No.	Variables		Categories	Severe Crashes (n=7470)	
	Abbreviation	Description		Freq.	%
1	ACCISEV	Crash Severity	Fatal	1756	23.5
			Non-fatal (injury)	5714	76.5
2	Veh_No	No. of vehicles involved in the crash	1	3183	42.6
			2	3407	45.6
			3 or more	880	11.8
3	Injuries_No	No. of Injuries resulted from the crash	1	876	11.7
			2	4708	63.0
			3	1038	13.9
			4 or more	848	11.4
4	week_day_ch	Week Day	Saturday	995	13.3
			Sunday	1005	13.5
			Monday	998	13.4
			Tuesday	987	13.2
			Wednesday	1495	20.0
			Thursday	1108	14.8
			Friday	882	11.8
5	DayNight	Crash time	Day	4394	58.8
			Night	3076	41.2
6	Private_Damages	Private damages	Yes	6793	90.9
			No	677	9.1
7	Public_Damages	Public damages	Yes	513	6.9
			No	6957	93.1
8	Acc_location	Crash Location	Not at Intersections	7319	98.0
			At intersections	151	2.0
9	road_surface_cond	Road surface condition	Dry	7293	97.6
			Wet / others	177	2.4
10	Lighting_status	Road Lighting conditions	With lighting	6836	91.5
			without lighting	634	8.5
11	weather	weather	clear	7418	99.3
			others	52	0.7
12	ACC_Reason_ch	Crash Reason	Sudden lane change	4357	58.3
			Distraction	992	13.2
			Speeding	567	7.6
			Others	1554	20.9
13	collision_point_ch	Collision point	Head-on	1481	19.8
			Rear end	550	7.4
			Angle	2174	29.1
			Sideswipe	977	13.1
			Others / unknown	2288	30.6
14	Acc_type_ch	Crash type	Single vehicle	1084	14.5
			Multiple vehicles	4468	59.8

			Pedestrians	1844	24.7
			others	74	1.0

Concerning the crash location, it was found that the majority of the Saudi severe crashes (about 98%) occur at non-intersection locations. In addition, the results revealed that the majority of the Saudi severe crashes occur on dry roadway surfaces, at clear weather conditions and on lighted segments. Sudden lane change, distraction and speeding were the main reasons for crash involvement in Riyadh.

With respect crash type, the results revealed that the larger percentage (60%) of severe crashes in Riyadh occur due to multi-vehicles crashes followed by pedestrian related crashes with about 25%.

3. METHODS

This section explains the basic idea, pros and cons of three statistical methods used in this study. The methods are decision trees, random forests and logistic regression, respectively.

3.1. Decision Trees

Decision Trees is one of the most popular data mining techniques that is used to identify significant variables affecting the target variable from a set of explanatory variables.

The advantages of decision tress method include: (1) It is a nonparametric statistical method that require few statistical assumptions and hence robust, (2) It ameliorates the "curse of dimensionality" and can be applied to various data structures involving both ordered and categorical variables in a simple and natural way, (3) It does variable selection, complexity reduction, and (implicit) interaction handling in an automatic manner, (4) Invariant under all monotone transformations of individual ordered predictors, (5) The output gives easily understood and interpreted information, and (6) Special features are available in handling missing values and obtaining ranking of variables in terms of their importance.

On the other hand, it has two main disadvantages: (1) sometimes, it is not doing so well for estimation or prediction tasks, (2) instability of tree models. These two weaknesses can be utilized and improved by developing bagging, boosting, or random forests.

In this study, SAS Enterprise Miner, Version 9.2 was used to develop decision trees model in order to identify significant variables affecting the binary target variable (crash severity).

3.2. Random Forests

Random Forest (RF) is one of the most recent and promising machines learning techniques, proposed by Breiman (2000), which is well known for selecting important variables from a set of variables. Random Forest was used in this study for selecting significant explanatory variables affecting the binary dependent variable (crash severity: fatal or non-fatal crash). The advantage of using RF instead of other data mining techniques such as decision trees is that there is no need for a separate cross-validation-test data set to obtain

unbiased error estimates, especially when the sample size is small (Abdel-Aty et al., 2008). In addition, RF handles missing values in the covariates efficiently (Grimm et al., 2008).

Random Forest is a refinement of bagged trees. The term came from random decision forests that were first proposed by Ho (1998). It combines Breiman's "bagging" idea and Ho's "random subspace method" to establish a collection of decision trees with controlled variations (Breiman, 2001). The main idea of Random Forest is that every tree is built using a deterministic algorithm based on two factors. First, at each node, a best split is chosen from a random subset of the predictors rather than all of them to determine the splitting decision. Second, every tree is built using a bootstrap sample of the observations. The Random Forest procedure eliminates the need of dividing the data into training and validation sub-datasets because when a particular tree is grown from a bootstrap aggregate sample, one third of the cases are left out and not used in the development of the tree. These cases are called out-of-bag (OOB) data. This OOB data becomes the validation dataset which are used to obtain an unbiased error estimate as well as the estimates of variable importance. To test whether the attempted number of trees is sufficient enough to reach relatively stable results, the plot of OOB error rate against various tree numbers is developed. The best number of trees is that having the minimum error rate along with a constant error rate nearby (Breiman, 2001; Pang et al., 2006; Grimm, 2008; and Kuhn et al., 2008).

To select the important variables affecting the binary target variable, the R package provides the mean decrease Gini "IncNodePurity" diagram. By means of the Gini Index, the quality (Node Purity) of a split for every variable (node) of a tree is measured. Every time a split of a node is made on a variable m, the Gini impurity criterion for the two descendent nodes is less than the parent node. Then, adding up the Gini decreases for each individual variable over all trees in the forest provides a variable importance. A higher IncNodePurity implies a higher variable importance (Kuhn et al., 2008).

3.3. Logistic Regression

As indicated earlier, the main objective of this study was to identify and quantify the significant factors that might affect the target variable (crash severity). The response variable (crash severity) is a binary (dichotomous) variable namely "ACCISEV" with two levels: 0 if the crash resulted in at least one injury but no fatality, and 1 if the crash resulted in at least one fatality (within 30 days after the crash date).

Since both decision trees and random forests are non-parametric methods, logistic regression was used in this study to estimate the effect of the statistically significant factors on accident severity (i.e. whether it was a fatal or non-fatal crash).The description and levels of dependent and independent variables used in the logistic regression model are given in Table 1. According to Agresti (2002), logistic regression model can be written in the following form:

$$E(Y/X) = \pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \qquad (1)$$

Where the transformation of the π (x) logistic function is known as the logit transformation:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \quad (2)$$

The logistic regression model estimated in this study can be expressed as follows:

$$P\ (fatal\ crash) = \pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad (3)$$

Where, $g(x)$ is the function of independent variables:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (4)$$

4. RESULTS

4.1. Decision Trees Results

In this study, SAS Enterprise Miner, Version 9.2 was used to develop decision trees in order to identify the significant variables affecting the binary target variable (crash severity). Figure 1 shows a flow chart explaining the steps of estimating decision trees models developed in this study. For nominal or binary targets, a choice of three splitting criteria can be used to build decision trees models. These splitting criteria are: (1) Chi-Square test (default) - the Pearson Chi-Square measure of the target vs. the branch node, with a default significance level of 0.20, (2) Entropy Reduction - the reduction in the entropy measure of node impurity, and (3) Gini Reduction - the reduction in the Gini measure of node impurity. As shown in Figure 1, these three splitting criterion were used to develop three decision trees models. These three models showed very similar goodness of fit indices as well as significant variables and hence the results of the Gini reduction model is only presented and discussed here.

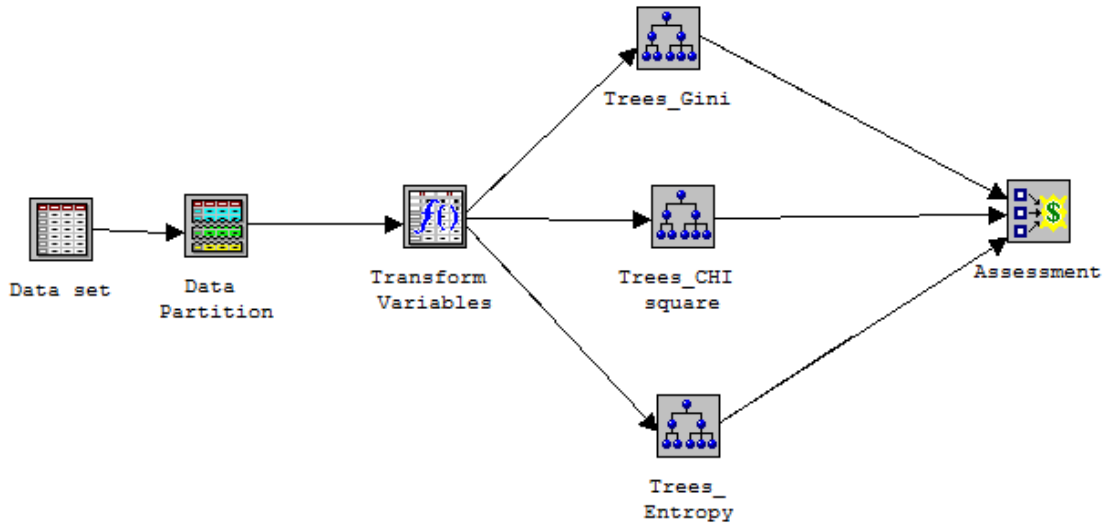

Figure 1: Flow chart showing steps of conducting decision trees models

Table 2: Goodness of fit measures of the three decision trees models

Tool	Name	Description	Target	Root ASE	Valid:Root ASE	Test:Root ASE	Misclassification Rate	Valid:Misclassification Rate	Test:Misclassification Rate
Tree	Tree	Gini	ACCISEV	0.4112495176	0.4264807636	0.424021311	0.2195448461	0.2431950022	0.239178938
Tree	Tree	Entropy	ACCISEV	0.4112495176	0.4264807636	0.424021311	0.2195448461	0.2431950022	0.239178938
Tree	Tree	chi	ACCISEV	0.4175161901	0.4302585992	0.4268211255	0.2248995984	0.2445336903	0.239178938

Table 2 shows the goodness of fit indices of the three developed decision trees models. In addition, Table 3 shows the ranking of the significant variables affecting the binary target variable (crash severity). As shown in Table 3, the most significant variables affecting crash severity were reason of crash, point of collision, crash type, damages in private and public properties, and day of the week. Moreover, Figure 2 depicts a flow chart of tress split based on Gini reduction decision trees model.

Table 3: Variable importance from the Gini decision Trees model

Name	Importance
ACC_REASON_CH	1.0000
COLLISION_POINT_CH	0.9118
ACC_TYPE_CH	0.8486
PUBLIC_DAMAGES	0.7596
PRIVATE_DAMAGES	0.6734
WEEK_DAY_CH	0.5905
ACC_LOCATION	0.4463
DAYNIGHT	0.2883
WEATHER	0.2641
LIGHTING_STATUS	0.2230
ROAD_SURFACE_COND	0.0000

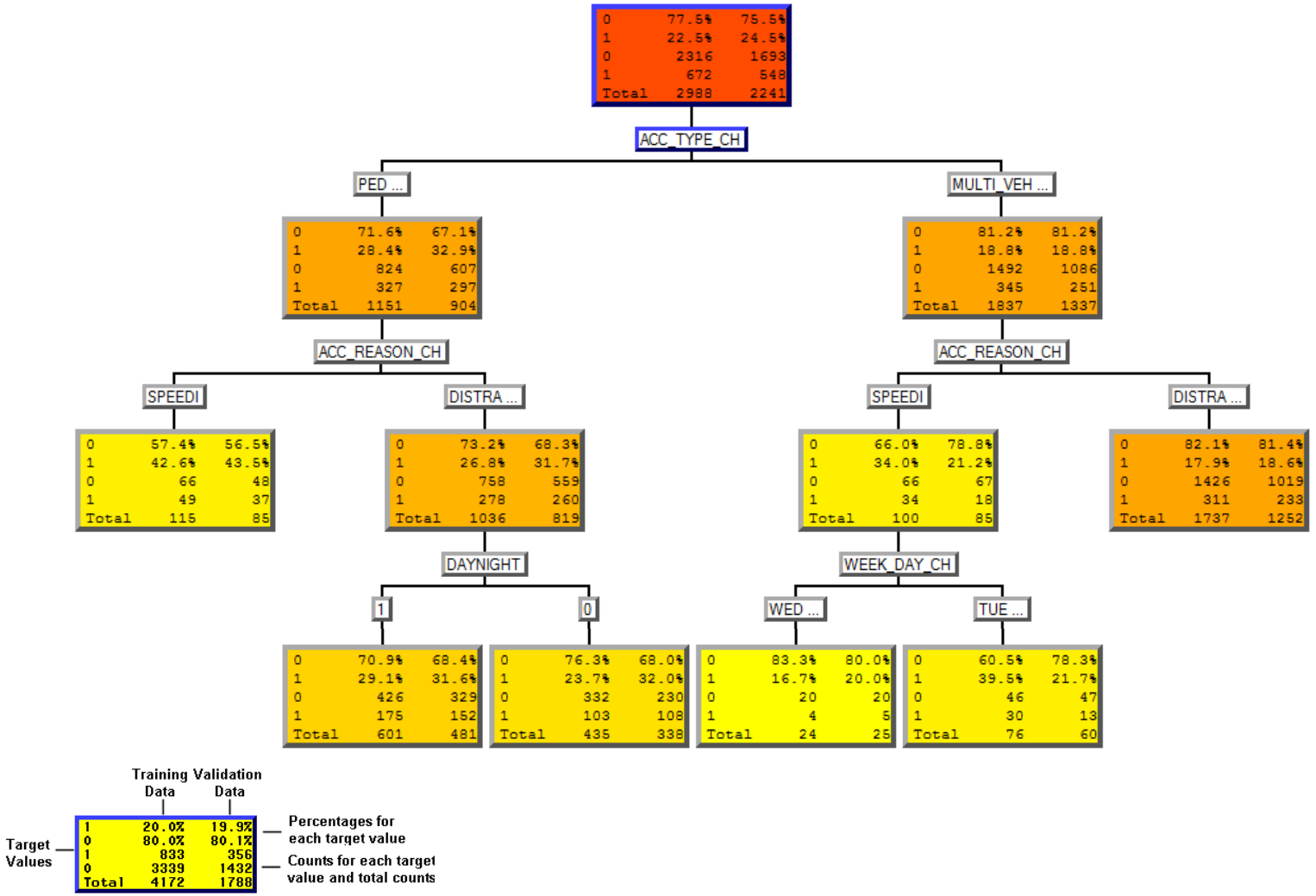

Figure 2: Trees split flow chart (Gini Trees)

4.2. Random Forests Results

In this study, the Random Forest technique was conducted using the R package (the reader is referred to Liaw and Wiener, 2002). Figure 3 shows the plot of OOB error rate against various tree numbers. Clearly, 100 trees are sufficient enough to reach relatively stable results. Additionally, the purity values for every covariate are shown in Figure 4.

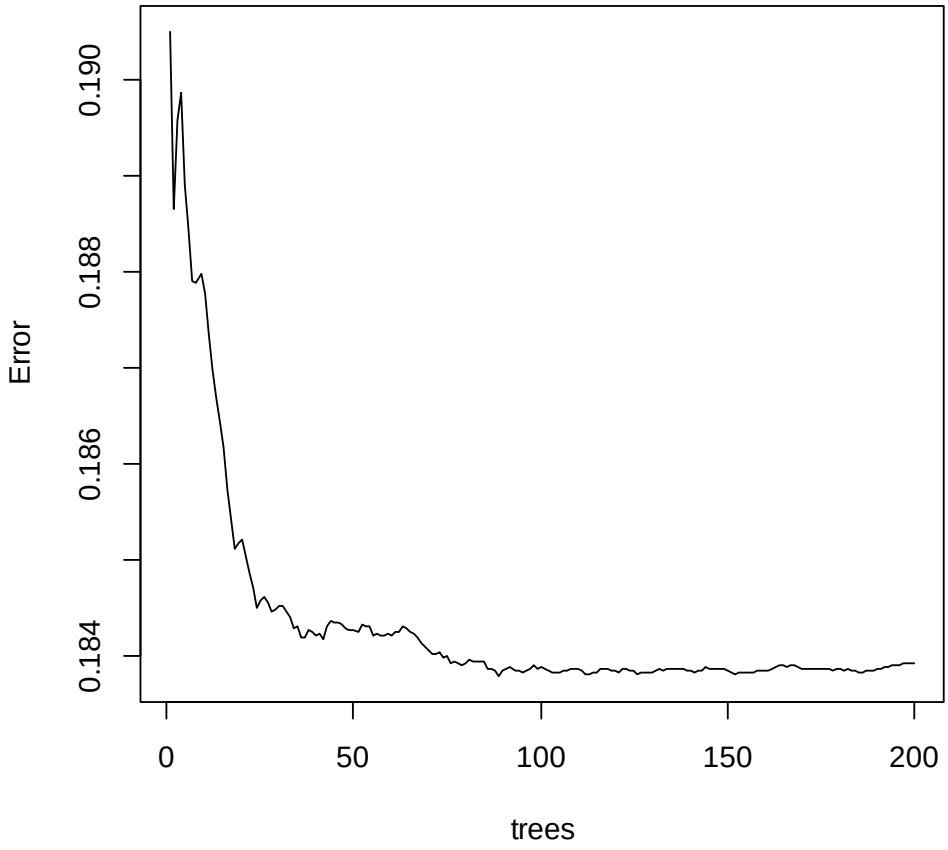

Figure 3: Plot of the OOB error rate against different number of trees

As shown in Figure 4, in order to choose the most important covariates affecting the binary target variable (fatal versus non-fatal crash), a cut-off purity value of "80" was chosen. This led to selecting four important covariates. These four variables have higher variable importance scores than the remaining variables. These variables are day of the week (week_day_ch), point of collision (collision_point_ch), reason of accident (ACC_Reason_ch) and crash type (Acc_Type_ch).

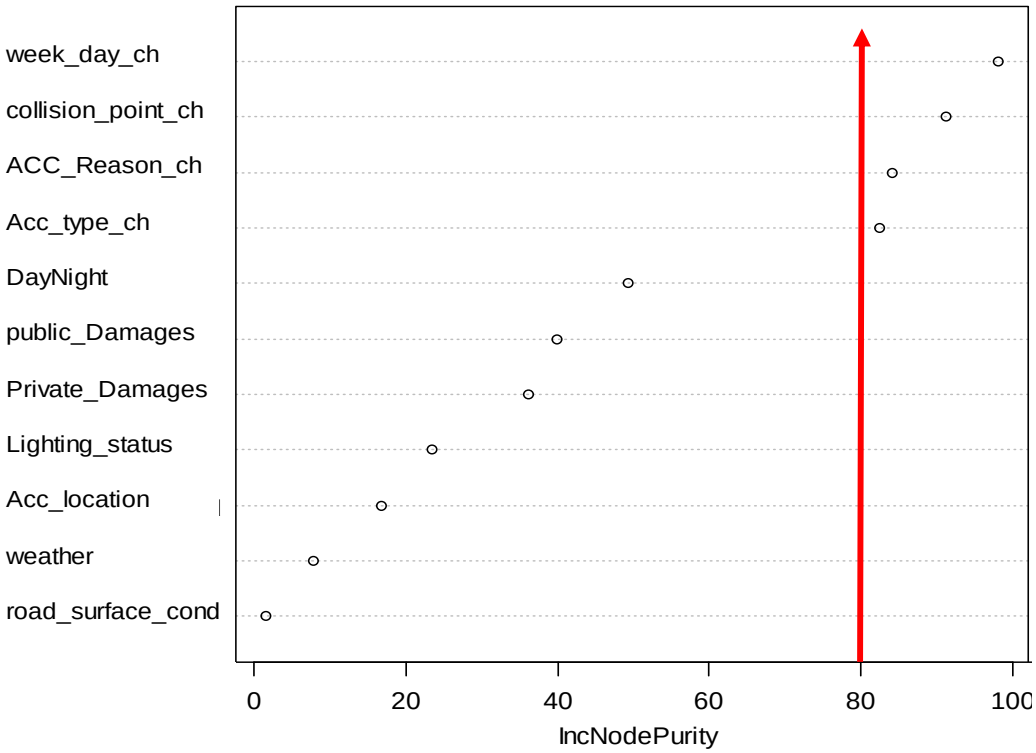

Figure 4: Variable importance ranking using node purity measure

4.3. Logistic Regression Results

Using SAS package, the logistic regression model can be built by using SAS procedure Logistic or Genmod. However, we developed the model in SAS Enterprise Miner since it's quite convenient to compare the results with Decision Tree in SAS Enterprise Miner framework.

Figure 5 shows the T-scores of the significant variables affecting the dependent variable (crash severity). The T-scores are ranked in decreasing order of their absolute values, which indicates that the higher the absolute value is, the more important the variable is. As shown in Figure 5, reason of crash (particularly speeding and distraction), type of crash (particularly pedestrian related crashes) were the significant variables affecting crash severity (fatal vs. non-fatal crashes). In addition, Table 4 shows the Logistic Regression Estimates.

Moreover, Table 5 presents the goodness of fit indices for the decision trees as well as logistic regression models.

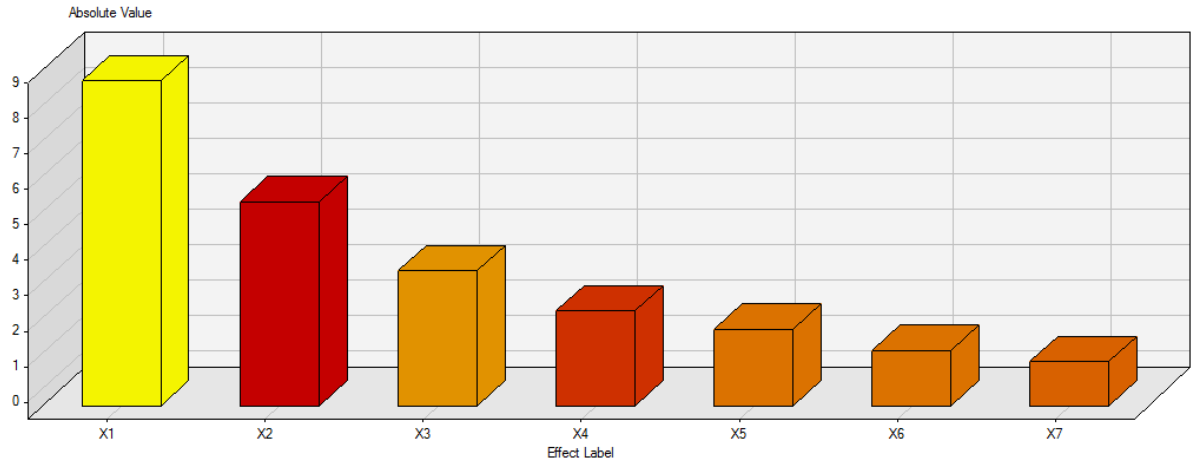

Figure 5: Variable importance ranking using node purity measure
(X1= Intercept: ACCISEV=1, X2=ACC_Reason_ch_speeding, X3= ACC_Reason_ch_distraction, X4= Acc_Type_ch_Ped, X5= Acc_Type_ch_others, X6= ACC_Reason_ch_others, X7= Acc_Type_ch_Multi vehicles)

Table 4: Logistic Regression Estimates

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.2312	0.1416	75.63	<.0001
ACC_Reason_ch	distraction	1	0.3645	0.1092	11.15	0.0008*
ACC_Reason_ch	others	1	0.0964	0.0876	1.21	0.2709
ACC_Reason_ch	Speeding	1	0.6076	0.1152	27.82	<.0001*
Acc_type_ch	Multi-vehicles	1	0.1135	0.1461	0.60	0.4372
Acc_type_ch	Ped	1	0.3344	0.1516	4.87	0.0274*
Acc_type_ch	others	1	0.6885	0.4050	2.89	0.0891**

* Significant at 0.05 level, ** significant at 0.10 level

Table 5: Goodness of fit indices of decision trees and logistic regression models

Goodness of fit indices	Model	
	Decision Trees	Logistic Regression
Root ASE	0.411	0.423
Valid: Root ASE	0.426	0.435
Test: Root ASE	0.424	0.433
Misclassification Rate	0.219	0.235
Valid: Misclassification Rate	0.243	0.264
Test: Misclassification Rate	0.239	0.245

5. DISCUSSION

As discussed earlier, two different data mining techniques (decision trees and random forests) were developed in this study to select significant variables affecting crash severity. Three significant variables were obtained using both methods. These three variables were reason of crash, point of collision, and crash type as shown in Table 3 and Figure 4.

Since both decision trees and random forests are non-parametric methods, logistic regression model was also estimated to quantify the effects of significant variables affecting the binary dependent variable (crash severity). The results showed that reason of crash and crash types were the significant variables. As shown in Table 5, the decision trees models showed a relatively better goodness of fit indices than logistic regression model.

It is worth mentioning that both models showed relatively high misclassification rates (about 23%) possibly due to the absence of important explanatory variables in the Saudi crash reports such as drivers' demographic and ADT. Certainly, the research team has made substantial effort to obtain the most complete available data.

6. CONCLUSIONS AND RECOMMENDATIONS

This study aimed to identify and quantify the significant variables affecting the severity of traffic crashes in Riyadh, the capital of Saudi Arabia. Thus, a total 7470 reported fatal and serious traffic crashes were used in the analysis. The target variable (crash severity) is of a binary nature (i.e., has two levels - fatal or non-fatal) and hence, two data mining techniques (decision trees and random forests) as well as logistic regression model were developed to achieve that goal. Crash reason (particularly speeding and distraction), type of crash (particularly pedestrian related crashes), point of collision and day of the week were the significant variables associated with crash severity.

Considering the results of this study, traffic safety agencies in Saudi Arabia are advised to focus their efforts at non-intersection locations rather than intersections (e.g., about 98% of severe crashes in Riyadh occurred at non-intersection locations). In addition, it is recommended that future education courses and campaigns should emphasize the negative effects of committing speeding, distraction and sudden lane change while driving and the strong association between these aberrant driving behaviours and increasing the crash severity.

Moreover, the findings of this study revealed that more than one quarter (26%) of the Saudi severe crashes were pedestrians-related crashes. Since the majority of the Saudi severe crashes occurred at non-intersection location, these results might shed light on the necessity for improving safety of pedestrian crossing at non-intersection locations (especially at locations where these crashes occur frequently) by providing pedestrians bridges/tunnels and/or improving marking and warning signs.

It is worth noting that the major contribution of this study was obtaining a complete data and performing these statistical methods on the Saudi crash dataset for the first time. Certainly, availability and completeness of crash data is seldom in developing countries.

Finally, the main limitation of this study is that the results described in this paper were estimated with no consideration for traffic exposure or the data that are not available or difficult to obtain in Saudi Arabia. However, the findings of this study can be considered as guidance for a future study when such data become available.

ACKNOWLEDGEMENTS: The authors would like to thank the Municipality of Riyadh for providing the Saudi crash dataset used in this study. All opinions and results are those of the authors.

7. REFERENCES

Abdel-Aty, M., Pande, A., Das, A., and Knibbe, W. (2008). Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. Transportation Research Record, 2083, pp. 153-161.

Agresti, A. (2002). Categorical Data Analysis, second ed. John Wiley & Sons, Inc., Hoboken, New Jersey.

Al-Ghamdi, A.S. (2002a). Pedestrian–vehicle crashes and analytical techniques for stratified contingency tables. Accident Analysis and Prevention, 34, pp. 205–214.

Al-Ghamdi, A.S. (2002b). Emergency medical service rescue times in Riyadh. Accident Analysis and Prevention, 34, pp. 499–505.

Al-Ghamdi, A.S., AlGadhi, S. A. (2004). Warning signs as countermeasures to camel–vehicle collisions in Saudi Arabia. Accident Analysis and Prevention, 36, pp. 749–760.

Al-Ghamdi, A.S. (2007). Experimental evaluation of fog warning system. Accident Analysis and Prevention, 39, 1065–1072.

Bendak, S. (2005). Seat belt utilization in Saudi Arabia and its impact on road accident injuries. Accident Analysis and Prevention, 37, 367–371.

Breiman, L. (2000). Some infinity theory for predictor ensembles. Technical Report 579, Statistics Dept. University of California Barkley.

Breiman, L. (2001). Random forests. Machines Learning, 45, 5–32.

Grimm, R., Behrens, T., Marker, M., and Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island. Digital soil mapping using Random Forests analysis. Geoderma, 146(1-2), pp. 102-113.

Ho, T. (1998). The random subspace method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(8), pp. 832–844.

Koushki, P.A., Al-Ghadeer, A.M. (1992). Driver non-compliance with traffic regulations in rapidly developing urban areas of Saudi Arabia. Transportation Research Record, 1375, pp. 1–7.

Kuhn, S., Egert, B., Neumann, S., and Steinbeck, C. (2008). Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. BMC Bioinformatics, 9 (400).

Liaw, A. and Wiener, M. (2002). Classification and Regression by random Forest. R News 2(3), pp. 18-22.

Pang, H. et al. (2006). Pathway analysis using random forests classification and regression. Systems biology, 22(16), pp. 2028–2036.