INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Quality Engineering and Technology

ISSN online: 1757-2185 - ISSN print: 1757-2177
https://www.inderscience.com/ijqet

Statistical analysis of factors associated with recent traffic accidents dataset: a practical study

Emad Imreizeeq, Jamal N. Al-Karaki, Amjad Gawanmeh

DOI: 10.1504/IJQET.2023.10054068

Article History:
Received: 08 April 2021
Accepted: 22 October 2021
Published online: 21 February 2023

Statistical analysis of factors associated with recent traffic accidents dataset: a practical study

Emad Imreizeeq

Academic Support Department,
Abu Dhabi Polytechnic,
Institute of Applied Technology,
P.O. Box 111499, Abu Dhabi, UAE
Email: Emad.Imreizeeq@adpoly.ac.ae

Jamal N. Al-Karaki

Zayed University,
P.O. Box 144534, Abu Dhabi, UAE
and
The Hashemite University,
P.O. Box 330127, Zarqa 13133, Jordan
Email: jamal.al-karaki@zu.ac.ae

Amjad Gawanmeh*

College of Engineering and IT,
University of Dubai,
Dubai, UAE
Email: amjad.gawanmeh@ieee.org
*Corresponding author

Abstract: In this paper, we propose a logistic model to fit accidents dataset of 10,000 road crash incidents for the Emirate of Abu Dhabi published in 2020. After cleaning up the dataset, we use descriptive and inferential statistical tools to study the attributes of each variable. Then, we identify the main independent variables that can be incorporated in a general logistic regression model which also includes the interactions between them. Our analysis using the significance level of (alpha = 0.05) found that there is a reduced logistic regression model that can fit the data in which the 'location of accident' can be represented using 'type of accident' and the 'age' of people involved in the accidents. Moreover, the results show that the interaction terms are not significant to be included in the model. Furthermore, the study shows that the odds for accidents by young age group (less than 40 years old) in external streets is 27% higher than the odds for internal streets, and that the odds for sequential type accidents in external streets is 13% higher than the odds for internal streets.

Keywords: logistic regression; traffic accidents; accident data analysis; chi-square test.

Reference to this paper should be made as follows: Imreizeeq, E., Al-Karaki, J.N. and Gawanmeh, A. (2023) 'Statistical analysis of factors associated with recent traffic accidents dataset: a practical study', *Int. J. Quality Engineering and Technology*, Vol. 9, No. 1, pp.1–19.

Biographical notes: Emad Imreizeeq is an Assistant Professor of Applied Mathematics. He worked at the University of Twente in The Netherlands and also as consultant with many companies in the Netherlands and abroad. Since 2014, he is working at the Academic Support Department at Abu Dhabi Polytechnic, UAE. He worked in developing new applied math courses and on teaching different topics in mathematics, in particular, calculus, probability and statistics and differential equations. His preferred fields of research are dynamical systems, statistical data analysis applied to finance, climate change and environmental science.

Jamal N. Al-Karaki is a Full Professor at the Zayed university, UAE and also at The Hashemite University of Jordan. He had a rich university career in education, service, and research. He holds a PhD (with distinction) in Electrical/Computer Engineering from the Iowa State University, USA, with Research Excellence Award. He has a proven record of progressive responsibility and authority including leadership positions as the College Dean at various institutes. He published 80+ refereed technical articles. He was listed among the top 2% highly cited researchers in his field worldwide. He is a senior member of IEEE and ACM.

Amjad Gawanmeh is an Associate Professor at the University of Dubai, UAE and Affiliate Adjunct Professor at the Concordia University, Montreal, Canada. He received his Bachelor's in Electrical and Computer Engineering for JUST, Jordan, 1998, and his MS and PhD from the Concordia University, Montreal, Canada, 2003 and 2008; respectively. He has two edited books, three book chapters, more than 40 peer reviewed Scopus indexed journal papers, and more than 70 peer reviewed conference papers. He is the EIC for *IJCPS*, AE for *IEEE Access* and *HCIS Journal*, and a guest editor for several SIs. He is senior IEEE member.

1 Introduction

The World Health Organization (2018) Global Status Report on Road Safety shows that annual road traffic accidents have reached 1.35 million people worldwide each year, leaving up to 50 million people with non-fatal injuries. In addition, these accidents have a significant impact on national economies, costing countries 3% of their gross domestic product annually. Road traffic injury (RTI) creates major human and economic pressure on countries around the world. In 2018, more than 1.5 million road users died as well as an additional 50 million, mainly among vulnerable road users, were injured or disabled for life, either pedestrians or motorcyclists. To minimise the associated morbidity and mortality of RTI, it is very important to understand the likelihood of traffic accidents by examining the main contributing factors and the relationship between these factors. Both intrinsic and extrinsic factors associated with the increased risk of fatal road accidents were explored in existing literature (Ghandour et al., 2020).

The work in Rolison et al. (2018) assessed the impact of some factors on road accidents such as lack of skills or experience and risk-taking behaviours of young drivers. Altwaijri et al. (2011) explored factors affecting the severity of road injury crashes in Riyadh City, Kingdom of Saudi Arabia. The study shows that there is no effect of age in slight injury crashes relative to serious injury crashes. The use of new methodologies and mathematical models to identify and assess essential factors that affect the severity or fatal of injuries have been extensively researched (Ghandour et al., 2020). An effective decision tree (DT) is presented in Abellán et al. (2013) as a potential method for studying the severity of traffic accidents.

Ghandour et al. (2020) used 8,482 road crash incidents from the Lebanese Road Accidents Platform (LRAP) database to define a machine learning-based intelligent to identify risk factors contributing to fatal road injuries. Moreover, other statistical approaches using regression analysis were adapted to analyse the associated variables with road accidents. Zajac and Ivan (2003) used ordered probit model to assess the effects of roadway and area type features on injury severity of pedestrian accidents in rural Connecticut. Cantillo et al. (2020) used exploratory analysis based on multinomial ordered discrete model to identify the main factors related with the severity levels of accidents. Wali et al. (2020) examined the relationship between driving volatility and the severity of accident-injuries and found a statistically significant positive correlation between them. Tarko and Azam (2011) proposed a bivariate ordered probit model to identify the pedestrian injury severity factors. Chen and Zhou (2016) used Bayesian hierarchical intrinsic conditional autoregressive model to test for the relationship between motorists and pedestrians using data for the city of Seattle. Li et al. (2021) identified on factors related to pedestrian-vehicle accidents in different time periods, using data from 2007 to 2018 in North Carolina. Theofilatos (2017) utilised real-time traffic and weather data collected from urban arterial in Athens, Greece. The findings can be used in designing effective traffic management strategies to reduce the severity of accidents. Furthermore, the authors presented a useful technique to measure the effectiveness of their proposal. Several other recent results used different multivariable method to analysis road accidents data, for instance, Gilardi et al. (2020), Abdella et al. (2019), Nazeri et al. (2021), Paul et al. (2021) and Xie (2020).

Bédard et al. (2002) used data for single-vehicle accidents with fixed objects to apply a multivariate logistic regression model. Valent et al. (2002) used logistic regression model to characterise the main factors for fatal injuries. The result shows that driver's injury was strongly related with lack of use of seat. Yau (2004) investigated factors affecting the occurrence of vehicle traffic accidents in Hong-Kong using stepwise logistic regression model. Similarly, Al-Ghamdi (2002) applied a logistic regression model with dichotomous dependent variable with two categories, fatal and non-fatal accident. Tay et al. (2011) estimated a multinomial logit model to identify the factors determining the severity of pedestrian-vehicle crashes in South Korea. Xie et al. (2009) compared between a Bayesian ordered probit (BOP) model and ordered probit (OP) model in the analysis for severity of injures. The work of Abdel-Aty (2003) and Abdel-Aty et al. (1998) used ordered probit model for factors affecting injury severity. Similar statistical modelling techniques that used nested logit model appeared in Chang and Mannering (1999), Nassar et al. (1994) and Shankar et al. (1996). Furthermore, from the recent studies using data mining approach to predict and find important factors, causing traffic accidents include Hazaa et al. (2019), El Tayeb et al. (2015), Kumar and Toshniwal (2016), and finally Mulay and Mulatu (2016).

Other interesting approach was presented in Umer et al. (2020) where a tree-based statistical models based on logistic regression stochastic gradient descent were used as voting classifiers in order to predict road accident severity. Finally, the work in Zhang et al. (2019) created a model using logistic regression in conjunction with Recursive Feature Elimination to determine the weights of different factors that caused recent traffic crashes in Massachusetts.

Overall, adopting an appropriate approach for data analysis in this area can significantly improve our understanding for the parameters being studied. Moreover, it also helps in enhancing accidents prediction accuracy. In addition, this can also help in finding the best paradigm for identifying factors affecting road accidents. For this reason it is important to understand how different factors can influence traffic accidents severity. This paper will try to identify a logistic model that takes into consideration not just different independent variables, but also the interaction between them. For this purpose, we analyse the association between the reason of the accident and the corresponding street where the accident happens (inside/outside) in the city of Abu Dhabi. Also, we checked the association between the age of the person who did the accident with the reason of the accident.

The key contributions of this paper is summarised as follows. First, the paper builds a logistic regression model (logit) based on a new real dataset published in 2020 for the emirates of Abu Dhabi, to infer meaningful patterns of factors causing car accidents. Second, the paper used only categorical variables to reach to the logit model due to lack of data values in the dataset (e.g., number of injuries in the accidents). Third, in the process of reaching to the model, we used descriptive statistics to reduce the dimensionality of the problem from 11 variables to four variables, by removing any redundant variables. Fourth, we go a step further by using inferential statistics to check for significance in the association between the four considered variables. This step gives the path to build a logit model considering the inputs (predictors), and the interactions between them. To the best of our knowledge, the literature related to traffic accident data did not consider the interactions between the inputs (independent) variables when building the logit model. Our work aims to fill this gap, and test and check if these interactions are statistically significant or not, when building the logit model. Final contribution is to reach to more consistent, reliable and above all simple model that can be shared with respective authorities and policy makers for better planning and decision making.

The rest of this paper is organised as follows. In Section 2, the problem statement and methodology of a solution are described. In Section 3, we cleaned the data and use descriptive statistics to remove the redundant variables and reduced the number of variables from 11 to 4 variables namely, the street, type of collision, reason of accident, and age of the drivers. In Section 4, we used inferential statistical tools to test for the significance of associations between the four variables. For that we used the chi square test of association six times to cover all possible combinations between the relevant 4 variables. In Section 5, a logit model is found by considering ‘street’ variable as the response variable, and the other three variables together with their interactions to represent the predictors. Section 6 presents the discussion on the impact of the results together with the limitations, and future research directions. Finally, Section 7 gives the conclusions and recommendations for policy makers.

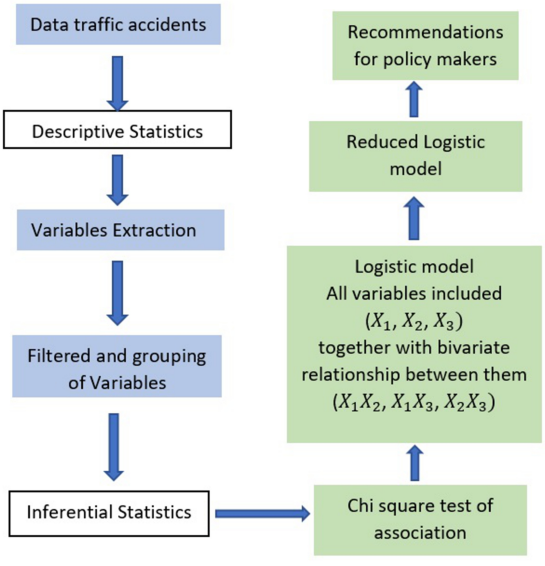
2 Materials and solution methodology

In this section, the research problem is defined, and the solution methodologies are also outlined.

2.1 Problem formulation

The characterisation of the factors related to traffic-accident data includes both continuous and discrete type of variables. Hence, analysing it, is harder than the standard linear/nonlinear regression models employed for only continuous data. Logistic type regression (logit) is so far the most logical path to handle such issue. However, logit model will lose its advantage because of another two main problems. The first is related to the high dimensionality of the model as there are too many involved variables. The second is related to the possible interaction effect between the independent variables, used as predictors in the logit model. This interaction is normally neglected in the literature on the assumption that interactions between the independent variables have no effect on the dependent variable. Hence, to overcome these issues when using a logit model, the above two issues should be resolved.

Figure 1 Proposed methodology (see online version for colours)



2.2 Solution methodology

Our objective is to find a model that can capture the interconnections between the variables without sacrificing the accuracy of the results. To achieve this objective, we perform the following steps. Firstly, we employed tools from descriptive statistics to clean the data, and get a general idea about the structure of the relevant variables. This step is useful in excluding any redundant variables and to reduce the number of attributes of each variable to get a categorical variable with a maximum of

2–4 attributes, which can be easily understood. In this step, we can reduce the dimensionality from 11 variables to four variables, namely street, type, reason and age. Secondly, we use inferential statistics to check the feasibility and the significance of association between the left four categorical variables by using chi-square test for association. Hence, in this step, the test is done six times to cover all possible combinations of interactions between these variables. Thirdly, and from the results of the second step, we search for the best logistic regression model after considering the variable 'street' as our response variable. In this search, we started with the three variables: type, reason and age and also the interactions between these variables as predictors in the model. The model then reduced further by removing all non-significant variables which do not contribute to the model (Agresti, 2006). From the simplified final model, we can draw conclusions and recommendations for policy maker to be incorporated when getting or renewing a driving license. Figure 1 shows a schematic diagram that represents the components of the solution methodology.

3 The dataset modelling and descriptive statistics

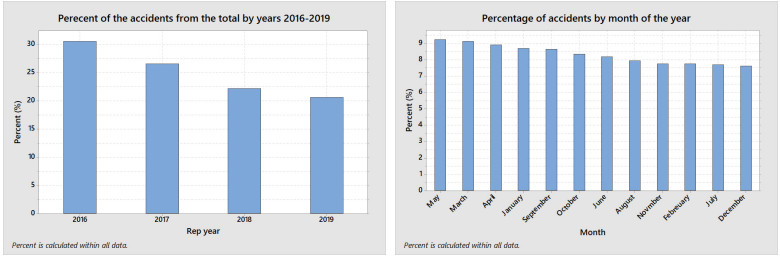
In this section, we describe the dataset used and how it is characterised and cleaned for proper analysis using descriptive statistical tools. Although the dataset contains many details, its critical challenge is the lack of accurate continuous variable representing the number of injuries. Hence we have to deal with categorical data only in this paper.

The dataset used in this paper includes a new dataset of traffic accidents which recently appears publicly for research in the Emirate of Abu Dhabi in UAE over the period 2016–2019 and contains around 10,000 traffic accidents entries (https://www.adda.gov.ae/) where every traffic accident entry is listed with the following attributes: number of injuries, type of collision, ages of drivers, weather condition, etc. A previous analysis for the Emirate of Abu Dhabi between for the period of 2012 to 2017 was analysed in Albuquerque and Awadalla (2020) where the authors used a multivariate logistic regression model to analyse the factors that cause fatal road crash severity. However, the new dataset that we are using contains the weather, visibility, traffic volume, speed and occupancy at a one minute resolution. Table 1 shows a sample of the data. The only problem of the data is the lack of accurate number of injuries represented by column 11. The table shows a sample of this column where some data for injuries is shown. However, most of the data in column 11 is missing and filled with zeros. Hence, we did not include this column in our analysis. Table 2 gives more details about the type of each variable and corresponding category related to it in this dataset.

3.1 Data augmentation: data filtering and cleansing

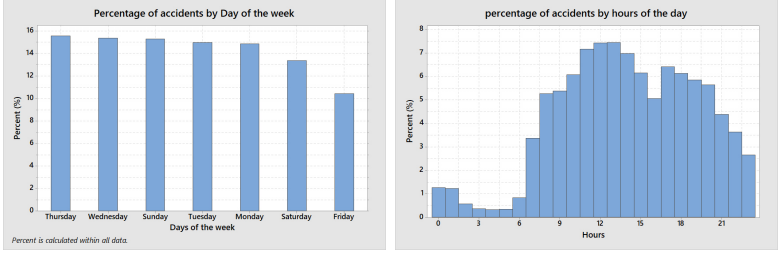
To clean up the data and take into consideration the most important attribute of each variable, we proceed as follows: First, we look at columns 1 and 2, which represent the date and time of the accidents, respectively. The summary of these data is shown in Figures 2(a) and 2(b). Figure 2(a) clearly shows that from the total number of accidents, the percentage number of accidents gradually decreased from 30% in 2016 to 20% in 2019. Figure 2(b), shows that the maximum number of accidents occurs in the spring starting with May, then March and April. The minimum percentage occurs in December and July.

Figure 2 Histogram of accidents with years and months in 2016–2019 (see online version for colours)



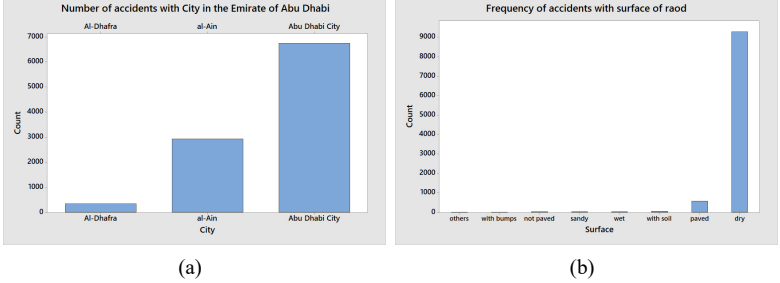
(a) (b)

Figure 3 Histogram of accidents by, (a) days from 2016–2019 (b) through 24 hours from 2016–2019 (see online version for colours)

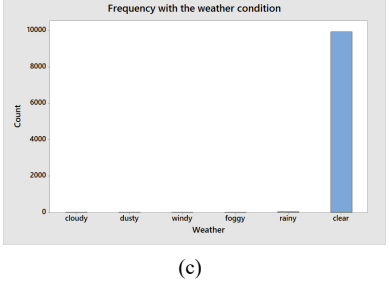


(a) (b)

Figure 4 Histogram of accidents by, (a) city (b) surface (c) weather (see online version for colours)



(a) (b)



(c)

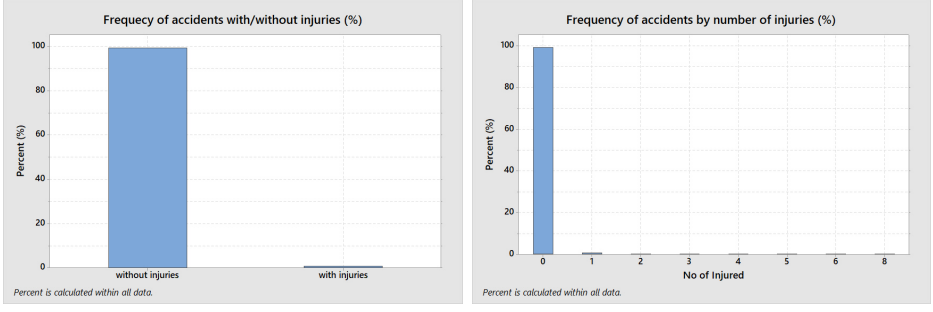
Table 1 Sample of the data in the used dataset

Col. 1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Date	*Time*	*Rep kind*	*City*	*Street*	*Rep. type*	*Reason*	*Surface*	*Weather*	*Age*	*No. of injured*
21/9/17	19:48	Accident with injury	Abu Dhabi	Tarif Road	Side-impact collision	No-lane discipline	dry	clear	41	8
16/1/16	7:30	Accident with injury	Al Ain	Abu Dhabi/Al Ain HW	Rear-end collision	Blurred vision	dry	foggy	54	6
4/7/16	7:00	Accident with injury	AlDhafra	Alsela'/Abu-Dhabi HW	Side-impact collision	No-lane discipline	dry	clear	43	5
8/7/17	5:30	Accident with injury	Abu Dhabi	Khalid Ibn Alwaleed/Zayed the first int.	Head-on collision	Jumping red lights	dry	clear	52	5
1/1/16	15:15	Accident with injury	Abu Dhabi	Abu-Dhabi Al-ain road	Rollover collision	Sudden swerving	dry	clear	54	4
17/11/17	7:00	Accident with injury	AlDhafra	(Alsela'-Abu-Dhabi) public road	Rollover collision	No-Lane discipline	dry	clear	36	4

Table 2 The variables in dataset and its attributes

Col. number	*Variable*	*Type and categories*
Column 1	Reported date	In days (date variable)
Column 2	Reported time	In minutes (time of day variable)
Column 3	Reported kind	with injury/without injury (binary)
Column 4	City	3 categories (Abu Dhabi, Al-Ain, Dhafra)
Column 5	Street	More than 200 different categories (highways and internal streets)
Column 6	Type	9 types of different categories of collisions (front, back, side, etc..)
Column 7	Reason	23 different categories (No lane, red light, …)
Column 8	Surface	5 categories (dry, paved, filled with sand, wet, dusty)
Column 9	Weather	3 categories (dry, foggy, rainy)
Column 10	Age	Ages of drivers (22–91)
Column 11	Number of injuries	In integer values (not reliable data, as most of the data represented by zeros)

Figure 5 Histogram of accidents by, (a) injuries/no injuries (b) number of injuries (see online version for colours)



(a) (b)

Figure 3(a) shows that the percentage of accidents by day of the week is uniform in the working days, which is around 15% each day. However, the percentage of accidents in the weekend is smaller especially on Friday with 10% and then Saturday with 13%. These results are logical as the number of vehicles on streets is relatively less on the weekend days in comparison with the working days. Figure 3(b) represents the frequencies of data within 24 hours for the whole period. As expected, most of the accidents occur between 9 AM to 6 PM.

The data in column 4 represents the cities in the Emirate of Abu Dhabi. This categorical variable has the following attributes: Abu Dhabi city, Al-Ain city, and Al-Dhafra region. The frequencies of the accidents in each are shown in Figure 4(a). The histogram shows that most of the accidents are in the city of Abu Dhabi and its surroundings. Figure 4(b) shows the frequency of accidents caused by surface situation related to column 8, which clearly shows that most of the accidents are attributed to dry surface. The frequency of accidents with the type of weather in column 9 is shown in Figure 4(c). The histogram shows that most of the accidents are attributed to clear weather condition.

The data in column 3 is related to the kind of accident (with/without injury). In addition, the data in column 11 is related to the number of injuries. This obviously shows that the data is not accurately recorded and hence requires cleansing and filtering in order to be processed properly. More than 99% of the accidents are recorded with no injuries, and it shows that the total number of injuries is only 115, which is not realistic. So, we did not use the data of these two variables in our analysis. See Figures 5(a) and 5(b).

3.2 Model development – choosing and encoding the relevant variables

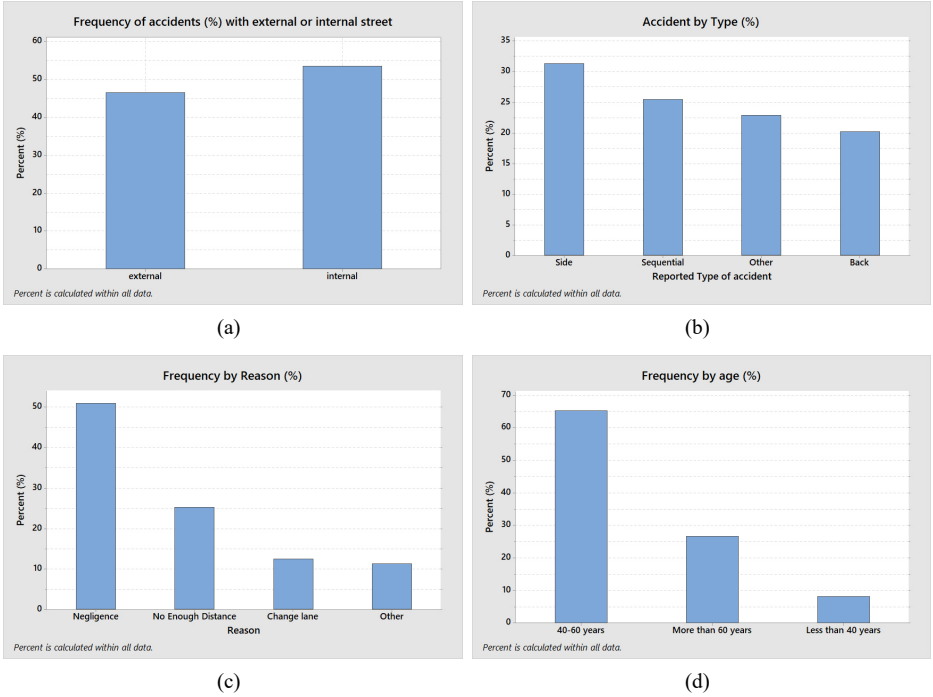
From these descriptive observations, we can reduce the dimensionality problem of the variables by fixing our attention and analysis to the accidents in the city of Abu Dhabi, on a dry surface, with clear weather. In other words, we are fixing the variables: city (column 4), surface (column 8), and weather (column 9). Moreover, the variable related to the reported kind of accident (column 3), and the variable related to the number of injuries (column 11), are both neglected as this data is not reliable as we have discussed in the previous subsection. Furthermore, as we do not consider any dynamics in the variables. The date (column 1) and time (column 2) are not included in the analysis part of this paper.

Table 3 Dependent and independent variables in the logistic model

Column no.	*Variable*	*Type and category*	*Combined attributes*
5	Street	More than 200 categories which include highways, internal streets, intersections	Internal, external dichotomous variable
6	Type	Nine types categories of collisions	Side, sequential, back, other categorical variables
7	Reason	23 categories	Negligence, no enough distance, change lane, other
10	Age	Drivers ages: 22–91	<40 years, 40–60, more than 60 years

Hence, we focused on the analysis of the interactions and connections between the other variables. Namely, columns 5, 6, 7, and 10, which are: street, type, reason, and age respectively. Starting with column 5, related to the street, the data registered more than 83 different streets/locations where the accidents occur. To analyse this variable, we combined all the attributes into a dichotomous variable (internal = 0, and external = 1). Similarly, we combined column 6 which has nine types of collisions to only four types of collisions (side, sequential, back, and other). For column 7, related to the reason of accident, 23 attributes are combined into four attributes (negligence, no enough distance, change lane, and other). Finally, Age variable in column 10 which is a continuous variable has been combined into three attributes (Less than 40 years, 40–60 years, and more than 60 years). Table 3 summarises the original and combined categories of each of the four variables. Figure 6 shows visually the attributes of the combined variables.

Figure 6 Histogram of accidents by, (a) street (b) type of accidents (c) reason (d) age (see online version for colours)



4 Pattern analysis of traffic accident using inferential statistics

4.1 Objective

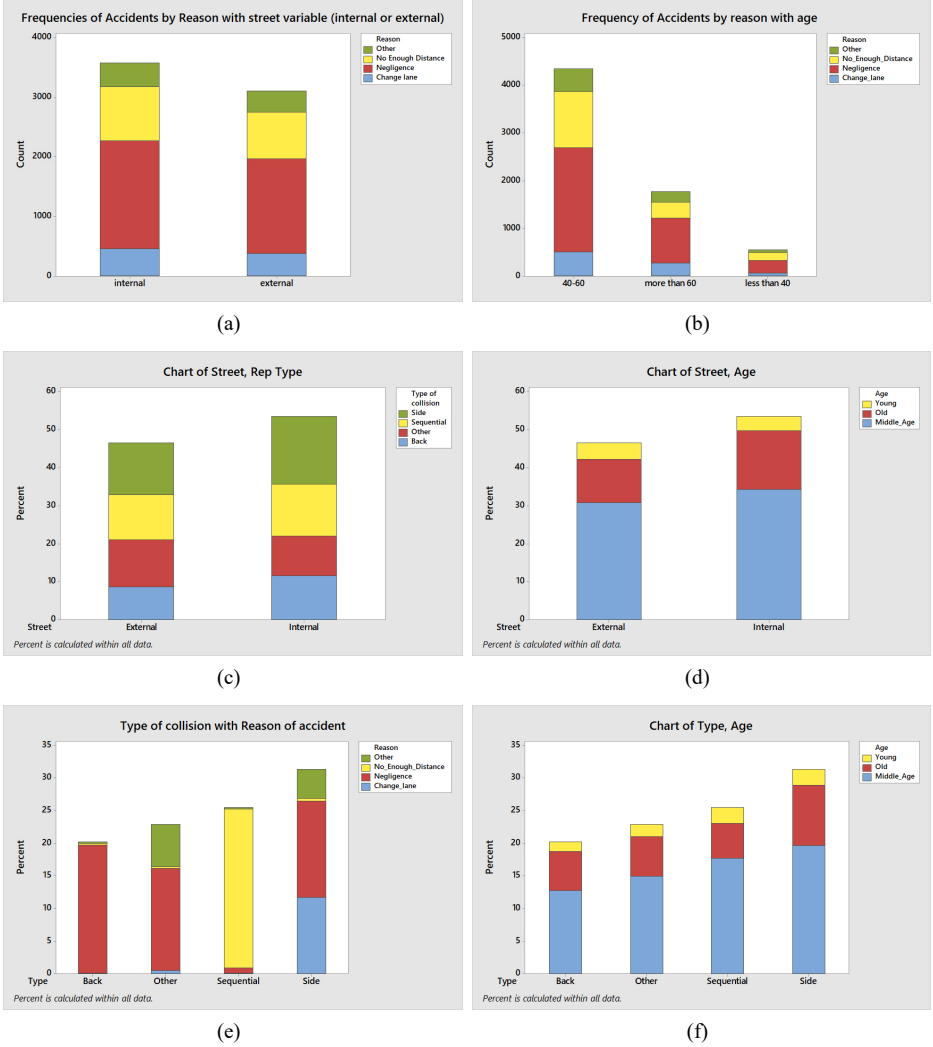
After cleaning up the initial data and encoding the relevant variables, our next step is to check if there are any associations or connections between the variables. A suitable measure for this purpose is to use the chi-square test for association (independence) between the considered variables. Namely, we test the association between the location (internal/external) with the reason of accidents, location with age, location with type, type with reason, type with age, and finally, age with reason. This step can help in the preparation for a more realistic logit model which will be discussed in Section 5.

4.2 Chi-square test for association

The chi-square test of association (or independence) is a non-parametric test that determines whether there is an association between categorical variables (Knoke et al., 2002). The test uses a contingency table to analyse the data by arranging the data according to two categorical variables. The categories for one variable appear in the rows, and the categories for the other variable appear in columns. The conditions of the test are already satisfied because our observations through time are independent, there is no relationship between the attributes (subjects) in each variable, and the sample size is large enough. Using a chi-square test of independence ($\alpha = 0.05$), where

the null Hypothesis $\mathbf{H}_0$ is used to represent that the two relevant variables are not associated (independent) while the alternative hypothesis $\mathbf{H}_A$ represents that the two relevant variables are associated (not independent). Table 4 summarises the result of the hypothesis test between the relevant four variables. Figures 7(a) to 7(f) shows the plots of the frequencies between the attributes for any two-combinations between the four variables. From this result, we can conclude that there is not enough evidence to suggest an association between the street and the reason of accidents. ($X_2(3) > 0.971$, $p = 0.808$). However, for all the other cases, we conclude that there is a significant association between them ($P < 0.05$).

Figure 7 Frequencies of accidents for, (a) streets vs. reason (b) age vs. reason (c) street vs. type (d) street with age (e) type with reason (f) type vs. age (see online version for colours)



(a) (b) (c) (d) (e) (f)

Table 4 Results of the chi-square test for association

Variables	*P value*	$X_2(DF)$	*Significant association*
Street/reason	0.808	$X_2(3) \geq 0.97$	Not reject
Age/reason	0.001	$X_2(5) = 60.49$	Reject
Street/type	0.001	$X_2(3) = 51.26$	Reject
Street/age	0.001	$X_2(2) = 23.03$	Reject
Type/reason	0.001	$X_2(9) = 8097.77$	Reject
Type/age	0.001	$X_2(6) = 47.44$	Reject

5 Logistic regression analysis

5.1 Introduction

In the preceding section we have discussed measures of association and tests of significance that help to determine whether two variables from our list of four considered variables systematically covary and whether the covariation observed in sample data is likely to reflect the population from which the sample was drawn. Although these results may be sufficient for some research purposes, but most researchers want to determine whether such bivariate relationships are affected by other independent factors. In such cases, the research problem changes from describing a two-variable relationship to considering more variables. In this case, one need to employ log-linear analysis where logistic regression is a special case, in which the dependent variable is assumed to be dichotomous that assume two outcomes. The basic dichotomous logistic regression equation for K independent variables is given as below (Knoke et al., 2002):

$$\begin{aligned} logit(\hat{L}) &= \ln\left(\frac{P}{1-P}\right) \\ &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_K \end{aligned} \tag{1}$$

It says that the expected natural log (logit) of the ratio of the two probabilities $\frac{p}{1-p}$, is a linear function of the K predictors. Taking the exponential function to both sides in the equation above, we see that the probability.

$$P(Y=1) = \frac{1}{1+e^{-\hat{L}}} \quad = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_K)}} \tag{2}$$

Notice that equation (1) can be also written as:

$$\begin{aligned} \frac{P(Y=1)}{P(Y=0)} &= \frac{p}{1-p} \\ &= e^{\hat{L}} \\ &= e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_K} \end{aligned} \tag{3}$$

5.2 A proposed logistic regression model for capturing variables interactions

Now, we are in a position to check whether the following categorical variables 'reason', 'type' and 'age' influence the dependent dichotomous variable 'street' which assumes the following two attributes (external = 1, internal = 0) with corresponding probabilities of $\{p, 1-p\}$, respectively, see Table 5.

In our analysis we also include the interactive terms X_1X_2, X_1X_3, X_2X_3 between the assumed independent variables in the logistic model to see if it is playing a role in improving the model. So, we use the following mathematical equation which considers this interaction, given by:

$$\frac{P_{Y=1}}{P_{Y=0}} = e^{\hat{L}} = e^{\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_1X_2+\beta_5X_1X_3+\beta_6X_2X_3} \tag{4}$$

5.3 Results of the bivariate logistic model

The outputs of the model represented in equation (4) are summarised in Table 6; chi-square value is used to test the significance of regression coefficients, where sig. reflects the significance level.

Table 5 Dependent and independent variables in the logistic model

Variable	*Description*	*Code*
Y	Street	0 = internal
		1 = external
X_1	Type	Side
		Sequential
		Back
		Other
X_2	Reason	Negligence
		No enough distance
		Change lane
		Other
X_3	Age	Less than 40 years
		40–60 years
		More than 60 years

Table 6 Variables in the interaction regression model in equation (4)

Independent variables	*Description*	*Chi-square*	*Sig.*
X_1	Type	7.21	0.066
X_2	Reason	5.17	0.16
X_3	Age	4.95	0.084
X_1X_2	Type × reason	12.55	0.184
X_1X_3	Type × age	8.45	0.207
X_2X_3	Reason × age	1.07	0.983

From Table 6, it can be found that the interaction variables have a probability of more than 0.05, so we started to remove the interaction terms sequentially starting with the one with the lowest chi-square value. The results show that in all cases, the interactive variables are not significant to be included in the model. So, we removed all of them. Hence, the reduced form of the model is shown in equation (5) below. Table 7 shows the results of this model.

$$\begin{aligned}\frac{P_{Y=1}}{P_{Y=0}} &= e^{\hat{L}} \\ &= e^{\alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}\end{aligned} \tag{5}$$

The results also show that the second variable X_2 representing the reason of the accident with sig. of 0.375 is not significant and has to be removed. This result for the reason variable in Table 7 emphasises our result in Table 4 when the coefficient of association test shows that this variable is significantly not associated with the street variable. The coefficients of the final reduced form of the logistic model are shown in Table 8. It has only 'type' of street and 'age' as the only independent variables that affect the 'street' which represents the dependent variable.

Table 7 Variables in the standard regression model in equation (5)

Independent variables	*Description*	*Chi-square*	*Sig.*
X_1	Type	51.61	0
X_2	Reason	3.11	0.375
X_3	Age	21.89	0

Table 8 Coefficients of the logistic regression equation

Term	*Variable*	*Coefficient* β_j	*Standard error*	*95% CI*	*Z value*	*V value*	*PIF*
Constants		–0.0435	0.0699	(–0.3667, –0.1382)	–4.33	0.000	test
Type	Back	0	0	(0.0000, 0.0000)	*	*	*
	Other	0.4479	0.0754	(0.3001, 0.5956)	5.94	0	1.65
	Sequential	0.1264	0.0737	(–0.0180, 0.2708)	1.72	0.086	1.7
	Side	0.0218	0.0706	(–0.1166, 0.1603)	0.31	0.757	1.76
Age	Middle age	0	0	(0.0000, 0.0000)	*	*	*
	Old	–0.1868	0.0572	(–0.2988, –0.0747)	–3.27	0.001	1.04
	Young	0.243	0.0914	(0.0638, 0.4221)	2.66	0.008	1.03

Table 9 The odds ratios for categorical predictors

Term	*Variable*	*Odds ratio*	*95% CI*
Type	$X_{1_{Other}}$	1.565	(1.3500, 1.8141)
	$X_{1_{Sequential}}$	1.1347	(0.9821, 1.3110)
	$X_{1_{Side}}$	1.0221	(0.8899, 1.1738)
Age	Old	0.8296	(0.7417, 0.9280)
	Young	1.275	(1.0659, 1.5252)

Now focusing on the attributes related to the relevant independent variables X_1 and X_3, Table 8 shows that the 'back' attribute in the type variable, and the 'middle age'

attribute in the age variable are almost zero also. So, we can eliminate these coefficients and the final reduced logistic regression model can be represented by the following equation:

$$\begin{aligned}\hat{L} &= \ln \frac{P}{1-P} \\ &= -0.253 + 0.448X_{1_{Other}} + 0.126X_{1_{Sequential}} \\ &+ 0.022X_{1_{side}} - 0.187X_{3_{old}} + 0.24X_{3_{Young}}\end{aligned} \tag{6}$$

The above equation resembles a linear, additive multiple regression equation, in that a β_i coefficient indicates by how much the log of the dependent variable's odds changes when the corresponding predictor variable changes by one unit (Knoke et al., 2002). Notice that the dependent variable is not a probability, but rather a logarithm of the odds of two probabilities. Hence, we can write the last equation in terms of probabilities as below:

$$\frac{P_{Y=1}}{P_{Y=0}} = e^{\hat{L}} \tag{7}$$

where $\hat{L}$ is defined in equation (6) above.

Table 9 shows the odds ratios of the variables in the second column and the 95% confidence interval of these ratios in the third column. Moreover, The chi-square test of fitting the model gives a value of 6,671.29 with a P-value of 0.476, so the fitting is significance using level of $\alpha = 0.05$.

6 Discussion

The interpretation of the result goes as follows: holding all the attributes of the variable '$type = X_1$' which are $(X_{1_{Other}}, X_{1_{Sequential}}, X_{1_{side}})$ together with the attribute $X_{3_{old}}$ which is part of the variable '$age = X_3$' at a fixed value, the odds of accident happens in the external street (Y=1) by young age group $X_{3_{Young}}$, over the odds of an accident in the internal street $(Y = 0)$ is equal to $e^{0.24} = 1.27$. In terms of percent change, we can say that the odd for accidents by the young age group in external streets is 27% higher than the odds for internal streets.

Moreover, because the coefficient of $X_{3_{old}}$ is negative which equals to -0.187 we can say that the odds of accident happens in the external street $(Y = 1)$ by the old age group $X_{3_{old}}$ over the odds of an accident in the internal street $(Y = 0)$ is equal to $e^{-0.187} = 0.829$. In terms of percent change, the result means that the odds for accidents by the old age group in external streets is 18% lower than the odds for internal streets.

Furthermore, the results of Table 9 suggest that holding the attributes of the variable $type = X_1$' which are $(X_{1_{Other}}, X_{1_{Side}})$ together with the two attributes of the variable '$Age = X_3$', namely, $X_{3_{old}}$ and $X_{3_{Young}}$ at fixed values, the odds of accident happens in the external street $(Y = 1)$ as sequential type accident $X_{1_{Sequential}}$, over the odds of the same type in internal street $(Y = 0)$ is equal to $e^{0.12} = 1.13$. In terms of percent change, we can say that the odds for accidents that happen as a result of a car hitting another car sequentially in external streets is 13% higher than the odds for internal streets. This result holds irrelevant to the age of driver.

The results and the model can be improved further by including the number of injuries/deaths in the analysis. This variable was not reliable in our data and we did not use it in the analysis. As this variable is presented by continuous values and not categorical. Future work including this variable will be interesting to see the effect on the results and compare it with the current work in this paper.

Another limitation is related to the high-dimensionality of the problem as there are many variables with many attributes. To overcome this problem, we combined many attributes into a limited number from each variable. This can result in missing some of the attributes which can have significant contributions to the model. However, this limitation can open the direction for future research in testing the inclusion of other factors and attributes. For example, in our analysis, the weather was not a factor, because the weather in Emirate of Abu Dhabi is dry most of the year. However, if the same setup of the model is to be implemented on a region where there are 4 seasons throughout the year, then the weather factor will play a major role in the model and it has to be included.

COVID-19 has affected many fields including transportation sector, including traffic accidents. Future research will be directed to fit the model to the period from March 2020 onward and compare it with the current results of this paper. Another path of research if number of injuries is available is to use time series analysis where dynamics is built in within the model. The results of the study can be used and tested also in the whole of UAE and other Arab Gulf countries, as they have the same type of streets and infrastructure, and most of the relevant variables are common in nature.

7 Conclusions and future research directions

In this paper, a practical reduced logistic regression model was proposed and derived after testing for relevant variables taking into account the interacting terms between them. The model is related to traffic accidents in a recent dataset of Abu Dhabi Emirate. Such understanding can advance our knowledge on how to improve road safety. We investigated the data using descriptive and inferential statistical methods. The results show that while there is no association between street and reason. It shows that there is an association between age and reason of accident, street and age, street and type of accident, age and type of accident and type of accident and reason. The analysis shows that some of the variables or its attributes are non-significant and are removed from the model. The variable ‘street’ used as the dependent variable in the reduced logistic model is dichotomous represented with two attributes: external (outside the city) and internal (inside the city). For the independent variables, we have ‘type’ of the collision and ‘age’. In the process of reaching to the final reduced model, the statistical results shows that the interactive terms included in the model were not significant.

The first result coming out from this study is that for young age group who are less than 40 years old, are more likely to be involved in car accidents on external streets (highway) than old age group by 18%. Moreover, another result from the model shows that sequential type accidents are more likely to occur on external streets than internal streets by 13%.

These new findings provide scientific support for new policies that should be considered when teaching practical driving lessons to new seekers for driving licenses. For example, adding compulsory driving test or adding extra compulsory driving lessons

on the highway specifically for young age candidates with age less than 40. In addition, it is recommended that more strict regulations should be put in place, to enforce enough distance between vehicles in the highways. This is to protect both groups either young or senior drivers. We believe that our analysis and conclusions will be of great benefit since it will save many lives.

References

Abdel-Aty, M. (2003) 'Analysis of driver injury severity levels at multiple locations using ordered probit models', *J. Safety Res.*, Vol. 34, No. 5, pp.597–603.

Abdel-Aty, M., Chen, C. and Schott, J. (1998) 'An assessment of the effect of driver age on traffic accident involvement using log-linear models', *Accid. Anal.*, Vol. 30, No. 6, pp.851–861.

Abdella, G.M., Al-Khalifa, K.N., Tayseer, M.A. and Hamouda, A.M.S. (2019) 'Modelling trends in road crash frequency in Qatar state', *International Journal of Operational Research*, Vol. 34, No. 4, pp.507–523.

Abellán, J., López, G. and de Oña, J. (2013) 'Analysis of traffic accident severity using ision rules via decision trees', *Expert Syst. Appl.*, Vol. 40, No. 15, pp.6047–6054.

Agresti, A. (2006) *An Introduction to Categorical Data Analysis*, 2nd ed., Wiley, Hoboken, New Jersey.

Al-Ghamdi, A.S. (2002) 'Using logistic regression to estimate the influence of accident factors on accident severity', *Accid. Anal.*, Vol. 34, No. 6, pp.729–741.

Albuquerque, F. and Awadalla, D. (2020) 'Characterization of road crashes in the Emirate of Abu Dhabi', *Transp. Res. Procedia*, Vol. 48, pp.1095–1110, ISSN: 2352-1465.

Altwaijri, S., Quddus, M.A. and Bristow, A. (2011) 'Factors affecting severity of traffic crashes in Riyadh City', *Proceedings of the Transportation Research Board 90th Annual Meeting*, Washington, DC, USA, pp.23–27.

Bédard, M., Guyatt, G.H., Stones, M.J. and Hirdes, J.P. (2002) 'The independent contribution of driver, crash, and vehicle characteristics to driver fatalities', *Accid. Anal.*, Vol. 34, No. 6, pp.717–727.

Cantillo, V., Márquez, L. and Díaz, C.J. (2020) 'An exploratory analysis of factors associated with traffic crashes severity in Cartagena, Colombia', *Accid. Anal.*, Vol. 146, No. 10, p.105749.

Chang, L-Y. and Mannering, F. (1999) 'Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents', *Accid. Anal.*, Vol. 31, No. 5, pp.579–592.

Chen, P. and Zhou, J. (2016) 'Effects of the built environment on automobile-involved pedestrian crash frequency and risk', *J. Transp. Heal.*, Vol. 3, No. 4, pp.1095–1110.

El Tayeb, A., Pareek, V. and Araar, A. (2015) 'Applying association rules mining algorithms for traffic accidents in Dubai', *Int. J. Soft Comput. Eng.*, Vol. 5, No. 4, pp.2231–2307.

Ghandour, A.J., Hammoud, H. and Al-Hajj, S. (2020) 'Analyzing factors associated with fatal road crashes: a machine learning approach', *Int. J. Environ.*, Vol. 17, No. 11, p.11.

Gilardi, A., Mateu, J., Borgoni, R. and Lovelace, R. (2020) *Multivariate Hierarchical Analysis of Car Crashes Data Considering a Spatial Network Lattice*, arXiv preprint arXiv:2011.12595.

Hazaa, A.S.M., Alnakhalny, R. and Alnaklani, M.A. (2019) 'Prediction of traffic accident severity using data mining techniques in Ibb Province, Yemen', *Eng. Comput. Syst.*, Vol. 5, No. 1, pp.77–92.

Knoke, D., Bohrnstedt, G. and Mee, A. (2002) *Statistics for Social Data Analysis*, 4th ed., F.E. Peacock Publishers, Illinois.

Kumar, S. and Toshniwal, D. (2016) 'A data mining approach to characterize road accident locations', *J. Mod.*, Vol. 24, No. 1, pp.62–72.

Li, Y., Song, L. and Fan, W.D. (2021) 'Day-of-the-week variations and temporal instability of factors influencing pedestrian injury severity in pedestrian-vehicle crashes: a random parameters logit approach with heterogeneity in means and variances', *Anal. Methods Accid. Res.*, Vol. 29, No. 3.

Mulay, P. and Mulatu, S. (2016) 'What you eat matters road safety: a data mining approach', *Indian J. Sci.*, Vol. 9, No. 15, pp.1–8.

Nassar, S.A., Saccomanno, F.F. and Shortreed, J.H. (1994) 'Road accident severity analysis: a micro level approach', *Can.*, Vol. 21, No. 5, pp.847–855.

Nazeri, A., GharehGozlu, H., Faraji, F. and Asakareh, S. (2021) 'Analysis of road accident data and determining affecting factors by using regression models and decision trees', *International Journal of Business Intelligence and Data Mining*, Vol. 18, No. 4, pp.449–471.

Paul, S., Alvi, A.M. and Rahman, R.M. (2021) 'An analysis of the most accident prone regions within the Dhaka metropolitan region using clustering', *International Journal of Advanced Intelligence Paradigms*, Vol. 18, No. 3, pp.294–315.

Rolison, J.J., Regev, S., Moutari, S. and Feeney, A. (2018) 'What are the factors that contribute to road accidents?', *An Assessment of Law Enforcement Views, Ordinary Drivers' Opinions, and Road Accident Records*, Vol. 115, No. 6, pp.11–24.

Shankar, V., Mannering, F. and Barfield, W. (1996) 'Statistical analysis of accident severity on rural freeways', *Accid. Anal.*, Vol. 28, No. 3, pp.391–401.

Tarko, A. and Azam, M.S. (2011) 'Pedestrian injury analysis with consideration of the selectivity bias in linked police-hospital data', *Accid. Anal.*, Vol. 43, No. 5, pp.1689–1695.

Tay, R., Choi, J., Kattan, L. and Khan, A. (2011) 'A multinomial logit model of pedestrian-vehicle crash severity', *Int. J. Sustain.*, Vol. 5, No. 4, pp.233–249.

Theofilatos, A. (2017) 'Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials', *J. Safety Res.*, Vol. 61, No. 6, pp.9–21.

Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Saher, N. and Madni, H.A. (2020) *Comparison Analysis of Tree Based and Ensembled Regression Algorithms for Traffic Accident Severity Prediction*, arXiv preprint arXiv:2010.14921.

Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S. and Barbone, F. (2002) 'Risk factors for fatal road traffic accidents in Udine, Italy', *Accid. Anal.*, Vol. 34, No. 1, pp.71–84.

Wali, B., Khattak, A.J. and Karnowski, T. (2020) 'The relationship between driving volatility in time to collision and crash-injury severity in a naturalistic driving environment', *Anal. Methods Accid. Res.*, Vol. 28, No. 12.

WHO (2018) *Global Status Report on Road Safety*.

Xie, L. (2020) 'Statistical analysis of fatal crash in Michigan using more than two time series models', *International Journal of Data Science*, Vol. 5, No. 1, pp.26–40.

Xie, Y., Zhang, Y. and Liang, F. (2009) 'Crash injury severity analysis using Bayesian ordered probit models', *J. Transp. Eng.*, Vol. 135, No. 1, pp.18–25.

Yau, K.K.W. (2004) 'Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong', *Accid. Anal.*, Vol. 36, No. 3, pp.333–340.

Zajac, S.S. and Ivan, J.N. (2003) 'Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural Connecticut', *Accid. Anal.*, Vol. 35, No. 3, pp.369–379.

Zhang, A., Patton, E.W., Swaney, J.M. and Zeng, T.H. (2019) *A Statistical Analysis of Recent Traffic Crashes in Massachusetts*, arXiv preprint arXiv:1911.02647.