Modeling Risk of Road Crashes Using Aggregated Data

Randa Oqab Mujalli

Department of Civil Engineering, The Hashemite University, 13115 Zarqa, Jordan.
E-Mail: randao@hu.edu.jo

ABSTRACT

Traffic crashes constitute a major burden on most governments, especially in countries classified as middle- or low-income countries. Most low- or middle-income countries suffer either from scarce or no crash records' data or unreliable data, where annual crash reports become the main source of data available to investigate crashes in order to find and take countermeasures to reduce both frequency and severity of crashes. This paper aimed at using aggregate annual crash reports for 18 years, in order to determine the main factors that contribute to characterizing crashes in a specific year according to severity. Identification of these factors was made possible using Bayesian networks, in which three different models were developed. The main contributing factors which were found to increase the likelihood of classifying crashes as severe or fatal were: higher number of traffic control device violations, speeding, higher number of run-off-road crashes and higher number of pedestrian crashes.

KEYWORDS: Crash annual reports, Classification, Severity, Aggregated data.

INTRODUCTION

Road crashes are described by the World Health Organization (WHO) as the leading cause of death among young people aged between 15 and 29 years and the 9th leading cause of death across all age groups. Further, they are predicted to be the seventh leading cause of death by 2030. In addition to loss of productivity and human lives caused by road crash fatalities, global economies are overburdened with an annual cost of road crashes of approximately 3% of GDP, rising up to 5% of GDP in low- to middle- income countries. More than 1.25 million fatalities are caused by road crashes, having a considerable effect on health and development with no effective countermeasures taken to reduce human or economic toll. On the other hand, some of the measures taken during the period from 2010 to 2013 proved to be successful in stabilizing road traffic deaths accompanied by a 4% increase in global population and 16% increase in registered vehicles over the same period.

Eighty two percent of world population reside in low- and middle-income countries. 90% of global road crash fatalities occur in these countries, accounting for 54% of the world's registered vehicles. Number of road crash fatalities occurring in low- and middle- income countries represent double the fatalities caused by road crash in high-income countries. Moreover, 45% of road crash deaths in low- and middle-income countries are pedestrians, cyclists and motorists (WHO, 2015).

Jordan is classified by the World Bank as a high-middle income country (World Bank, 2016), with a population of 9.5 million and more than 1.4 million registered vehicles (Department of Statistics, 2016). Total number of road crashes on Jordanian roads is estimated at more than 195 thousands with more than 92 thousand property damage-only crashes. Number of

Received on 10/4/2017.
Accepted for Publication on 2/6/2017.

road crashes with fatalities is estimated at 449 crashes resulting in 688 fatalities, added to 9310 crashes with injuries resulting in 2063 severe injuries and 12727 slight injuries (Police Security Directorate, 2014). Over the past decade, population increased by 70% and number of registered vehicles increased by 76%, while road crashes with injuries increased for the same period by approximately 5% and fatalities decreased by 30% as compared to 2006. It should be mentioned that the trend of the number of fatalities over the past decade was not a declining trend over the whole period. It witnessed a series of inclines and declines and was thus unstable.

Traffic crash studies in developing countries are very limited in findings because of the limited data available at time of research. Most of the data collected do not provide enough information of the single crash record, where causes of the crash are normally "guessed" by the police officer at site, who is in most cases not well trained to define and decide what caused the crash.

In order to monitor, assess, compare and tailor prevention efforts, reliable traffic crash registration data should become available for all safety analysts and researchers all around the world. However, crash data is not collected and stored in a unified global manner; variations in type of data collected and the way they are collected are ambiguous in some low- and middle-income countries. Therefore, results and findings of research carried out are affected by the level of confidence of the original data collected and might lead to erroneous decisions and consequently to improper countermeasures and strategies. Therefore, aggregated traffic crash data represent the only source of information used by safety analysts and researchers in many developing countries around the world.

Aggregated traffic crash data consist of time series in which crashes are recorded over time (annually, monthly, weekly) and described based on one or more of the following factors: total number of traffic crahes, traffic crash fatalities, traffic crash injuries, total vehicle-kilometers,… etc. Therefore, one of the most classical used traffic crash analysis techniques is linear regression. However, linear regression models are generally not well suited to handle dependencies between the consecutive observations making up a time series, since they likely yield serially correlated residuals.

In general, regression models that are usually used to model traffic crashes assume that the sample data comes from a population that follows a specific distribution (i.e., normal distribution). However, if the assumption is incorrect, the statistical model used will lead to wrong conclusions (Chang and Wang, 2006).

In Jordan, studies were so far based on aggregate data available from annual reports issued by the Police Security Directorate, which is the agency responsible for collecting, maintaining and analyzing traffic crash data. Researchers are normally encountered with the difficulty of obtaining disaggregate data and hence they tend to use aggregated crash data.

Al-Suleiman and Al-Masaeid (1992) studied the fatality rates of traffic crashes in Jordan using aggregated traffic crash reports. They developed a descriptive model using both fatality rates and motorization levels for the years 1970-1988, where the realtionship was found to be non-linear and thus they used the standard regression technique to linearize the relationship. The results indicated that heavy vehicles are more likely to get involved in traffic crashes.

Pedestrain crashes were studied by Al-Masaeid and Nelson (1996). The researchers used two datasets; the first dataset was used to study the pedestrian relative involvement rates and the spatial distribution of pedestrian crashes using data on pedestrian crash information, pedestrian behavior and population. The second dataset was used to establish the relationship between pedestrian crashes and both street geometrics and operation variables on 30 streets with high pedestrian crashes in Amman city. Their results indicated that 78% of pedestrian crashes were not nearby intersections, where a strong relationship was found between pedestrian crashes and street geometry and operation variables.

In their study, Al-Masaeid et al. (1997) collected traffic, land use and geoemtric characteristics of 64 arterial midblocks in order to find out what effect these variabels have on the ocuurence of pedestrian crashes. Their results indicated that traffic flow during peak hours, commercial and public buildings on arterial roads and lack of or insufficient sidewalk width all increase the probability of pedestrian crash occurrence.

To analyze the factors that affect both number of fatalities and number of crashes, Al-Masaeid (2009) used both descriptive analysis and regression to estimate number of annual traffic crashes and number of annual fatalities based on totals of population, number of registered vehicles, number of accidents and number of fatalities for the years 1998 to 2007. It was found that number of registered vehicles per 1000 of population was the only factor affecting both number of annual crashes and number of annual fatalities. This result was also found by Al-Omari et al. (2013) who studied the same variables, but for the years between 1998 and 2010.

In addition, Al-Masaeid (2016) evaluated the effects of modified policy measures and high fuel prices on traffic safety using aggregated traffic crash data from 1998 to 2008 on Jordanian roads. A trend analysis was used to find out the effects of new safety policies and high fuel prices on traffic safety improvement. The results indicated that strict safety policies and stiff penalties both assisted in reducing traffic crashes by 14% and traffic fatalities by 25%, while no significant effect of fuel prices was found to affect traffic safety.

The only study found in Jordan which used disaggregate crash record data was performed by Mujalli et al. (2016), who used crash records for the years from 2009 to 2011 for crashes which occurred in urban and suburban areas. They used Bayesian networks to model injury severity of traffic crashes using 13 variables. They found that number of vehicles involved, accident pattern, number of directions, accident type, lighting, surface condition and speed limit are the variables that contribute to the occurrence of fatalities and severe injuries in traffic crashes.

To this end, the goal of this research work is to determine the primary causes of traffic crashes in developing countries using aggregated traffic crash data as obtained from annual crash reports employing Bayesian networks in order to determine the individual level of contribution of each cause in traffic crashes. The methodology followed herein will help researchers in developing countries forecast the probablity of traffic crash occurrence and develop safety strategies to mitigate the growing traffic safety problem.

Traffic Crashes in Jordan

Statistical reports issued by Police Security Directorate indicate that the percentage of increase in traffic crashes since 1997 is estimated by 163%. This increase in the number of crashes in the period 1997-2014 was accompanied by an increase in the number of registered vehicles by 267%. However, crashes with casualties and number of fatalities have slightly decreased in 2014 as compared to 1997.

The change in the total number of crashes with casualties since 1997 has been fluctuating. The maximum number of crashes with casualties was in 2005. After 2005, a decline in the number of crashes with casualties was noticed until 2009, where the number of crashes with casualties continued increasing for 3 years, followed by a decrease over the period from 2012 to 2014.

Bayesian Networks

Bayesian networks (BNs) consist of a directed acyclic graph (DAG), in which some variables (nodes) might be connected by directed edges (arcs). In this case, a relationship is said to exist between these connected variables. When the arc is directed from the variable, the variable is called a parent, whereas the variable to which the arc is directed is called a child (Neapolitan, 2009). Edges between variables can represent a causal or dependence relationship (Acid et al., 2004).

A DAG is a pair (X, A), where X is a finite, nonempty set whose elements are called nodes, where $X=\{X_1, X_2, \ldots, X_n\}$ and A are a set of arcs, where $A=\{a_{ij}\}$ and a_{ij} describes a direct dependence

relationship between X_i and X_j. Elements of A are called directed edges and if $(X_i, X_j) \in X$ and there is an edge from X_i to X_j, then X_i is called a parent (non-descendent) of X_j, where X_j is called a child (descendent) of X_i. The conditional probabilities of a BN quantify the dependencies between variables and their parents in the DAG, where a BN can be described using joint probability distribution of the variable set $X= \{X_1,X_2,\ldots X_n\}$:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \left| \prod (X_i) \right.)$$

There are two ways to develop a BN; either using expert prior knowledge about variables and dependencies among them, which is normally used in medical diagnosis, or learning BN from data, as used herein, in which A BN over variables $X_1, \ldots, X_n$ can be learned from a data set over these variables, which is a table with each row representing a partial instantiation of variables $X_1, \ldots, X_n$. There are two steps in order to learn a BN; the first step is to learn the network structure, in which a structure might either be known or unknown. The second step is to learn the probabilities. For this research work, a structure of the BN is assumed to be unknown and thus this type of BN will be discussed. Readers interested in learning from known structures are advised to refer to Darwiche (2008).

In order to learn the structure of a BN, iterative algorithms are used, starting with the empty graph and incrementally modifying this structure until reaching some termination condition. There are also two types of algorithm: score-based algorithms (which are used herein) and conditional independence-based algorithms. Conditional independence-based algorithms are mainly aimed at uncovering causal structure and are sensitive to errors in individual tests. The assumption is that there is a network structure that exactly represents the independencies in the distribution that generated the data (Darwiche, 2008). It follows that if a conditional independency can be identified in the data between two variables, there is no arrow between those two variables. Once locations of edges are identified, the direction of the edges is assigned such that conditional independencies in the data are properly represented (Bouckaert, 2004). The advantage of score-based methods over constraint-based methods is that they are less sensitive to errors in individual tests; compromises can be made between the extent to which variables are dependent in the data and the cost of adding the edge.

Score-based algorithms employ either local or global search, where both use cross-validation which provides an out of sample evaluation method to facilitate this by repeatedly splitting the data in training and validation sets. A Bayesian network structure can be evaluated by estimating the network parameters from the training set and the resulting Bayesian network performance determined against the validation set. The average performance of the Bayesian network over the validation sets provides a metric for the quality of the network.

Local search algorithms evaluate the current BN structure (which might be an empty structure), as well as every structure formed by some simple modification and climb to the new structure with the highest score, which can be considered an optimization problem, where a quality measure of a network structure gives the training data needs to be maximized. The quality measure can be based on a Bayesian approach, minimum description length, information and other criteria. Those metrics have the practical property that the score of the whole network can be decomposed as the sum (or product) of the scores of the individual nodes (Bouckaert, 2004). Unfortunately, this does not always work because of overfitting that shows up in learning Bayesian networks, as it would favor a fully connected network and hence this complete graph would maximize the probability of data due to its large set of parameters (maximal degrees of freedom) (Darwiche, 2008).

On the other hand, global search measures how well a BN performs on a given data set, by predicting its

future performance through estimating expected utilities, such as classification accuracy. Cross-validation differs from local scoring metrics in that the quality of a network structure often cannot be decomposed in the scores of the individual nodes and thus the whole network needs to be considered in order to determine the score. In this research work, global search algorithms are used, which are briefly described below:

1. *Hillclimbing search* (Bouckaert, 2004): is an iterative algorithm that starts with an arbitrary solution to a problem, then attempts to find a better solution, by incrementally changing a single element of the solution.

2. *Repeated hill climber* starts with a randomly generated network and then applies hill climber to reach an optimum (Bouckaert, 2004).

3. *Tabu search* (Bouckaert, 1995): the basic principle of tabu search is to pursue search whenever it encounters a local optimum by allowing non-improving moves; cycling back to previously visited solutions is prevented by the use of memories, called tabu lists that record the recent history of the search.

4. *Simulated annealing*: it tries to avoid being trapped in local optima by accepting both "good" and "bad" moves at the beginning of the iterations and gradually lowering the probability of accepting "bad" moves. Even though in theory simulated annealing can find global optima, if we represent the above probability slowly in exponential time, its performance in a practical time frame depends heavily on the parameters comprising its "cooling schedule". In general, simulated annealing is time-consuming, but has been successfully applied to many optimization problems (Tao et al., 1992).

5. *TAN* (Cheng and Greiner, 1999; Friedman et al., 1997): Tree Augmented Naive Bayes, where the tree is formed by calculating the maximum weight spanning tree using algorithm of Chow and Liu (1968).

6. *Genetic search* (Bouckaert, 2004): applies a simple implementation of a genetic search algorithm to network structure learning. A BN structure is represented by an array of $n \cdot n$ (n = number of nodes) bits, where bit $i \cdot n + j$ represents whether there is an arrow from node $j \rightarrow i$.

Weka software (Witten and Frank, 2005) was used in this study to build the BN. This software is freely available. It is implemented in Java language and contains a collection of data processing and modeling techniques. It also contains a graphical user interface.

Road Crash Data

Aggregated road crash data was obtained from annual reports issued by Police Security Directorate. The analyzed years were from 1997 to 2014 for traffic crashes which occurred on Jordanian roads, in which the variables found to be available in all years were considered. Other variables related to socio-economic and demographic characteristics were obtained for the same years from the annual reports issued by the Department of Statistics. A detailed illustration of the data obtained is found in Tables 1, 2 and 3.

Table 1. Road crash data obtained from PTD (number of casualties and number of crashes)

	Number of casualties			Number of crashes according to crash type		
Year	**Fatalities**	**Severe injuries**	**Slight injuries**	**Collision**	**Pedestrian**	**run-off-road**
1997	577	4286	11973	27602	5669	2308
1998	612	4535	12642	30799	5964	2545
1999	676	4688	14327	36514	6105	2793
2000	686	5167	13675	39162	5840	2630
2001	783	2189	16643	40890	5525	2582
2002	758	2247	15134	42011	5417	2206
2003	832	2514	15854	51159	5345	2174
2004	818	2451	14276	58803	5079	1971
2005	790	2598	14981	76261	4866	2002
2006	899	2941	15078	91075	4826	2154
2007	992	2882	15087	104576	4178	1876
2008	740	2527	11386	95085	4146	1835
2009	676	1556	14106	117119	4054	1620
2010	670	2964	14439	134328	4091	1595
2011	694	2456	15666	137889	3223	1476
2012	816	1966	15177	107957	3313	1547
2013	768	2258	13696	102185	3954	1725
2014	688	2063	12727	96756	3839	1846

Table 2. Road crash data obtained from PTD (number of violations)

	Number of violations committed by drivers according to type of fault				
Year	**Speeding**	**Crossing lanes**	**Driving too close (tailgating)**	**Disobeying traffic control devices**	**Not respecting priority**
1997	2314	6108	7642	1780	4808
1998	1895	7121	7986	2196	4772
1999	2283	9734	8912	3280	4514
2000	2475	9889	9644	3103	5161
2001	2261	10219	10213	3043	5176
2002	1967	12775	12378	4654	459
2003	1970	9991	12901	2880	5554
2004	1771	10647	13225	4770	7511
2005	1561	11076	14591	2527	11542
2006	1114	13403	14100	2681	14339
2007	1525	17817	22196	2877	17804
2008	n.d.*	n.d.	n.d.	n.d.	n.d.
2009	2620	25847	27211	4403	18413
2010	2125	25417	30720	3637	21137
2011	2915	26019	33041	4537	21757
2012	2557	19892	23553	2901	18544
2013	1070	18838	22236	2112	16055
2014	1189	9584	22444	2207	13578

***** n.d.: no data available.

Table 3. Socio-economic and demographic characteristics obtained from DOS

Year	Registered vehicles	Population	Number of foreigners	Gross domestic product (GDP)	Length of road network
1997	362811	4444000	1132000	7244402962	7022
1998	389196	4564000	1772000	7910621157	7133
1999	418433	4680000	1790000	8147494358	7200
2000	473339	4797000	1580000	8457923956	7245
2001	509832	4917000	1672000	8972965058	7259
2002	535112	5038000	2384000	9580161861	7301
2003	568096	5164000	2353000	10193023676	7364
2004	614614	5290000	2853000	11407566734	7500
2005	679731	5411000	2987000	12588665303	7601
2006	755477	5536000	3225000	15056936954	7694
2007	841933	5661000	3431000	17110615283	7768
2008	905592	5786000	3729000	21972870921	7816
2009	994753	5915000	3789000	23818322958	7878
2010	1075453	6046000	4207000	26425379437	7100
2011	1147258	6181000	3960000	28840263380	7204
2012	1213882	6318000	4162000	30937277606	7234
2013	1263754	6460000	3945000	33593843662	7299
2014	1331563	6607000	4178896	35826925775	7339

Data Filtering and Processing

Initially, data was filtered in order to be suitable for modeling purposes. As illustrated in Table 2, there is a missing record for the year 2008 in the variables under (number of violations committed by drivers according to type of fault). To this end, a filter was used to replace missing values with means of other values in the same variables.

Next, all variables were categorized according to PKI discretization, which discretizes numerical variables, where the number of categories is equal to the square root of the number of non-missing values (Yang and Webb, 2001).

Since the target of this research work is to find out the variables that increase the probability of having highly severe crashes, a new variable was introduced, named: high or fatility severity risk (FSR). The variable is calculated according to the following equation:

$$FSR = \frac{number\ of\ fatalities\ and\ high\ severe\ injuries}{total\ number\ of\ injuries}$$

After FSR was calculated, the average of FSR values for all years used in the analysis was computed and values above the average (average=0.20) FSR were classified as "highly severe" with the abbreviation "H" and those below or equal to the average FSR were classified to be of "low severity" with the abbreviation "L". The reason for this classification is that the classifier used needs to have nominal classes in order to perform correctly. The results are shown in Tables 4, 5 and 6.

Table 4. Road crash data categorized according to FSR with corresponding abbreviations (number of crashes)

Number of crashes according to crash type				
FSR	**Class**	**Collision (COL) (10^3)**	**Pedestrian (PED) (10^3)**	**Run-off-road (ROR) (10^3)**
0.29	H	≤41.45	>5.60	(1.99-2.43]
0.29	H	≤41.45	>5.60	>2.43
0.27	H	≤41.45	>5.60	>2.43
0.30	H	≤41.45	>5.60	>2.43
0.15	L	≤41.45	(4.85-5.60]	>2.43
0.17	L	(41.45-83.67]	(4.85-5.60]	(1.99-2.43]
0.17	L	(41.45-83.67]	(4.85-5.60]	(1.99-2.43]
0.19	L	(41.45-83.67]	(4.85-5.60]	(1.78-1.99]
0.18	L	(41.45-83.67]	(4.85-5.60]	(1.99-2.43]
0.20	L	(83.67-106.27]	(4.07-4.85]	(1.99-2.43]
0.20	H	(83.67-106.27]	(4.07-4.85]	(1.78-1.99]
0.22	H	(83.67-106.27]	(4.07-4.85]	(1.78-1.99]
0.14	L	>106.27	≤4.07	≤1.78
0.20	L	>106.27	(4.07-4.85]	≤1.78
0.17	L	>106.27	≤4.07	≤1.78
0.15	L	>106.27	≤4.07	≤1.78
0.18	L	(83.67-106.27]	≤4.07	≤1.78
0.18	L	(83.67-106.27]	≤4.07	(1.78-1.99]

Table 5. Road crash data categorized according to FSR with corresponding abbreviations (number of violations)

Number of violations committed by drivers at fault				
Speeding (10^3)	**Crossing lanes (WL) (10^3)**	**Driving too close (tailgating) (REAR) (10^3)**	**Disobeying traffic control devices (TCD) (10^3)**	**Not respecting priority (ROW) (10^3)**
(1.97-2.39]	≤9.94	≤ 11.30	≤2.60	≤5.17
(1.67-1.97]	≤9.94	≤ 11.30	≤2.60	≤5.17
(1.97-2.39]	≤9.94	≤ 11.30	(2.97-4.02]	≤5.17
>2.39	≤9.94	≤ 11.30	(2.97-4.02]	≤5.17
(1.97-2.39]	(9.94-11.93]	≤ 11.30	(2.97-4.02]	(5.17-11.39]
(1.67-1.97]	(11.93-19.37]	(11.30-14.35]	>4.02	≤5.17
(1.67-1.97]	(9.94-11.93]	(11.30-14.35]	(2.60-2.97]	(5.17-11.39]
(1.67-1.97]	(9.94-11.93]	(11.30-14.35]	>4.02	(5.17-11.39]
≤1.67	(9.94-11.93]	(14.35-23.00]	≤2.60	(11.39-18.11]
≤1.67	(11.93-19.37]	(11.30-14.35]	(2.60-2.97]	(11.39-18.11]
≤1.67	(11.93-19.37]	(14.35-23.00]	(2.60-2.97]	(11.39-18.11]
(1.97-2.39]	(11.93-19.37]	(14.35-23.00]	(2.97-4.02]	(5.17-11.39]
>2.39	>19.37	>23.00	>4.02	>18.11
(1.97-2.39]	>19.37	>23.00	(2.97-4.02]	>18.11
>2.39	>19.37	>23.00	>4.02	>18.11
>2.39	>19.37	>23.00	(2.60-2.97]	>18.11
≤1.67	(11.93-19.37]	(14.35-23.00]	≤2.60	(11.39-18.11]
≤1.67	≤9.94	(14.35-23.00]	≤2.60	(11.39-18.11]

Table 6. Socio-economic and demographic characteristics categorized according to FSR with corresponding abbreviations

FSR	Class	Registered vehicles (RV) (10^3)	Population (POP) (10^3)	Number of foreigners (FA) (10^3)	Gross domestic product (GDP) (10^6)	Length of road network (LEN) (10^3)
0.29	H	≤522.5	≤4977.5	≤2071.5	≤9276.6	≤7.22
0.29	H	≤522.5	≤4977.5	≤2071.5	≤9276.6	≤7.22
0.27	H	≤522.5	≤4977.5	≤2071.5	≤9276.6	≤7.22
0.30	H	≤522.5	≤4977.5	≤2071.5	≤9276.6	(7.22-7.3]
0.15	L	≤522.5	≤4977.5	≤2071.5	≤9276.6	(7.22-7.3]
0.17	L	(522.5-717.6]	(4977.5-5473.5]	(2071.5-3106.0]	(9276.6-13822.8]	(7.30-7.65]
0.17	L	(522.5-717.6]	(4977.5-5473.5]	(2071.5-3106.0]	(9276.6-13822.8]	(7.30-7.65]
0.19	L	(522.5-717.6]	(4977.5-5473.5]	(2071.5-3106.0]	(9276.6-13822.8]	(7.30-7.65]
0.18	L	(522.5-717.6]	(4977.5-5473.5]	(2071.5-3106.0]	(9276.6-13822.8]	(7.30-7.65]
0.20	L	(717.6-1111.4]	(5473.5-6113.5]	(3106.0-3952.5]	(13822.8-27632.8]	>7.65
0.20	H	(717.6-1111.4]	(5473.5-6113.5]	(3106.0-3952.5]	(13822.8-27632.8]	>7.65
0.22	H	(717.6-1111.4]	(5473.5-6113.5]	(3106.0-3952.5]	(13822.8-27632.8]	>7.65
0.14	L	(717.6-1111.4]	(5473.5-6113.5]	(3106.0-3952.5]	(13822.8-27632.8]	>7.65
0.20	L	(717.6-1111.4]	(5473.5-6113.5]	>3952.5	(13822.8-27632.8]	≤7.22
0.17	L	>1111.4	>6113.5	>3952.5	>27632.8	≤7.22
0.15	L	>1111.4	>6113.5	>3952.5	>27632.8	(7.22-7.30]
0.18	L	>1111.4	>6113.5	(3106.0-3952.5]	>27632.8	(7.22-7.30]
0.18	L	>1111.4	>6113.5	>3952.5	>27632.8	(7.30-7.65]

Research Methodology

After the data was filtered and processed, the data was divided into 3 subsets in order to develop 3 different models with the aim to classify crashes according to FSR. The models used are summarized below:

1. Model 1: using variables related to violations committed by drivers according to type of fault (speeding, wrong lane, driving too close, disobeying traffic control devices and not giving right-of-way).
2. Model 2: using variables related to socio-economic and demographic characteristics (number of registered vehicles, population, foreigners, GDP and length of road network).
3. Model 3: using the crash type variables (number of collisions, number of pedestrian crashes and number of ROR crashes).

Modeling Results and Discussion

Seven different BNs were developed using each of the 3 subsets and applying different search algorithms described in previous sections. For each BN, accuracy of the models developed and AUCs were used to compare the different BNs in each subset. The results obtained are summarized in Table 7.

Table 7. Developed BNs using the 3 subsets and applying different search algorithms

Search algorithms	Subset 1		Subset 2		Subset 3	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Genetic	0.83	0.81	0.72	0.86	0.78	0.94
Hillclimber	0.56	0.75	0.72	0.86	0.78	0.9
K2	0.67	0.78	0.72	0.86	0.78	0.92
Repeated hillclimber	0.72	0.76	0.72	0.86	0.78	0.91
Simulated annealing	0.72	0.78	0.72	0.86	0.78	0.94
Tabu	0.5	0.75	0.72	0.85	0.78	0.92
TAN	0.78	0.81	0.61	0.78	0.78	0.86

These subsets were used to develop 3 different models using BNs and applying global search simulated annealing algorithm.

As illustrated in Table 7, the BNs developed using subsets 1 and 3 had higher accuracies and AUCs when using genetic search algorithm, while for subset 2 the results obtained were similar for all search algorithms except for TAN. Consequently, the BNs developed using genetic search algorithm will be used to determine variables affecting FSR.

Bayesian Models

The models developed using genetic algorithm are used herein, since they lead to obtain highest values in both accuracy and AUC as previously described. As shown below, the joint probability distributions of the models are presented in order to investigate the dependencies between variables, as well as between the class variable and the rest of variables.

Joint Probability Distribution of Model 1

The following equation is the joint probability developed for model 1 when using genetic algorithm to learn the model:

$$P(FSR, WL, ROW, REAR, TCD, Speeding) = P(WL) * P(REAR \mid WL) * P(FSR \mid WL, REAR) * P(ROW \mid FSR, WL) * P(TCD \mid FSR, REAR) * P(Speeding \mid FSR, ROW, TCD, REAR)$$

As shown above, there are dependencies among the independent variables, as well as between the class variable and the rest of variables. The following sets of variables have dependencies among each other: (REAR, WL), (ROW, WL), (TCD, REAR), (Speeding, ROW), (Speeding, TCD) and (Speeding, REAR).

On the other hand, all the variables used have direct dependence with the class variable FSR.

Joint Probability Distribution of Model 2

The following equation is the joint probability developed for model 2 when using genetic algorithm to learn the model:

$$P(FSR, FA, LEN, POP, GDP, RV) = P(FSR) * P(FA \mid FSR) * P(LEN \mid FA) * P(POP \mid FSR, LEN) * P(RV \mid FSR, POP) * P(GDP \mid FSR, POP, RV)$$

As shown above, there are dependencies among the independent variables, as well as between the class variable and the rest of variables. The following sets of variables have dependencies among each other: (LEN, FA), (POP, LEN), (RV, POP), (GDP, POP) and (GDP, RV).

On the other hand, the following variables have direct dependence with the class variable FSR: POP, RV and GDP. The variables that do not have direct dependence with FSR are: LEN and FA.

Joint Probability Distribution of Model 3

The following equation is the joint probability developed for model 3 when using genetic algorithm to learn the model:

$$P(FSR, PED, COL, ROR) = P(PED) * P(FSR \mid PED) * P(ROR \mid FSR, PED) * P(COL \mid FSR, PED, ROR)$$

As shown above, there are dependencies among the independent variables, as well as between the class variable and the rest of variables. The following sets of variables have dependencies among each other: (ROR, PED), (COL, PED) and (COL, ROR).

On the other hand, all the variables used have direct dependence with the class variable FSR.

Using Partial Evidence

Using partial evidence, the way that class probabilities are affected by using different variables and the extent to which each variable value is relevant to the classification of a particular instance are investigated.

This can be carried out by setting the observation for one variable at a time and then checking the class probabilities. The following sub-sections show the results of partial evidence set to each variable used in each of the 3 models developed.

Model 1: Drivers at Fault and FSR

Table 8 shows the posterior probability of (H) outcome when prior evidence is set to 1.0 for each category in each variable used to develop the model.

It is clear that REAR is the strongest variable to support (H) outcome, in which when setting a probability of 1.0 to (REAR ≤ 11.30), a posterior probability of 0.7087 resulted for (FSR=H), indicating that lower number of REAR increases the probability of having an (H) outcome in a crash. The second strongest variable to support (H) is WL, where setting an evidence to (WL ≤9.94) resulted in a posterior probability of 0.7035 for (FSR=H). It is now evident that lower WL number increases the probability of having an (H) outcome.

Both ROW and TCD had lower effect on FSR, where lower number of ROW (ROW≤5.17) resulted in 0.6754 probability of having an (H) outcome. On the other hand, higher number of TCD (2.97-4.02] resulted in 0.5688 probability of having an (H) outcome in a crash.

Model 2: Socio-economic and Demographic Characteristics and FSR

Table 9 shows the posterior probability of (H) outcome when prior evidence is set to 1.0 for each category in each variable used to develop the model. It is clear that FA is the strongest variable to support (H) outcome, in which when setting a probability of 1.0 to (FA ≤2071.5), a posterior probability of 0.7319 resulted for (FSR=H), indicating that lower number of FA increases the probability of having an (H) outcome in a crash. The second strongest variable to support (H) is POP, where setting an evidence to (POP ≤4977.5) resulted in a posterior probability of 0.6190 for (FSR=H). It is now evident that lower POP number increases the probability of having an (H) outcome.

Both RV and GDP had lower effects on FSR, where lower number of RV (RV≤522.5) resulted in 0.6076 probability of having an (H) outcome. In addition, lower number of GDP (GDP≤9276.6) resulted in 0.5531 probability of having an (H) outcome in a crash.

Table 8. Partial evidences in model 1

Variable	Evidence given to categories	Probability of FSR=H
Speeding	P(Speeding ≤1.67) = 1.0	0.3743
	P(Speeding = (1.67-1.97]) = 1.0	0.4170
	P(Speeding = (1.97-2.39]) = 1.0	0.4676
	P(Speeding >2.39) = 1.0	0.3943
WL	P(WL ≤9.94) = 1.0	**0.7035**
	P(WL = (9.94-11.93]) = 1.0	0.2361
	P(WL = (11.93-19.37]) = 1.0	0.4434
	P(WL >19.37) = 1.0	0.2000
REAR	P(REAR ≤ 11.30) = 1.0	**0.7087**
	P(REAR = (11.30-14.35]) = 1.0	0.2222
	P(REAR = (14.35-23.00]) = 1.0	0.4572
	P(REAR >23.00) = 1.0	0.2023
TCD	P(TCD ≤ 2.60) = 1.0	0.4504
	P(TCD = (2.60-2.97]) = 1.0	0.3701
	P(TCD = (2.97-4.02]) 1.0	**0.5688**
	P(TCD >4.02) =1.0	0.2387
ROW	P(ROW ≤5.17) = 1.0	**0.6754**
	P(ROW = (5.17-11.39]) = 1.0	0.3747
	P(ROW = (11.39-18.11]) = 1.0	0.3301
	P(ROW >18.11) = 1.0	0.2387

Table 9. Partial evidences in model 2

Variable	Evidence given to categories	Probability of FSR=H
RV	P(RV≤522.5) = 1.0	**0.6076**
	P(RV = (522.5-717.6]) = 1.0	0.1985
	P(RV = (717.6-1111.4]) = 1.0	0.3554
	P(RV >1111.4) = 1.0	0.2025
POP	P(POP ≤4977.5) = 1.0	**0.6190**
	P(POP= (4977.5-5473.5])= 1.0	0.2002
	P(POP= (5473.5-6113.5])= 1.0	0.3417
	P(POP >6113.5) = 1.0	0.2063
FA	P(FA ≤2071.5) = 1.0	**0.7319**
	P(FA = (2071.5-3106.0]) = 1.0	0.0918
	P(FA = (3106.0-3952.5]) = 1.0	0.3939
	P(FA >3952.5) = 1.0	0.0918
GDP	P (GDP≤9276.6) = 1.0	**0.5531**
	P (GDP = (9276.6-13822.8]) = 1.0	0.2362
	P (GDP = (13822.8-27632.8]) = 1.0	0.3441
	P (GDP >27632.8) = 1.0	0.2397
LEN	P (LEN ≤7.22) = 1.0	0.4285
	P (LEN = (7.22-7.3]) = 1.0	0.4300
	P (LEN = (7.30-7.65]) = 1.0	0.1577
	P (LEN >7.65) = 1.0	0.3707

Model 3: Crash Type and FSR

Table 10 shows the posterior probability of (H) outcome when prior evidence is set to 1.0 for each category in each variable used to develop the model. In this model, PED had two categories that resulted in the two highest posterior probabilities of FSR=H, indicating that PED is the strongest variable to support an (H) outcome. Setting a probability of 1.0 to (PED = (4.85-5.60]), a posterior probability of 0.9166 resulted for (FSR=H), while setting a probability of 1.0 to (PED >5.60) resulted in 0.8999 as a posterior probability of an (H) outcome, indicating that larger numbers of PED increase the probability of having an (H) outcome in a crash. Both ROR and COL had lower effects on FSR, where higher number of ROR (ROR >2.43) resulted in 0.6103 probability of having an (H) outcome. On the other hand, lower number of COL (COL ≤41.45) resulted in 0.5693 probability of having an (H) outcome in a crash.

Table 10. Partial evidences in model 3

Variable	Evidence given to categories	Probability of FSR=H
COL	P (COL ≤41.45) = 1.0	**0.5693**
	P (COL = (41.45-83.67]) = 1.0	0.2413
	P (COL = (83.67-106.27]) = 1.0	0.3830
	P (COL >106.27) = 1.0	0.2399
PED	P (PED ≤4.07) = 1.0	0.0833
	P (PED = (4.07-4.85]) = 1.0	0.5000
	P (PED = (4.85-5.60]) = 1.0	**0.9166**
	P (PED >5.60) = 1.0	**0.8999**
ROR	P (ROR ≤1.78) = 1.0	0.1568
	P (ROR = (1.78-1.99]) = 1.0	0.4357
	P (ROR = (1.99-2.43]) = 1.0	0.2841
	P (ROR >2.43) = 1.0	**0.6103**

DISCUSSION AND CONCLUSIONS

This research work aimed at analyzing traffic crashes using annual crash reports' aggregated data by developing 3 different models to characterize traffic crashes according to severity. All the data used in the research was obtained for roads in Jordan. It should be stressed out that the analysis carried out herein is not based on disaggregate crash records, but on aggregated annual data. Thus, the results can be used only to describe the general trend of annual crashes committed and their severity as described by FSR.

When developing a model to analyze the factors that characterize crashes as being more prone to cause fatality or highly severe injury using number of violations committed, it was found that the only single violation that when committed in large numbers indicates larger likelihood of severe crashes is disobeying traffic control devices (TCD), where more TCD violations resulted in a probability of 0.5688 of characterizing crashes as severe or fatal, as illustrated in Table 8. These results indicate that committing traffic violations increases the likelihood of being involved in a severe or a fatal crash, which is in accordance with other studies found in literature (Valent et al., 2002; Yau, 2004; Yau et al., 2006; Kim et al., 2008).

The second violation which was found to contribute to the occurrence of a fatality or highly severe injury was speeding, which is consistent with Zhang et al. (2013) and Al-Masaeid (2009). This, however, indicates that there are other violations, such as TCD that contribute to higher numbers of severe crashes than those attributed to speeding.

The second developed model is used to analyze the

relationship between FSR and socio-economic and demographic characteristics. As illustrated in Table 9, the variables that were found to be most significant in characterizing crashes as severe or fatal were (in order of importance): lower number of foreigners (FA), less population (POP), lower number of registered vehicles (RV) and less gross domestic product (GDP), indicating that the effect of these variables on severe or fatal crashes is minimal.

The last developed model is used to analyze the crash type and its relationship with FSR. The results, as described in Table 10, indicate that the most significant variable was higher number of pedestrian crashes (PED) involved in crashes, which increases the probability of characterizing crashes as severe or fatal with probabilities larger than 0.890. The second significant variable found to increase the probability of severe or fatal crashes with a probability of 0.6103 was run-off-road (ROR). These results are in accordance with Fréchède et al. (2011), who found that ROR crashes accounted for 35% of all fatalities in single-vehicle crashes. Also, the largest group of road user fatalities worldwide is pedestrians being hit by motorized vehicles (Peden et al., 2004; Naci et al., 2009; Rosén et al., 2011).

To this end, the results found indicate that the most significant variables that are associated with higher number of severe or fatal crashes are: TCD, speeding, ROR and PED.

The results imply a need of comprehensive and in-depth investigation of driver behavior as proved by increased severity with increased number of TCD and speeding violations. Also, ROR and PED crashes are attributed to more than one factor, such as inadequate roadway design, mechanical problems, environmental conditions and driver behavior. Thus, there is a need to investigate thoroughly possible causes of these types of crash in order to effectively take actions to reduce their number and consequently reduce the severity of traffic crashes.

In addition, it was proved that BNs can effectively be used to analyze relationships between variables in smaller datasets without any prior assumptions regarding the distribution that the sample of data follows. Nonetheless, the accuracy of the results was not affected, which can be of a great help for governments that lack reliable data or do not have traffic crash records on a microscopic level. These governments can use the aggregated annual traffic crash data to develop models using BNs that are capable of finding dependencies between variables and accurately deciding which of these variables significantly affect the outcome of a traffic crash. It should also be mentioned that BNs were capable of analyzing behavioral drivers' aspects related to reckless driving and thus they can be used to calculate the percentage of contribution of human behavior to the occurrence of a specific outcome in a traffic crash.

REFERENCES

Acid, S., de Campos, L.M., Fernández-Lunam, J.M., Rodríguez, S., and Salcedo, J.L. (2004). "A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service". Artificial Intelligence in Medicine, 30, 215-232.

Al-Masaeid, H., and Nelson, D.C. (1996). "Pedestrian accidents and their relationship to street geometrics and operation variables in Jordan". Journal of Indian Highways, 24 (9), 49-57.

Al-Masaeid, H., Obaidat, M.T., and Gharaybeh, F.A. (1997). "Pedestrian accidents along urban arterial midblocks". Journal of Traffic Medicine, IAATM, 25 (3-4), 65-70.

Al-Masaeid, H. (2009). "Traffic accidents in Jordan". Jordan Journal of Civil Engineering, 3 (4), 331-343.

Al-Masaeid, H. (2016). "Safety impact of enforcement, penalty level and fuel prices". International Journal of Constructive Research in Civil Engineering, 2 (2), 27-31.

Al-Omari, B., Ghuzlan, K., and Hasan, H. (2013). "Traffic accidents and charactersitics in Jordan". International Journal of Civil and Environmental Engineering, 13 (5), 9-16.

Al-Suleiman, T.I., and Al-Masaeid, H. (1992). "Descriptive model for fatality rates of traffic accidents in Jordan". Institute of Transportation Engineers Journal, 62 (4), 37-39.

Bouckaert, R.R. (1995). "Bayesian belief networks: from construction to inference". Ph.D. Thesis, University of Utrecht.

Bouckaert, R.R. (2004). "Bayesian networks in Weka". Technical Report 14/2004. Computer Science Department, University of Waikato.

Chang, L.Y., and Wang, H.W. (2006). "Analysis of traffic injury severity: an application of non-parametric classification tree techniques". Accident Analysis and Prevention, 38, 1019-1027.

Cheng, J., and Greiner, R. (1999). "Comparing Bayesian network classifiers". In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers, Inc., San Francisco, 101-108.

Chow, C.K., and Liu, C.N. (1968). "Approximating discrete probability distributions with dependence trees". IEEE Transactions on Information Theory, 14 (3), 426-467.

Darwiche, A. (2008). "Bayesian networks". In: Harmelen, F.V., Lifschitz, V. and Porter, B. (Eds.), Handbook of Knowledge Representation (467-509). Amesterdam, The Netherlands: Elsevier Science.

Department of Statistics. (2016). "Jordan in numbers". [Online]. Available at: http://dos.gov.jo/dos_home_a/jorfig/2015/8.pdf

Fréchède, B., McIntosh, A.S., Grzebieta, R., and Bambach, M.R. (2011). "Characteristics of single vehicle rollover fatalities in three Australian states (2000-2007)". Accident Analysis and Prevention, 43, 804-812.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). "Bayesian network classifiers". Machine Learning, 29, 131-163.

Kim, K., Brunner, I.M., and Yamashita, E. (2008). "Modeling fault among accident-involved pedestrians and motorists in Hawaii". Accident Analysis and Prevention, 40, 2043-2049.

Mujalli, R.O., López, G., and Garach, L. (2016). "Bayes classifiers for imbalanced traffic accident datasets". Accident Analysis and Prevention, 88, 37-51.

Naci, H., Chisholm, D., and Baker, T.D. (2009). "Distribution of road traffic deaths by road user group: a global comparison". Injury Prevention, 15, 55-59.

Neapolitan, R.E. (2009). "Probabilistic methods for bioinformatics". Morgan-Kaufmann Publishers, San Francisco, CA.

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hydar, A.A., Jarawan, E., and Colin, M. (2004). "World report on road traffic injury prevention". Available at: http://apps.who.int/iris/bitstream/10665/42871/1/9241562609.pdf. Accessed in February, 2016.

Police Security Directorate. (2014). "Traffic crash study for 2014". [Online] Available at: https://www.psd.gov.jo/images/traffic/docs/derasah2014.pdf

Rosén, E., Stigson, H., and Sander, U. (2011). "Literature review of pedestrian fatality risk as a function of car impact speed". Accident Analysis and Prevention, 43 (1), 25-33.

Tao, L., Zhao, C., Thulasiraman, K., and Swamy, M. (1992). "Simulated annealing and tabu search algorithms for multi-way graph partition". Journal of Circuits, Systems and Computers, 2 (2), 159-185.

Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S., and Barbone, F. (2002). "Risk factors for fatal road traffic accidents in Udine, Italy". Accident Analysis and Prevention, 34, 71-84.

WHO. (2015). "Global statuts report on road safety 2015". Geneva: World Health Organization.

World Bank. (2016). "World bank country and lending groups". [Online]. Available at: https://datahelpdesk.worldbank.org/ knowledgebase/articles/906519-world-bank-country-and-lending-groups

Yang, Y., and Webb, G. (2001). "Proportional k-Interval discretization for naive-Bayes classifiers". In: Proceeding of the 12th European Conf. on Machine Learning, Freiburg, Springer, 564-575.

Yau, K.K.W. (2004). "Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong". Accident Analysis and Prevention, 36, 333-340.

Yau, K.K.W., Lo, H.P., and Fung, S.H.H. (2006). "Multiple-vehicle traffic accidents in Hong Kong". Accident Analysis and Prevention, 38, 1157-1161.

Zhang, G., Yau, K.K.W., and Chen, G. (2013). "Risk factors associated with traffic violations and accident severity in China". Accident Analysis and Prevention, http://dx.doi.org/10.1016/j.aap.2013.05.004