Industrial Engineering and Innovation Management (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/ieim.2022.050705
ISSN 2522-6924 Vol. 5 Num. 7

Research on the Causes of Road Traffic Accidents Based on Cause and Effect Diagrams

Zixuan Li[1], Xiangteng Ma[2]

[1]*Transportation Institute of Inner Mongolia University, Hohhot, 010000, China*
[2]*School of Mechanical and Automotive Engineering, Xiamen University of Technology, Xiamen, 361024, China*

Keywords: traffic accidents, Bayesian network, cause and effect diagram

Abstract: To reduce the probability of traffic accidents and enhance the safety of traffic travel, this study aims to analyze the strength of the causal effect of the causes of traffic accidents. This study collected some road and social attribute feature data in California, used the Bayesian network to build a dependency graph, used the causal science to evaluate the causal effect between each attribute, then marked the weak correlation between attributes and ranked the importance. The final conclusion: Based on this data, the total local population has the strongest causal effect on causing traffic accidents, the Spanish account for the largest number of the local population, and the Principal Arterial added under MAP-21 has an impact on the average annual daily traffic (AADT) and annual average daily traffic of vehicles with 2 axles and 6 wheels and above (AADTT).

1. Introduction

Traffic accidents, as a key social issue of global concern, have received significant attention from scholars in transportation worldwide. Both domestic and foreign scholars have analyzed and explored the causes of traffic accidents in different regions in different directions and dimensions. Some domestic scholars have studied and explored the differences in road attributes and social attributes that may lead to traffic accidents [1] and found that some inherent attributes of local roads and social factors such as local economic conditions have a certain relationship to the occurrence of traffic accidents; other scholars have applied Bayesian networks and fault tree models to study and summarize the causes of traffic accidents [2]. For issues such as the driver's speeding behavior during driving and the protective measures of the motor vehicle, research is conducted to find that there is a causal relationship with the occurrence of traffic accidents; Pei Yulong and Ma Ji studied the inherent properties of roads, proposed that high roadbeds and steep slope roadbeds can threaten traffic safety, and gave certain accident prevention countermeasures [3]. Sun Ping, Song Rui, and Wang Haixia used the hierarchical analysis method to comprehensively consider the people, vehicles, roads, and environment on the influence of traffic accidents occur, objectively analyze the causes of traffic accidents, and give preventive countermeasures [4]. Some foreign scholars have studied traffic accidents that cause traumatic brain injury, analyzed them in the direction of traffic accident causes, and confirmed that increased helmet use helps reduce the risk of traumatic brain injury caused by traffic accidents [5]. Keiichi Yamad and other scholars discussed human error, a key factor in traffic

accidents, including "distraction" in human factors, and found that the driver's reaction time is indeed affected by the human factor after an unexpected situation, and proposed the effectiveness of warning systems to prevent distractions [6]. S. AlKheder et al. conducted a study on the effect of highway design on fatal and non-fatal traffic accidents and found that the number of lanes has the highest sensitivity index in this system and has the greatest effect on accident frequency and road safety after modifying the number of lanes on the highway [7].

In this study, based on previous research inspired by the collection of highway characteristics and some social attributes in California, USA, a graphical model built using Bayesian Networks was used to analyze the causes of traffic accidents using a causal science approach. It is hoped to provide the transportation sector, including universities, design institutes, traffic engineers, and planners, with a distribution of causes and spatial inequalities associated with traffic accidents to help them better mitigate traffic accidents and reduce the occurrence of accidents. In this paper, we use the road traffic accident data of some regions in the United States to pre-process and integrate different attributes of the data, determine the strength of the "cause" and "effect" relationships, and put the integrated attributes into a Naive Bayesian Network for analysis. The hierarchical relationship of each node is obtained. As shown in figure 1.

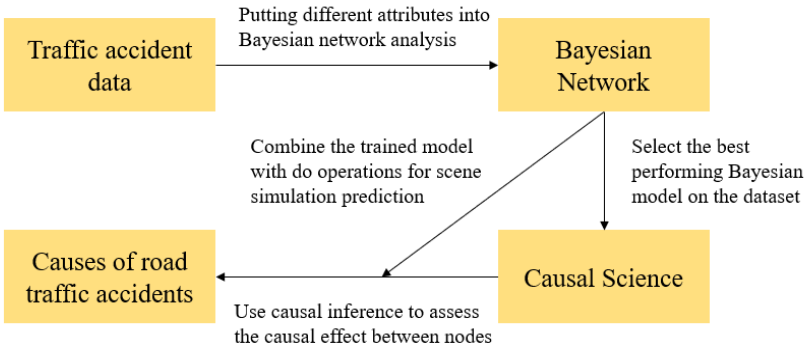


Figure 1: Flow chart of research theory

2. Analysis of the causes of traffic accidents

2.1. Bayesian method

Bayesian nets portray the dependencies between attributes with the help of directed acyclic graphs and use conditional probability tables to describe the association of attributes with probabilities. The Bayesian net structure effectively expresses conditional independence between attributes, given the set of parent nodes, and the Bayesian net assumes that each attribute is independent of its non-descendant attributes. Thus $B=<G,\theta>$ define the joint probability density distribution of the attributes $x_1,x_2,\ldots,x_d$ as:

$$P_B(x_1,x_2,\ldots,x_d)=\prod_{i=1}^{d}P_B(x_1|\pi_i)=\prod_{i=1}^{d}\theta_{x_1|\pi_i} \quad (1)$$

Typical dependencies between three variables in a Bayesian net: same-parent structure: given the value of the parent node x_1, then x_3 x_4 and are independent of each other. V-structure: also called impulse structure, given values of x_4, x_1 and x_2 are not independent; If x_4 is unknown, x_1 and x_2 are independent of each other. Sequential structure: given the value of x, y and z are conditionally independent. As shown in figure 2.

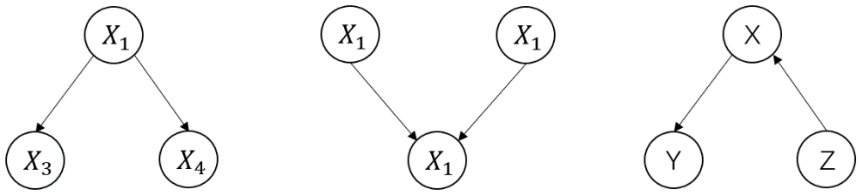


Figure 2: Typical dependencies between three variables in Bayesian net

The Naive Bayesian model belongs to the supervised learning algorithm, which is a model specifically used to solve classification problems. The prior probability of a sample is calculated from a training dataset with known categories, and then the posterior probability of an unknown category sample belonging to a category is measured using Bayesian formula, and finally, the category sample corresponding to the maximum posterior probability is used as the predicted value of the sample. By means of a plain Bayesian classifier, the road traffic accident data in the United States are trained, and since the independent variables in the dataset are all continuous-type values, the independent variables are assumed to obey a Gaussian distribution in the calculation of $P(X \mid C_i)$

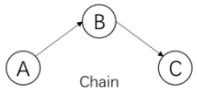
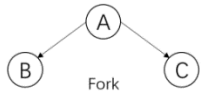
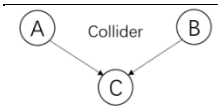
$$P(X|C_i) = \frac{1}{\sqrt{2\pi}\,\sigma_{ji}} \exp\left(-\frac{(x_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right) \tag{2}$$

Where x_j is the value of the first independent variable, μ_{ji} is the mean of the independent variable x_j belonging to a category C_i in the training data set, and σ_{ji} is the standard deviation of the independent variable x_j belonging to the category C_i in the training data set.

2.2. Causal Inference Method

In causal inference, there will be different ways of connection pointing between each different node, among which the most common and important three graph structures are chain structure, cross structure, and collision structure, in which when two events (nodes) do not exist a path makes the two events have a certain causal relationship, then become these two events are called D-Separated (Directed- Separated). On the contrary, if there is a path that makes the two events causally related, then it is called D-Connected. as shown in Table 1.

Table 1: Summary of the structure of the three cause-effect diagrams

Structure type	Two-end node relationship	Relationship between two end nodes given intermediate variables
	D-Connected Between A and C	Given the value of node B, D-Separated between A and C
	D-Connected Between B and C	Given the value of node A, D-Separated between B and C
	D-Separated Between A and B	Given the value of node C, D-Connected between A and B

In practical applications, there may be unobserved paths between two events to interact with each other in addition to the observed paths. It is also not possible to eliminate the interference of the interacting paths between the studied variables. So, at this point human intervention (also called calculation in the model) is necessary. When intervening in a variable in a model, i.e., fixing the value

of this variable, the values of the other variables are changed accordingly.

There is a fundamental difference between interventions and conditioning on variables, which are not equivalent cases. When conditioning a variable, only the overall situation of this variable at a certain value is observed, which is part of all situations. An intervention, on the other hand, observes the overall situation by fixing the value of the variable to a certain value in all cases. Thus, conditioning on a variable is changing the researcher's perspective of observing the overall event, and intervention is artificially changing the overall event. When intervening on a variable, i.e., fixing the value of this variable, it restricts other causal relationships that affect the outcome for this variable, and all arrows pointing to this variable should be deleted in the graphical model. This is shown in Figure 3.

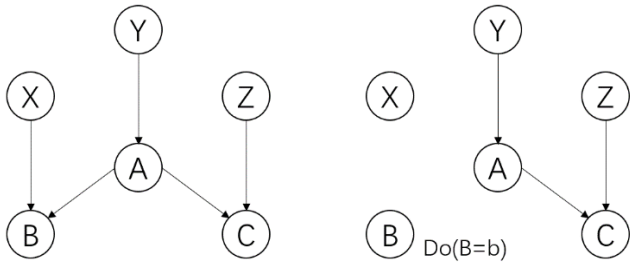


Figure 3: Graph model changes in the intervention case

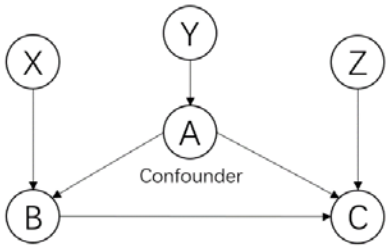


Figure 4: Obfuscated variables diagram

The applied interventions are generally in cross-structures with confounding variables. As in Figure 4, since we want to explore the causal relationship between B and C, but due to the existence of the cross-structural path of B ← A → C, the confounder directly explores the relationship between B and C. The A variable is the confounder variable. At this point, we need to block the cross-structural path of B ← A → C. We can intervene artificially on the B variable, that is, do(B = b), to block all arrows pointing to B, in order to explore the existence and importance of this path of B → C.

In research, it is common to use the difference between probabilities to characterize the strength of the causal effect.

$$\sigma = P\big(A = a\big|do(B = high)\big) - P\big(A = a\big|do(B = low)\big) \tag{3}$$

The difference between the probability of the A feature at a certain value when the human intervention B features are all values of HIGH characteristic and the probability of the A feature at a certain value when the human intervention B features are all values of LOW is indicated by the larger difference indicating that the causal effect of the two characteristic values of the B feature is large enough for the A feature when taken to a. The smaller the difference the opposite.

For the degree of influence of the secondary factors on the important factors, the influence size is compared using the weight assessment method. Where ω_i denotes the importance of the ith secondary cause on its major factor and β_i denotes the sum of the number of the ith secondary cause in all the data.

$$\omega_i = \frac{\beta_i}{\sum_i \beta} \tag{4}$$

3. Analysis of results

3.1. Case Overview

The data used in this study describe the highway characteristics and local social attribute characteristics of each roadway in California, USA, during the selected period, including the annual average daily traffic volume, total number of people, employment, and number of accidents on the corresponding roadway in the collection area. The data attributes that were integrated and utilized after the data were collected for the study are shown in the Table 2.

Table 2: Traffic data table for integrated use

Data attribute description	Symbols	Unit	Data range description (min/max)
Annual average daily traffic volume	AADT	Vehicle	0/15154457
Annual average daily traffic volume of vehicles with 2 axles and 6 wheels and above	AADTT	Vehicle	0/2950407
Minimum grade national highway	L	Mile	0/305.224
Interstates	S	Mile	0/197.024
Strategic Highway Corridor Network of Non-Interstate Highways	F	Mile	0/68.638
Strategic Highway Corridor Network Interchange Section	N	Mile	0/33.575
Other National Highways	O	Mile	0/119.302
Approved multimodal interchange sections	P	Mile	0/5.532
Major trunk lines added under MAP-21	M	Mile	0/50.342
Total number of regions	T	People	0/39454
Spanish	X	People	0/14613
White race	W	People	0/23885
Black race	B	People	0/5808
Asian	A	People	0/12174
Total local employment	E	People	0/129745
Industrial employment	G	People	0/24052
Service industry employment	R	People	0/7273
Transportation employment	TS	People	0/30765
Number of accidents	AC	Times	0/24
Total number of deaths	D	People	0/6
Total number of injuries	J	People	0/39
Number of seriously injured	BJ	People	0/9
Number of traumatic injuries	VJ	People	0/10

3.2. Building a Bayesian Network

Based on the data collected in this paper, studying the Causal Effect of the AADT, the total number of people, the number of employees, and other attributions on the corresponding road section and traffic accidents, a Bayesian Network is established to describe the dependencies between various attributes, and the integrated processing will be done. Different attributes are put into the Bayesian Network for analysis, and the hierarchical relationship of each node is obtained, as shown in Figure 5.

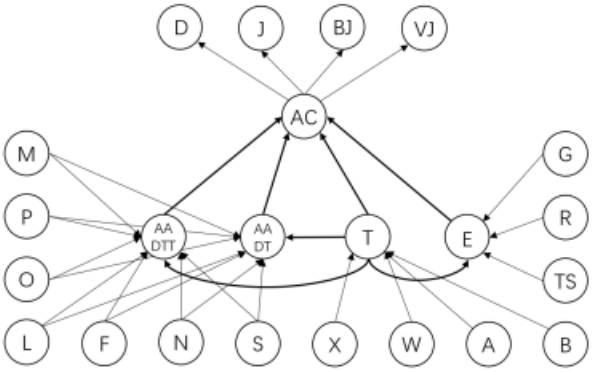


Figure 5: Bayesian Network for Road Traffic Accidents

3.3. Naive Bayes Model

By constructing a Gaussian Naive Bayes classifier, the prediction on the test data set is realized. According to statistics, there are a total of 1623 samples that predict traffic accidents as negative cases, and a total of 186 samples are predicted as positive cases. To test the prediction effect of the model on the test data set, it is necessary to construct a confusion matrix and draw a ROC curve. As shown in Figure 6, the confusion matrix is visualized, in which the value of the main diagonal represents the number of correctly predicted samples, and the remaining 478 samples are wrongly predicted samples. After calculating the confusion matrix, we obtained that the overall prediction accuracy of the model is 73%. As shown in the ROC curve shown in Figure 7, the calculated AUC value is 0.94.

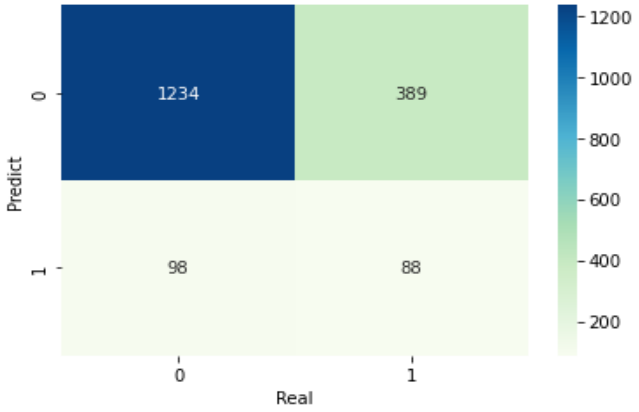


Figure 6: Visualization of Confusion Matrix

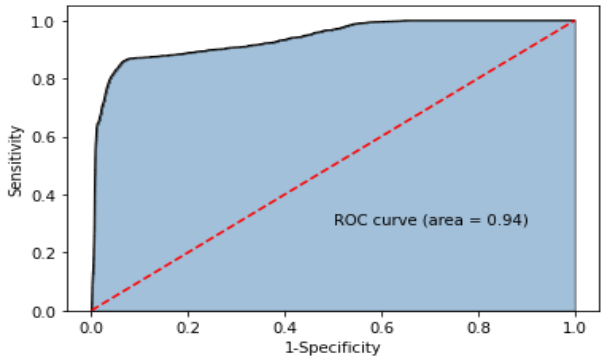


Figure 7: ROC curve of Gaussian Bayes classifier

3.4. Assessing Attribute Node Causal Effects

In the obtained Bayesian Network, the attribute nodes involved can be roughly divided into secondary causes, main causes, and after-effects of causing traffic accidents. This paper mainly studies the cause analysis of traffic accidents, so it mainly evaluates the strength of the causal effect of the main cause on the traffic accident and the influence of the secondary cause on the main cause.

3.4.1. Assessing the Strength of the Causal Effect of the Main Cause

The main factors that cause traffic accidents include the total number of local people, AADT, AADTT, and the proportion of employed persons. This study firstly uses the 'do calculation' in causal inference to quantify the impact of the total local population on traffic accidents, so uses the equation (9) to evaluate the probability difference. For continuous variables of attributes such as the total local population and AADT, the high and low differences with the same proportion should be used for the 'do calculation', and the two groups of values before and after the numerical quartile should be used as the high and low values.

Without processing other eigenvalues, artificially revise the local total number, change all the local total number to the first and third quartiles, obtain two sets of data, and use Gaussian Naive Bayesian Classifier to fit and classifies these two sets of data to predict whether a traffic accident occurs or not. When the total local population is in the first quartile, it is predicted that 851 of the 7999 locations will have traffic accidents, then the probability formula (3) can be used to calculate $P(AC = 1|do(total = \alpha_{75\%})) = 10.6388\%$. With the same method, we can obtain $P(AC = 1|do(total = \alpha_{25\%})) = 11.1139\%$ and get the difference $\sigma = 0.4751\%$.

Since traffic accidents are not events with a high probability in daily traffic travel, the three sets of values calculated above are all acceptable. Using the same method, the improvement values of the other main reasons that can affect the probability of traffic accidents at the 25% and 75% quartiles can be obtained, as shown in Table 3.

Table 3: Influence of the Main Reasons on Traffic Accidents under 'do calculation

Feature attribute	When $\alpha_{25\%}$, the number of predicted locations	When $\alpha_{75\%}$, the number of predicted locations	The difference in the probability of affecting the occurrence of traffic accidents σ
Total local population	851	889	0.4751%
AADT	604	633	0.3625%
AADTT	620	628	0.1000%
Employment rate	978	978	0%

We can easily know that $\sigma_{Total} > \sigma_{AADT} > \sigma_{AADTT} > \sigma_{Employment} = 0\%$. Therefore, the total local population has the strongest causal effect on causing traffic accidents. And it is found that the impact probability of employment rate on the occurrence of traffic accidents is 0, indicating that the local employment rate does not affect the occurrence of traffic accidents under this data.

Looking back at the Bayesian Network diagram, the simple relationship diagram can be obtained by extracting the relationship between the main reason and the traffic accident, and the main reason and the main reason, as shown in Figure 8.

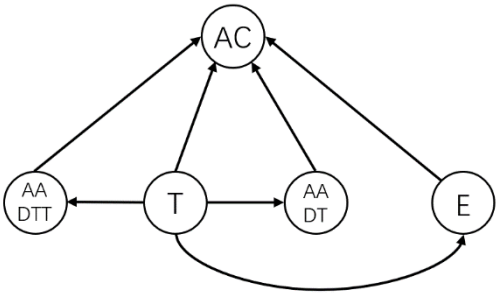


Figure 8: Extraction of the Relationship between the Main Reasons and Traffic Accidents

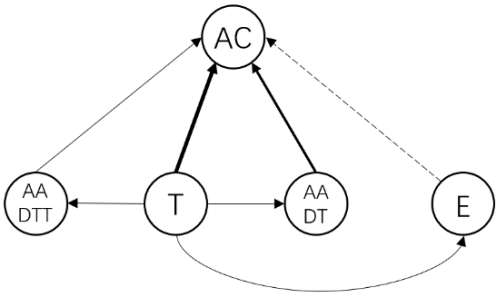


Figure 9: Extraction of the Relationship after Improvement

Since the total local population has an impact on AADT, AADTT, and the employment rate, and among the four, the total local population has the strongest causal effect on traffic accidents. Therefore, it can be inferred that the total local population has little effect on the other three characteristics, which is a weak causal effect, and the employment rate has no effect on traffic accidents, so it can be removed. The above figure can be changed to Figure 9, where the thickness of the solid line indicates the relative strength of the causal effect, and the dotted line indicates that there is no causal effect but only statistical correlation.

3.4.2. Assessing the Magnitude of the Impact of the Secondary Cause on the Main Cause

When assessing the influence of secondary causes on the main cause, since the employment rate does not affect the occurrence of traffic accidents, the secondary factors affecting the number of employed persons are not considered when studying the secondary causes. Therefore, we mainly consider the structure that affects the total local population (different races) and the structure of highways that affect AADT and AADTT (including Not_on_NHS, Interstate, Non-Interstate STRAHNET, STRAHNET Connector, Other_NHS, Approved Intermodal Connector, Principal Arterial added under MAP-21). Respectively, the magnitude of the impact on their corresponding main causes.

Since the essential reasons for the influence of the above two groups of structures are not the same, but there is no complex structure and there is no correlation between the attributes, they are both independent, so use the sum weight method to evaluate the influence, and use the formula (4) to evaluate the main cause influence assessment. Assessing the secondary causes of the "local total population" feature, the population structure table can be obtained, as shown in the Table 4:

Table 4: The Influence of the Secondary Cause on the Main Factor in the Population Structure

Population structure	Spanish	White	Black	Asian	Total
counts	4587274	3803556	654850	1319033	10744446
proportion	42.6944%	35.4002%	6.0948%	12.2764%	100%

From the observation of the population structure, we can know that in this region, the proportion

of Spanish is more than that of White, Asians, and Black. Using the same method, the highway structure table can be obtained as shown in the Table 5:

Table 5: Influence of Secondary Causes on Main Causes of Highway Structure

Highway Structure	Not_on_NHS	Interstate	Non-Interstate STRAHNET	STRAHNET Connector
Length	2773.934	574.061	357.911	38.415
proportion	29.7952%	6.1661%	3.8444%	0.4126%
Highway Structure	Other_NHS	Approved Intermodal Connector	Principal Arterial added under MAP-21	Total
Length	943.482	55.439	4567.21	9310.45
proportion	10.1341%	0.5955%	49.0570%	100%

Looking at the highway structure, it can be seen that Principal Arterial added under MAP-21 in this area structure account for the main part of all highway structures, that is, has the greatest influence on the AADT and AADTT.

4. Conclusion

The team used California road attributes and local social attribute data, based on Bayesian Network, and used causal inference to evaluate the causal effect of each attribute on causing traffic accidents. According to the evaluation of the causal effect of major factors on the causal effect of traffic accidents and the influence of secondary factors on the main cause, the following causal effect evaluation results can be obtained: (1) The total local population has the strongest causal effect on causing traffic accidents; (2) The local employment rate has no effect on the occurrence of traffic accidents; (3) Spanish accounted for the largest proportion of the total local population, and had the strongest impact on it; (4) The Principal Arterial added under MAP-21 in the impact on AADT and AADTT The largest proportion and the strongest impact on it.

References

[1] Quan Yuan, Jueyu Wang, Goods movement, road safety, and spatial inequity: Evaluating freight-related crashes in low-income or minority neighborhoods, Journal of Transport Geography, Volume 96,2021,103186, ISSN 0966-6923

[2] Luo X, Sun Shimei, Chen Kun, Fu Gui. Behavioral cause analysis of highway traffic accidents based on Bayesian network [J]. Journal of Jilin University of Architecture, 2022, 39(02): 49-53.

[3] Ei Yulong, Ma Ji. Analysis of the causes of road traffic accident road conditions and prevention countermeasures [J]. Chinese Journal of Highways, 2003(04): 78-83. DOI: 10.19721/j.cnki.1001-7372.2003.04.017.

[4] Sun P, Song R, Wang HX. Analysis of the causes of road traffic accidents and preventive measures in China [J]. Safety and Environmental Engineering, 2007(02): 97-100

[5] Sumanth P. Reddy, Maura S. Walsh, Robert Paulino-Ramirez, Jomar Florenzán, Jaime Fernández, Fiemu E. Nwariaku, Abier Abdelnaby, Neurologic injuries following road traffic accidents in the Dominican Republic: Examining causes and potential solutions, Traffic Injury Prevention, Volume 20, Issue 7, 2019, Pages 690-695, ISSN 1538-9588

[6] Keiichi Yamada, Yumie Minakami, Keisuke Suzuki, Analytical Study of Human Errors causing Traffic Accidents from the viewpoint of Consciousness Transition, IFAC Proceedings Volumes, Volume 41, Issue 2,2008, Pages 8526-8531, ISSN 1474-6670, ISBN 9783902661005

[7] S. AlKheder, H. Al Gharabally, S. Al Mutairi, R. Al Mansour, An Impact study of highway design on casualty and non-casualty traffic accidents, Injury, Volume 53, Issue 2,2022, Pages 463-474, ISSN 0020-1383.