How to cite this article:

Zamzuri, Z. H., Shabadin, A., & Ishak, S. Z. (2019). Bayesian network of traffic accidents in Malaysia. *Journal of Information and Communication Technology, 18*(4), 473-484.

BAYESIAN NETWORK OF TRAFFIC ACCIDENTS IN MALAYSIA

[1]Zamira Hasanah Zamzuri, [2]Akmalia Shabadin & [2]Siti Zaharah Ishak
[1]Faculty of Science and Technology,
Universiti Kebangsaan Malaysia, Malaysia.
[2]Malaysia Institute of Road Safety Research, Selangor, Malaysia.

zamira@ukm.edu.my, akmalia@miros.gov.my, sitizaharah@miros.gov.my

ABSTRACT

Exploring the causes and effects of a hazardous event such as traffic accidents have been of vital importance to society. Statistical analyses have been widely implemented to understand and deduce inferences on the cause-effect analysis, and to anticipate the occurrences of accidents in the future. One of the issues that has not been solved through conventional statistical modelling is the existence of interrelationships between variables in the data set. However, with the advent of technology and the wide application of machine learning algorithm, this problem can be solved through the application of Bayesian network analysis, which is a directed acyclic probabilistic graphical model. By using Hill Climb (HC) and Tabu algorithms, the structure of the data was studied and the relationship was estimated through conditional probability, that is based on the Bayes' theorem. The results suggests that weather plays a major role in the increase of traffic accidents, and occurs by disrupting lighting conditions which then disrupts the traffic systems. Furthermore, the results indicate that fatal accidents have a higher likelihood to occur in head-on, turn over and out of control accidents. The use of the Bayesian network creates probability estimates to enable the identification of the risk and the necessary precaution needed to be implemented.

Keywords*:* Bayesian network, HC algorithm, Tabu algorithm, traffic accidents.

Received: 31 Mei 2018 **Accepted**: 13 February 2019 **Published:** 26 September 2019

INTRODUCTION

Understanding the cause and effect of traffic accident occurrences is essential to enable policy decisions and safety actions to be implemented. Ongoing studies involving traffic accidents are observed worldwide from various issues and perspectives. The consequences of an accident can be related to financial loss, psychologically damage and worse, death. One of the important aspects of traffic accident analysis is the statistical analysis on predicting future occurrences and understanding the impact or influence of certain factors, particularly on the consequences of traffic accidents.

In the traffic accident statistical analysis, the main focus is to develop and implement the suggested model, and conduct assessments through the inferential and accuracy of the predictions produced. Generally, Poisson and negative binomial regression models were used to fit the frequency of traffic accidents as found in Maycock and Hall (1984) and Hauer et al. (1988). Over time, more complex models were proposed to handle various issues such as the heterogeneity through random effects models and spatial correlation, for example, in studies conducted by Guo (2010) and Plug, Xia and Caulfield (2011). Ulfarsson and Shankar (2003) and Quddus (2008) focused on the temporal correlation, meanwhile work on spatio-temporal correlation can be referred to Wang, Quddus and Ison (2011) and Castro, Paleti and Bhat (2011). Another growing interest in traffic accident models are the multivariate models as suggested in Hosseinpour et al. (2018) and Zamzuri (2018), and the model of extra zeros count by Zamri and Zamzuri (2017) and Zamzuri (2015).

Typically, the occurrence of traffic accidents is linked to many variables, such as traffic flow (Priambodo & Ahmad, 2018). The generalized linear model used is based on the relationship between the dependent variable (the traffic accident frequency) and the independent variables (weather, road condition, traffic flow), and is considered as many (independent variables) to one (dependent variable). This paper aims to find the relationships between all the variables observed, and to understand the sequential events that lead to a traffic accident.

Hence, this paper will examine the interrelationships between the variables that are involved in the traffic accident occurrence. The relationships are described through causal probability. For example, what would be the probability of A if its caused by B? This can be achieved by applying the Bayesian network analysis. The use of the Bayesian network is proven to be efficient in determining the relationship between variables (Ahmad-Azami et al., 2017). Past research that utilized the Bayesian network to traffic accidents data was conducted in Iran by Karimnezhad and Moradi (2014). Hongguo et al. (2010) performed

a similar analysis using K2 and junction tree algorithms. Moreover, similar works of different countries data sets were studied in Australia by Zou and Yue (2017) and Switzerland by Deublein et al. (2015). However, no previous work was found that applied the Bayesian network to Malaysian traffic data. The advantage of the Bayesian network is it produces comprehensive information that is obtained through the conditional probabilities of the interrelationships between the variables. Additionally, the proposed method reveals the structure and dependencies of the factors or variables involved in the traffic accidents occurrences. By obtaining valuable insights on the occurrences of traffic accidents through these interrelationships, traffic engineers and policy makers can plan and build an improved road system that minimizes the risk of traffic accidents.

METHODOLOGY

Bayesian Network

Bayesian network is a graphical probability model with directed acyclic that is based on the Bayes' theorem. In the Bayesian network, a node denotes each variable, and the relationship between a pair of nodes is represented by the edge. This relationship refers to the conditional probability of the event of interest from the given information. For example, let A be the event that the collision type is head-on and B is the event that the accident severity is fatal; then the conditional probability of B given A is as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (1)$$

This edge can be directional, and does not depend on how the relationship is defined. Typically, the variables in the Bayesian network are categorical, as the ones used in this paper.

A Bayesian network, $B = (G, X_V)$, is composed by a Directed Acyclic Graph (DAG), $G = (V, E)$ in which VV: vertices and EE:edges and a random vector $X_v = \{X_i\}_{i \in V}$ with a probability density factorizes according to the DAG as given by equation (2):

$$P(B|D) = P(G, \theta|D) \propto P(G|D)\, P(\theta|G, D) \quad (2)$$

where θ: parameters and D:data.

The main tasks of the Bayesian network are to infer unobserved variables, parameter learning and structure learning. This paper focuses on the first

and last tasks, which are inferring unobserved variables and identifying the likelihood of an accident to happen, given the current information. From the examination of the structure of dependencies between variables, the outcome could either be specified or structured by the field experts; or machine learning algorithms can be used when no information was provided from the experts. The objective of structure learning is to find the DAG that maximizes $P(G|D)$, whereby it can be rewritten as:

$$P(G|D) \propto P(G)P(D|G) \propto P(G) \int P(D|G,\theta)\, P(\theta|G) d\theta \quad (3)$$

In this paper, two machine-learning algorithms are employed: Hill Climb (HC) and Tabu. The Hill Climb algorithm is used for solving optimization problem as can be found in Shehab et al. (2018). Both of these algorithm are categorized as score-based, for which the Bayesian Information Criterion (BIC) is used as the score.

Hill Climb Algorithm

Hill Climb algorithm is a hybrid between constraint based and score methods that uses optimization to find a local search, as explained by Tsamardinos et al. (2006). This algorithm identifies the parent and the children for each variable. Then, the algorithm maximizes the minimum association between a variable and a target from the given candidate parents and children. The steps for the Hill Climb algorithm are:

Hill Climb Algorithm

Step 1:	Start with an initial graph structure, G
Step 2:	Perform the operation that gives an acyclic graph, G^* - add, delete or reverse an arc of G
Step 3:	Compute the score of the new graph Score (G^*)
Step 4:	If Score (G^*) > Score (G), set $G = G^*$and Score (G) = Score (G^*)
Step 5:	Repeat steps (2) to (4) as long as Score (G) increases

Haff et al. (2016) proposed that the algorithm is among the simplest method used and hence, its application have been widely popularized in artificial intelligence, and is used to find the goal state starting with the initial node. However, the main disadvantage of this algorithm is that it tends to lag in local optimum point. This weakness is overcome by the next algorithm, Tabu.

Tabu Algorithm

The Tabu algorithm has similar function and characteristics as Hill Climb, but with an added advantage. This algorithm utilizes memory that describes the visited solutions. Dunder et al. (2014) explained that this algorithm will add a short term memory for the links added between moves. There are two main conditions in this algorithm, which are the relaxing and prohibitive conditions. In relaxing condition, any move that does not improve the solution is accepted, given that there is no more improving moves available. The essence of this condition is to avoid the search to lag at a strict local minimum. Meanwhile, the purpose of a prohibitive condition is to prevent the search to revisit any previously visited solutions. Figure 1 summarizes the steps involved in this algorithm.

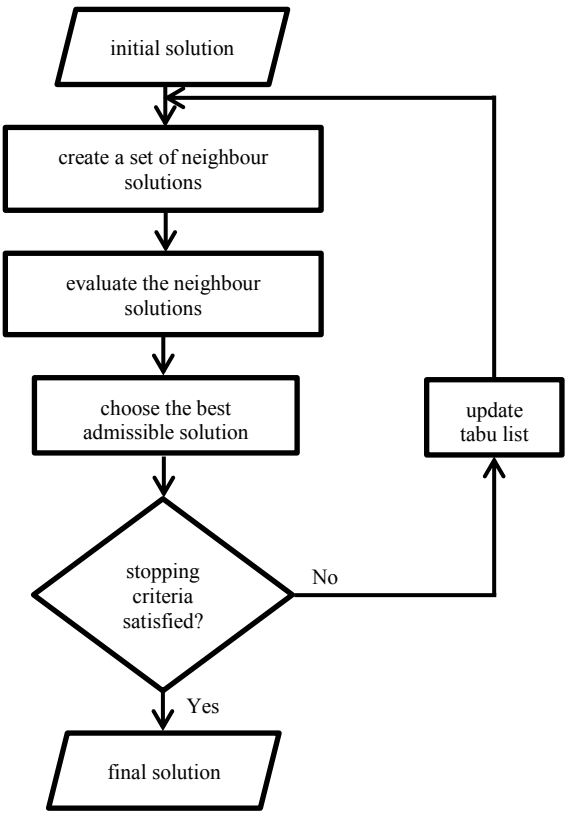

Figure 1. Flow chart of tabu algorithm.

RESULTS

The Data

The dataset used in this analysis is the collection of accident reports from police stations in Malaysia for the year 2014. Eight categorical variables are recorded on the details of the accidents. Table 1 lists the details of the eight variables.

Table 1

Description of the data

Variables	Levels
Day of Week	Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday
Light Condition	day, dawn/dusk, dark with street light, dark without street light
Weather	clear, windy, foggy, rain
Lane Marking	double, single, one way, divider, u-turn, no marking
Road Geometry	straight, bend, roundabout, cross junction, t/y junction, staggered junction, interchanges
Traffic System	one way, two way, three-lane, dual carriage away
Collision Type	head on, rear, right angle side, angular, sideswipe, forced, hitting animal, hitting object on/off road, hitting pedestrian, overturned, out of control, others
Accident Severity	fatal, serious, slight

The Network

By using Tabu and HC algorithms as described in the previous section, the network is constructed as observed in Figure 2. Human intervention is also permissible in the Bayesian network, in which a user can specify any

relationship to be included or dismissed from the network. Based on the relationships in Figure 2, the day of the week has no direct relationship with the other variables.

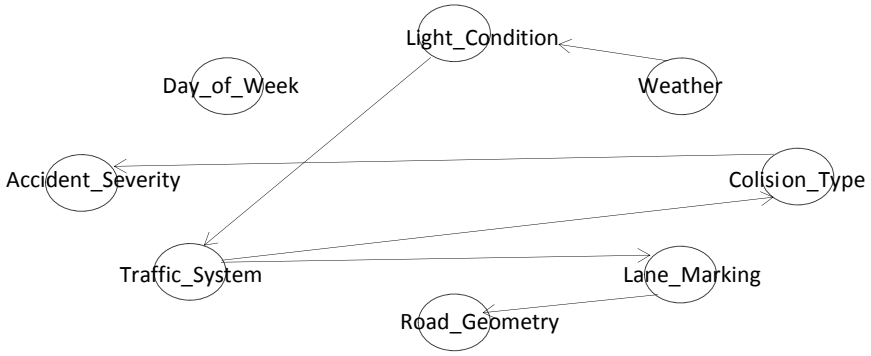

Figure 2. The bayesian network of Malaysian Traffic accident data constructed through HC and tabu algorithms.

The causal relationships that were obtained are:

1) Weather --> Light Condition --> Traffic System --> Collision Type --> Accident Severity
2) Traffic System --> Lane Marking --> Road Geometry

Out of the two sequences of causal relationships, the first relationship is more useful as it relates to the occurrence and severity of the accidents. Based on the network in Figure 2, the weather implies a change in the light condition, and the light conditions will then imply a change in the traffic system. Subsequently, the traffic system implies the collision type and the collision type will imply the severity of the accident. Based on the network developed from the data, it is shown that day of the week has no contribution to the occurrence of any event that is associated to a traffic accident.

From the network obtained, the interrelationships between the variables can be discussed. For instance, through the conditional probability table of weather-light conditions as shown in Table 2, it can be observed that the probability of light condition to be daylight in any given clear weather is 0.6369. Furthermore, the *P*(Day|any weather) has a high probability for which

most drivers are driving in daylight. For light condition in windy or dusk, the networks correlates these to windy and foggy weather.

Table 2

The conditional probabilities of light condition given the information on weather

Light condition/ Weather	Clear	Windy	Foggy	Rain
Day	0.6369	0.5294	0.4491	0.5274
Dawn/Dusk	0.0502	0.2353	0.2275	0.0645
Dark with street light	0.1893	0.1176	0.0898	0.2184
Dark w/o street light	0.1235	0.1176	0.2335	0.1894

Next, a conditional probability table of the accident severity based on the information of the type of collision is presented in Table 3.

Table 3

The conditional probabilities of the accident severity given the collision type

Collision type / Accident severity	Fatal	Serious	Slight	N
head-on	0.457	0.264	0.279	2172
rear	0.378	0.207	0.415	2166
right angle side	0.271	0.283	0.446	968
angular side	0.216	0.257	0.527	3791
sideswipe	0.247	0.164	0.589	1814
forced	0.432	0.068	0.500	44
hitting animal	0.391	0.187	0.422	225
hitting object on road	0.372	0.144	0.484	180
hitting object of road	0.516	0.125	0.359	153
hitting pedestrian	0.372	0.215	0.413	1239
overturned	0.490	0.121	0.389	108
out of control	0.569	0.139	0.292	2974
N	6186	3472	6986	16644

From Table 3, the results suggest that the top three types of collisions that commonly occurs are angular side, out of control and head-on. By focusing on the most severe outcomes, given that the accident is head on, the probability for a fatality is 0.457. Fatality outcomes are also high in the chance for out of control and overturned accidents. As for resulting to serious injury, the top three types of accidents are right angle side, head-on and angular side. The rear end accidents mostly end up with a slight injury consequence, given by the probability value of 0.415.

Query for Probability

Another interesting and informative output from the Bayesian network is the ability to compute the probability for an event to occur, from the given information. Based on the network constructed from Figure 2; Assume that the following information are provided:

1) Collision Type = Head-on (*HO*)
2) Weather = Rain (*R*)
3) Traffic System = Two way (*T*)
4) Light Condition = Day (*D*)

Given the above conditions, the probability of a fatal accident (Accident severity = Fatal (*F*)) to occur can be estimated computationally. Hence, the probability can be computed as:

$$\begin{aligned} P(F|HO,R,T,D) &\propto P(F,HO,R,T,D) \\ &\propto P(F|HO)P(H|T)P(T|D)P(D|R) \\ &= 0.1888 \end{aligned}$$

Therefore, based on the speculated conditions; weather, traffic system and light condition, the probability of a head-on accident to occur is around 19% (prob. = 0.1888).

CONCLUSIONS

This paper examined the application of the Bayesian network in traffic accident analysis, specifically in the context of Malaysia. One of the advantages of the Bayesian network is its ability to estimate the interrelationships that exist between the variables in the data set. The information obtained is essential

in the causal analysis to determine the following: which event causes the occurrence of a subsequent event; identify the event that occurred first; and which event has the highest chance to occur when given the perceived information. These information are obtained through the Bayesian network that makes this analysis effective, especially in deducing inferences on the unobserved variables.

Two algorithms were considered in this paper, which are HC and Tabu algorithms. Both algorithms produced the same conclusion on the network structure. Based on the Bayesian network analysis conducted on the data of reported accidents at Malaysian police stations in 2014, it is identified that the traffic system will imply the type of accidents to happen and subsequently imply the level of severity of that particular accident. In addition, the traffic system is influenced by the lighting condition, which in turn was influenced by the weather. Through the fitted network, the conditional probability can be estimated for the event of interest when the given information are available.

ACKNOWLEDGEMENT

This research was funded by a grant from Ministry of Higher Education of Malaysia (FRGS/1/2015/ST06/UKM/02/1).

REFERENCES

Ahmad-Azani, N. I., Yusoff, N., & Ku-Mahamud, K. R. (2018). Fuzz discretization technique for Bayesian flood disaster model. *Journal of Information and Communication Technology*, *18*(2), 167–189.

Castro, M., Paleti, R., & Bhat, CR. (2013). A spatial generalized order response model to examine highway crash injury severity. *Accident Analysis & Prevention, 52*, 188-203.

Deublein, M., Schubert, M., Adey B. T., & García de Soto, B. (2015). A Bayesian network model to predict accidents on Swiss highways. *Infrastructure Asset Management, 2*, 145-158.

Dunder, E., Cengiz, M. A., & Koc, H. (2014). Investigation on the impacts of constraint –based algorithms to the quality of Bayesian network structure in hybrid algorithms for medical studies. *Journal of Advanced Scientific Research, 5*, 8-12.

Guo, F., Wang, X., & Abdel-Aty, M. (2010). Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis & Prevention*, *42*(1), 84-92.

Haff, I. H., Aas, K., Frigessi, A., Laval, V. (2016). Structure learning in Bayesian network using regular vines. *Computational Statistics and Data Analysis, 101*,186-208.

Hauer, E., Ng, J. C. N., Lovell, J. (1988). Estimation of safety at signalized inter sections. *Transportation Res. Rec*, *1185*, 48-61.

Hosseinpour, M., Sahebi, S., Zamzuri, Z. H., Yahaya, A. & S., Ismail, N. (2018). Predicting crash frequency for multi vehicle collision types using multivariate poisson lognormal spatial model: A comparative analysis. *Accident Analaysis & Prevention, 118*, 277-288.

Hongguo, X., Huiyong, Z., & Fang, Z. (2010). Bayesian network-based road traffic accident causality analysis. In *Proceedings of 2010 WASE International Conference on Information Engineering ICIE*. 413-17.

Karimnezhad, A. & Moradi, F. Road. (2017*)*. Accident data analysis using Bayesian networks. *Transportation Letters*, *9*, 12-19.

Maycock, G., & Hall, R. D. (1984). *Accidents at 4-arm roundabouts*. Laboratory Report LR1120, Transport Research Laboratory, Crowthorne, Berks, UK.

Plug, C., Xia, J., & Caulfield, C. (2018). Spatial and temporal visualisation techniques for crash analysis. *Accident Analysis & Prevention, 43*, 1937-1946.

Priambodo, B., & Ahmad, A. (2018). Traffic flow prediction model based on neighbouring roads using neural network and multiple regression. *Journal of Information and Communication Technology*, *17*(4), 513-535.

Quddus M. A. (2008). Time series count data models: An empirical application to traffic accidents. *Accident Analysis & Prevention*, *40*, 1732-1741.

Shehab, M., Khader, A. T., & Laouchedi, M. (2018). A hybrid method based on cuckoo search algorithm for global optimization problems. *Journal of Information and Communication Technology*, *17*(3), 469-491.

Tsamardinos, I., Brown, L. E., & Aliferis, C.F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning, 65*, 31-78.

Ulfarsson, G. F., & Shankar, V. N. (2003) Accident count model based on multiyear cross-sectional roadway data with serial correlation. *Transportation Research Record, 1840*,193-197.

Wang, C., Quddus, M.A. & Ison, S.G. (2011). Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis & Prevention*, *43*, 1979-1990.

Zamzuri, Z. H. (2018). The spatio-temporal multivariate Poisson lognormal model in *Proceeding of the 25th National Symposium on Mathematical Sciences*. 020013.

Zamri, N. S. N., & Zamzuri, Z. H. (2017). A review on models for count data with extra zeros in *The 4th International Conference on Mathematical Sciences, ICMS*, 080010.

Zamzuri, Z. (2015). An alternative method for fitting a zero inflated negative binomial distribution. *Global Journal of Pure and Applied Mathematics, 11*, 2461-2467

Zou, X., & Yue, W. L. (2017). A Bayesian network approach to causation analysis of road accidents using netica. *Journal of advanced transportation*, *2017*, 1-18.