Decision Tree Model for Count Data

Yap Bee Wah, Norashikin Nasaruddin, Wong Shaw Voon and Mohamad Alias Lazim

***Abstract*— The Poisson Regression and Negative Binomial Regression models are the conventional statistical models for count data. This paper presents using decision tree to model motorcycle accident occurrences and compared its classification performance with Poisson Regression and Negative Binomial Regression model. The frequency of motorcycle accidents that involve death or serious injury based were converted into a categorical dependent variable (zero, low and high frequency) and the factors considered are collision types, road geometry, time, weather condition, road surface condition and type of days. Based on classification accuracy, results show that the decision tree model using CART (Classification and Regression Tree) slightly performs better (78.1%) than Poisson Regression (76.3%) with Negative Binomial Regression (77.6%) models. The CART decision rules revealed that the most significant factor contributing to high frequency of motorcycle accidents that result in death or serious injury is when the accidents happen on a straight road, junction or bend.**

Index Terms*— count data, decision tree, motorcycle accidents, Poisson regression, Negative Binomial regression

I. INTRODUCTION

Road accident occurrences are one of the major issues in the news. In 2002, it is reported that more than 1.2 million people died due to road traffic accidents and it is ranked as the eleventh top causes of death in the world [1]. Furthermore, in 2004 it is reported that traffic accidents is the top three leading causes of deaths for people aged between 5 and 44 years old [2]. According to MIROS (Malaysian Institute of Road Safety Research) the highest number of registered vehicles in Malaysia is motorcycles. For the year 2009, the total number of registered motorcycles in Malaysia was estimated at 8,940,230, where 113,962 involved in accidents and about 4070 deaths (include motorcyclists and pillion riders) [3]. It is much more dangerous to ride a motorcycle than to drive an automobile in terms of injury or death when an accident occurs.

Therefore, it is no surprise that the percentage of fatalities that involved motorcyclists and the pillion riders are high. MIROS also reported that in terms of fatalities by mode of transport, motorcycle fatalities ranked the highest as shown in Table 1, the motorcycle fatalities (58.21%), (59.72%) and (60.30%) for 2002, 2008 and 2009 respectively [4].

TABLE I
FATALITIES BY MODE OF TRANSPORT

Road User	2002	%	2008	%	2009	%	Previous year	% change over 2002
Pedestrian	650	11.03	598	9.16	589	8.73	-1.51	-9.38
Motorcycle	3429	58.21	3898	59.72	4067	60.30	4.34	18.61
Bicycle	261	4.43	203	3.11	224	3.32	10.34	-14.18
Car	1023	17.37	1335	20.45	1405	20.83	5.24	37.34
Van	156	2.65	96	1.47	91	1.35	-5.21	-41.67
Bus	45	0.76	48	0.75	31	0.46	-35.42	-31.11
Lorry	197	3.34	195	2.99	213	3.16	9.23	8.12
4Wheel	74	1.26	106	1.62	78	1.16	-26.42	5.41
Other	56	0.95	48	0.74	47	0.7	-2.08	-16.07
Total	5891	100	6527	100	6745	100	3.34	14.50

(Source: Road facts, Retrieved from: http://www.miros.gov.my/web/guest/road)

The general definition of the term road accident in Malaysia is "*An occurrence on the public or private roads due to the negligence or omission by any party concerned (on the aspect of road users conduct, maintenance of vehicle and road condition) or due to environmental factor (excluding natural disaster) resulting in a collision (including "out of control" cases and collisions or victims in a vehicle against object inside or outside the vehicle e.g. bus passenger) which involved at least a moving vehicle whereby damage or injury is caused to any person, property, vehicle, structure or animal and is recorded by the police*". According to the types of road accidents, the description of the term fatal road accident is "*road accidents in which one or more person were killed within 30 days from the date of event*", serious injury road accident is "*A road accident in which at least a person sustained serious injury but none killed*", minor injury road accident is "*A road accident in which one or more person were injured but not killed or seriously injured*" and non-injury road accident is "*A road accident in which no person was killed or injured*" [5], [6].

Most studies involving cross-sectional count data use the conventional Poisson Regression and Negative Binomial Regression models. This paper focus on using decision tree to model motorcycle accident frequency and compared it with the conventional Poisson Regression and Negative Binomial Regression model. This paper is organized as follows. Section 2 provides a review on previous studies on traffic accidents. The methodology and techniques are explained in Section 3 while the description of the data is given in Section 4. The results for decision trees, Poisson and Negative Binomial Regression model are then presented and compared in Section 5. Finally the conclusion and some recommendations for future work are given in Section 6.

Manuscript received Mar 21, 2012; revised Apr 12, 2012. This work was supported in part by the Malaysian Fundamental Research Grant, FRGS/2/2010/SG/UITM/02/34

B. W. Yap. Faculty of Computer and Mathematical Sciences, UiTM Shah Alam, 40450 Shah Alam, Selangor, Malaysia. (corresponding author: phone: +603-5543-5461; fax: +603-5543-5501; e-mail: beewah@tmsk.uitm.edu.my).

N. Nasaruddin. Faculty of Computer and Mathematical Sciences, UiTM Shah Alam, 40450 Shah Alam, Selangor, Malaysia. (e-mail: see_keen86@yahoo.com).

S. V. Wong. Department of Mechanical and Manufacturing Engineering, University Putra Malaysia, 43400 UPM, Serdang, Malaysia. (e-mail: wongsv@eng.upm.edu.my).

M. A. Lazim. Faculty of Computer and Mathematical Sciences, UiTM Shah Alam, 40450 Shah Alam, Selangor, Malaysia. (e-mail: dralias@tmsk.uitm.edu.my).

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. LITERATURE REVIEW

Previously, numerous studies have addressed various aspects of traffic accidents, including accidents involving motorcycles. These global or national studies focus mainly on predicting the critical factors that influence the occurrence of accidents. However, the results of each study vary depending on the model and the variables considered in the study.

Since road accidents data are count data, most of the previous studies focus on studying the effect of some variables that influence accidents occurrence. For instance, [7] studied on the effect of the "daytime running headlight" intervention (RHL) to improve the conspicuity-related motorcycle accidents in Malaysia. The variables that are taken into account to investigate the effectiveness of the "daytime running headlight" (RHL) and the regulations introduced to reduce the conspicuity-related motorcycle accidents are the influence of time trends, changes in recording and analysis systems, the effect of fasting during Muslim holy month of Ramadhan and the "balik kampung" (returning to hometown) culture during the annual festival celebrations. Their study have found conspicuity-related motorcycle accidents decreased by 29% when the "daytime running headlight" (RHL) was introduced. Reference [8] studied the level of injury of motorcycle riders when an accident occurred at T-junctions in the UK for different collision types which are rarely conducted. The collision types considered in this study are head-on, sideswipe, rear-end, approach-turn and angle. The predictor variables included in this study are human (gender and age of rider), weather (fine, bad or other), road (junction control, light condition, speed limit), month (spring, summer, autumn, winter), time (midnight, early mornings, rush hours, non-rush hours and evening), day of week and vehicle (engine size, crash partners). The results of their study indicate that the level of injury of motorcycle riders depends on different collision types and are associated with all independent variables in different ways. For example, injuries tend to be greatest at the uncontrolled junction due to sideswipe crashes and rear-end B collisions type. However, for those who were involved in approach-turn A at signalised junctions, the level of injuries will be much higher. In addition, mid-night or early morning is most significantly associated with more severe injuries; except for rear-end collision type A / B while crossing without lights causes serious injuries to the motorcyclist in head-on collisions compared to motorcyclist involved in other types of collisions.

The contributing factors of average hourly traffic flow, number of lanes, crashes involving one or more than two vehicles, crash types, weekend/weekdays and night or day on accident frequencies were investigated by [9]. They found that the incidence rates are more likely to increase as the traffic flow (vehicles per hour) increases. For light traffic travel, the number of crashes is higher on three-lane than on two-lanes and higher at weekends. In heavy traffic travel, the number of crashes is higher at weekdays. The number of deaths caused by accidents is higher at night even though the occurrences of accidence at night are lower as compared with day.

Almost all of the data on road traffic accidents are count data. Conventional models (Poisson or negative binomial regression model) have long been used to analyze accident frequency [9-12]. However, for modeling this type of data, Poisson and negative binomial model does not take into account the fact when there are many observed zero in accident data. For data with many observed zeros [13] and [14] used extended conventional models using Zero Inflated Poisson or Zero Inflated Negative Binomial models. It is found that, conventional models using Zero-Inflated model are much better in dealing with accident data when there are many observed zeroes. However, the models discussed above are not broadly used for accident data in Malaysia. When we have a combined time series and cross sectional data (also known as panel data), the appropriate count model are Fixed Effects Poisson, Fixed Effects Negative Binomial, Random Effects Poisson, Random Effects Negative Binomial and Dynamic Panel model [15]. The Fixed Effects Negative Binomial model was used on a panel count data of 25 countries from 1970 to 1999 and results showed that the implementation of road safety regulation, improvement in the quality of political institutions and medical care as well as technology developments have contributed to reduce motorcycle deaths [16]. Reference [17] presented an analysis on road accident occurrence using panel data analysis approach. The Fixed Effect Poisson and Negative Binomial model were used to analyze the accident data on 14 states in Malaysia for the period of 1996 to 2007. Among the factors considered in this study are the monthly registered vehicle in the state, the amount of rainfall, the number of rainy day, time trend and the monthly effect of seasonality. Their results indicate that the road accident occurrence are positively associated with the increase in the number of registered vehicle, the amount of rain and time while according to seasonality, the accident occurrence is higher in the month of October, November and December.

Recently, data mining techniques has been used to model motorcycle accidents and other data related to accidents. Data mining appears as a useful tool to analyze and interpret a large amount of data and maximum information can be gathered. Data mining techniques were applied to study the relationship between road characteristics and accident severity in Ethiopia using decision tree, naïve Bayes and K-nearest neighbor classifiers. The results of their studies indicate that the performances of the three models are almost equal [18]. Reference [19] conducted logistic regression, CART and multivariate adaptive regression splines (MARS) to analyze accident data. Their study indicated that CART and MARS are attractive models since these two models can display the results in graphical manner and able to determine the group of people who are of high risk to be involved in motorcycle accidents. The two models were then compared and the results showed that the MARS model is the best in providing information of potential risky areas according to age and number of years driving experience [19]. Reference [20] also used adaptive regression trees to analyze the real data obtained from Addis Ababa city traffic office. Their study focused on predicting the injury severity levels based on driver's age and gender, age and type of vehicle, road, light and weather condition as well as type and cause of accident. Their results show that the decision tree model accurately classifies the injury severity levels with 87.47% predictive accuracy which is

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

reasonably high. Similarly, Classification and Regression Tree (CART) one of the most widely applied data mining techniques and negative binomial model were used to analyze the accident data in Taiwan for the year 2001-2002. CART was found to perform better in analyzing the freeway accident frequencies. CART is different from the Poisson and Negative Binomial regression model since it does not require any predefined assumption about the data [21].

III. METHODOLOGY

This section explains in detail the modeling techniques used. The target variable is frequencies of motorcycle accident. For regression count model, the target variable is of count data type while for decision tree model, the target will be coded into three categories: 0=zero frequency, 1=low frequency (1-19) and 2=high frequency (20 and above). The sample data set was randomly split to create the training and testing samples. The number of observations for training sample is 896 observations (70% of the total observations) while testing sample is 384 observations (30% of the total observations). SPSS 16.0 were used to build a Poisson and Negative Binomial regression model while SPSS Clementine 12.0 (now known as IBM SPSS Modeler) was used to build the decision tree model.

A. Poisson Regression

The benchmark model for count data is the Poisson Regression model. Previously, researchers analyzed count data by using ordinary linear regression. Poisson regression has the advantage of being precisely tailored for discrete dependent variable which is highly-positively skewed. The Poisson regression model is appropriate for target variable that have only non-negative integer values such as motorcycle accident frequencies. Besides, the data y_i is assumed to be independent and follows a Poisson distribution. An unusual property of the Poisson distribution is that the mean and variance are equal:

$$E(y) = \text{var}(y) = \lambda$$

Let the dependent variable (y) be motorcycle accident frequencies which is drawn from a Poisson distribution with conditional mean of μ_i, given vector X_i for case i. Thus the density function of y_i can be expressed as;

$$f(Y_i \mid X_i) = \frac{\ell^{-\mu}\mu^{y_i}}{y_i!}, \text{ for } y = 0,1,2.. \quad (1)$$

Where $\mu_i = \exp(X^{'}\beta)$. In order to develop a Poisson regression model, μ_i is expressed as a function of some explanatory variables through a log link function in the following form;

$$\ln \mu_i = X^{'}\beta \text{ or } \mu_i = \exp(X^{'}\beta) \quad (2)$$

Given the independent observations assumption, with density function (1), the regression parameters β is estimated using the maximum likelihood method.

B. Negative Binomial Regression

In certain situations, overdispersion may arise and Poisson regression model is then not appropriate. Overdispersion arises when the observed variance of Y is greater than the mean. The most common parametric model for overdispersion is Negative Binomial which introduced a dispersion parameter to accommodate for unobserved heterogeneity in count data. This model is a generalization of Poisson Regression which assumes that the conditional mean μ_i of Y_i is not only determined by X_i but also a heterogeneity component of ε_i unrelated to X_i. The formulation can be expressed as:

$$\hat{\mu}_i = \exp(X^{'}\beta + \varepsilon_i) = \exp(X^{'}\beta)\exp(\varepsilon_i)$$

Where $\exp(\varepsilon_i) \sim Gamma\ (\alpha^{-1}, \alpha^{-1})$. As a result, the density function of Y_i can be derived as

$$f(Y_i \mid X_i) = \frac{\Gamma(Y_i + \alpha^{-1})}{\Gamma(Y_i + 1)\Gamma(\alpha^{-1})}.(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i})\alpha^{-1}.(\frac{\mu_i}{\alpha^{-1} + \mu_i})Y_i$$

The Negative Binomial distribution is derived as a gamma mixture of Poisson random variables with conditional mean and variance of $E(y_i \mid x_i) = \mu = \exp(x_i^{'}\beta)$ and $Var(y_i \mid x_i) = \mu_i + \alpha\mu_i^2$. Note that when $\alpha = 0$, the model becomes the Poisson regression model. Thus, Negative Binomial model has greater flexibility in modeling the relationship between the expected value and variance of Y_i. The smaller the α, the closer the Negative Binomial approaches the Poisson model [22-23].

C. Decision Tree

A *decision tree* model consists of a set of rules for dividing a large collection of observations into smaller homogeneous group with respect to a particular target variable. The target variable is usually categorical and the decision tree model is used either to calculate the probability that a given record belongs to each of the target category, or to classify the record by assigning it to the most likely category. Decision tree can also be used for continuous target variable although multiple linear regression models are more suitable for such variable. Given a target variable and a set of explanatory variables, decision algorithms automatically determine which variables are most important, and subsequently sort the observations into the correct output category [24]. The common decision tree algorithms in data mining software are CHAID (Chi-Square Automatic Interaction Detector), CART (Classification and Regression tree) and C5. The splitting criteria for CART, C5 and CHAID are gini, entropy and chi-square test respectively. These algorithms will produce the tree-like structure diagram and the decision rules whereby important information can be extracted [25].

Fig. 1 illustrates the construction of the decision tree model. The data partition node in SPSS Clementine 12 (now known as IBM SPSS Modeler) was not used because the data was initially partitioned using SPSS 16. Three algorithms were used to obtain the decision trees which are

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

CART, CHAID and C5.0. From these generated models, the best decision tree will be selected using the analysis and evaluation node.

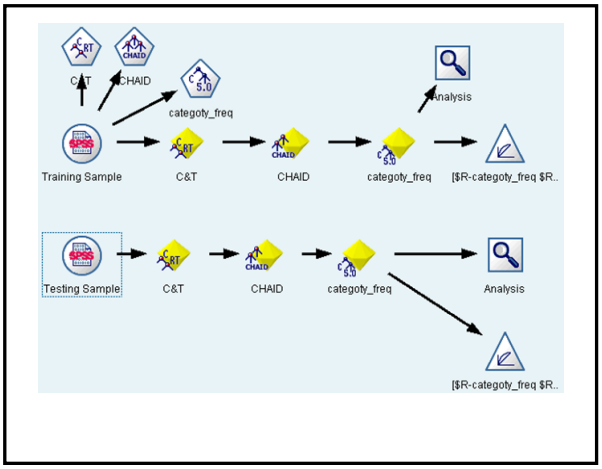

Fig. 1. Process Flow Diagram

Firstly, the decision trees (CART, CHAID and C5.0) nodes were connected to the Source node which contains the Training sample. The decision tree models for Training sample were then assessed and compared using the Analysis and Evaluation node. Subsequently, the CART, CHAID and C5.0 model nodes were connected to the Source node which contains the Testing (or validation) sample. The performance of the decision tree model for testing sample were then assessed and compared using the Analysis and Evaluation node. The best decision tree is then compared with the conventional models (Poisson and Negative Binomial Regression model).

IV. THE DATA

The motorcycle accident data were obtained from MIROS (Malaysian Institute of Road Safety Research) while MIROS obtained the accident data from the police department for the generation and dissemination of road safety data and information. The scope of this study only covers motorcycle accidents occurrences in Malaysia and the data provided contained the motorcycle accidents data that involved deaths or serious injury for the year 2008 and 2009. The data obtained from MIROS were messy and some variables have outliers and missing values. Hence, data cleaning, which involves checking completeness of data records, missing values, removing for errors were first performed. At this stage, the outliers, redundancy and cases with missing values are removed. Next, the accident data was reorganized based on the six independent variables and the target variable in this study which is the frequencies of motorcycle accidents. The final data set for modeling consists of 1280 observations. The variables included in this study are presented in Table 2.

TABLE II
DESCRIPTION OF VARIABLES

Variable Name	Role	Variable Type	Description
Motorcycle Accident	Target	Count/Categorical	Frequencies of motorcycle accidents / Category of Frequencies of motorcycle accidents 1: Zero frequency (0) 2: Low frequency (1-19) 3: High frequency (20 and above)
Collision type(CT)	Input	Categorical	Collision Type 1: Head on 2: Rear 3: Right Angle Side 4: Angular 5: Sideswipe 6: Hitting 7: Out of control 8: Forced/Overturned
Road geometry(RG)	Input	Categorical	Road geometry where motorcycle accident occur: 1: Straight 2: Bend 3: Roundabout 4: Junction 5: Interchanges
Time (T)	Input	Categorical	Time motorcycle accident occur: 1: Midnight/early morning (0000-0559) 2: Rush hours (0600-0859; 1600-1759) 3: Non-rush hours (0900-1559) 4: Evening (1800-2359)
Weather	Input	Categorical	Weather condition when motorcycle accident occur: 1: Clear 2: Not Clear (windy, foggy and rain)
Road surface condition(RSD)	Input	Categorical	Road surface condition where motorcycle accident occur: 1: Dry 2: Not Dry (flood, wet, oily, sandy and reconstruction work)
Days (WDAYS)	Input	Categorical	Days where motorcycle accident occur: 1: Weekdays 2: Weekends

V. RESULTS

In this section, the results of statistical modeling using Poisson, Negative Binomial Regression and decision trees model are presented.

A. Poisson and Negative Binomial Regression Results

The range of motorcycle accident frequencies for the training sample is from 0 to 418, with mean 10.28 and standard deviation of 39.081. The deviance for Poison regression model is 5.414, far from the value of 1, indicating overdispersion problem. .The deviance value for Negative Binomial regression is 0.779 which is much closer to 1. Besides that, the likelihood ratio chi-square is 915.915 with $p<0.05$ indicating that the negative binomial regression model is significant. Thus, at least one independent variable is a significant predictor of frequency of accidents. Based on Wald chi-square tests, all the six variables (collision type, road surface condition, time, weather, road geometry and

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

type of days) are significant since all or at least one category for each variable is significant ($p<0.05$).

In Table 3, the value of AIC and SIC for Negative Binomial regression model is lower than Poisson regression model. Hence, the Negative Binomial model demonstrates a better fit than the Poisson model.

TABLE III
ESTIMATED COEFFICIENTS OF POISSON AND NEGATIVE BINOMIAL REGRESSION MODELS

Variable	Poisson Regression Model	Negative Binomial Regression Model
Intercept	-9.417*	-7.508*
Collision type = head on	3.396*	3.360*
Collision type = rear	3.310*	2.950*
Collision type = right angle side	3.088*	2.946*
Collision type = angular	3.886*	3.742*
Collision type = sideswipe	2.902*	2.457*
Collision type = hitting	2.114*	1.851*
Collision type = out of control	3.013*	3.364*
Road geometry = straight	5.322*	5.319*
Road geometry = bend	3.877*	3.945*
Road geometry = roundabout	0.868*	0.740*
Road geometry = junction	4.509*	4.077*
Time = midnight/early morning	-1.267*	-1.207*
Time = rush hours	-0.295*	-0.223
Time = non rush hours	-0.008	-0.337*
Weather	2.784*	1.824*
Road surface condition=dry	2.577*	1.511*
Weekdays	0.859*	0.869*
Log Likelihood	-3071.72	-1512.13
No. of Parameters	17	17
AIC	6179.45	3062.26
BIC	6265.81	3153.42

** Significant at $\alpha = 0.05$*

Hence, the estimated Negative Binomial regression model is written as follow:

$$Log\ \mu = -7.508 + 3.360CT1 + 2.950CT2 + 2.946CT3 + 3.742CT4 + 2.457CT5 + 1.851CT6 + 3.364CT7 + 5.319RG1 + 3.945RG2 + 0.740RG3 + 4.077RG4 - 1.207T1 - 0.223T2 - 0.337T3 + 1.824CW + 1.511RSD + 0.869WDAYS$$

B. Decision Tree Results

There are three types of decision tree algorithms used in this study, the CART, CHAID and C5.0. All three decision trees found that the six variables (collision types, road geometry, time, weather, road surface condition and type of days) are significant. Results show that the most important variables is road geometry. The three decision trees were then compared based on the accuracy rate. Table 4 summarizes the accuracy rate for the three decision tree models applied on the training and testing sample.

TABLE IV
SUMMARY OF DECISION TREE RESULTS

Model		Accuracy Rate (%)
CART	Training	**83.37**
	Testing	78.12
CHAID	Training	80.25
	Testing	74.74
C5.0	Training	82.59
	Testing	**78.65**

The accuracy rate in Table 4 shows that CART has the best accuracy rate for the training sample compared to other models while C5.0 has the best accuracy rate for testing sample. However, the rules for C5.0 are complicated and the terminal nodes have very small number of cases. Overfitting refers to the situation in which the induction algorithm generates a classifier which perfectly fits the training data but has the lost capability of generalizing to instances not presented during training [26]. Despite the slight overfitting problem, the CART model is still acceptable. The decision tree model is too large to be displayed. Hence, only the interpretations of the CART rules for high frequency of motorcycle accidents that involve death or serious injury are presented in Table 5.

TABLE V
CART RULES FOR HIGH FREQUENCY OF ACCIDENTS

1. The motorcycle accidents happened on straight road, when the weather is clear, the road surface condition is dry and the accident involve head on/rear/right angle side/angular/sideswipe/out of control collision.
2. The motorcycle accidents happened at bend/junction road, when the weather is clear, the road surface condition is dry and the accident involve head on/angular collision.
3. The motorcycle accidents happened at bend/junction road, when the weather is clear, the road surface condition is dry, the accident involve rear/right angle side/sideswipe/out of control collision and the accidents occurred in the weekdays

C. Model Comparison

The comparisons of Poisson, Negative Binomial Regression and CART model are summarized in Table 6.

TABLE VI.
SUMMARY OF POISSON, NEGATIVE BINOMIAL AND CART RESULTS

Model		Accuracy Rate (%)
Poisson	Training	73.60
	Testing	76.30
Negative Binomial	Training	75.28
	Testing	77.60
CART	Training	**83.37**
	Testing	**78.12**

Based on the accuracy rate for Poisson and Negative Binomial Regression model, the prediction accuracy for testing sample are higher than the training sample. This is known as underfitting. Undefitting may occur when we mistakenly exclude important variables [27]. From the results presented in Table 6, CART was chosen as the best predictive model in predicting the category of occurrences of motorcycle accidents since it gave the best accuracy rate for training and testing sample even though there is slight overfitting.

VI. CONCLUSION AND RECOMMENDATIONS

Results of this study show that decision tree can be used to model motorcycle accident frequencies. The classification performance of decision tree model is quite comparable with conventional statistical models and the rules are easy to interpret. The results of this study also provide valuable information on how the collision types, road geometry, time, weather conditions, road surface conditions and type of days are related with motorcycle accident occurrences. The most important variable in predicting the occurrence of

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

motorcycle accident is road geometry with straight road contributing to the highest number of accidents that involve deaths or serious injury. This could be due to the possibility that when the road is straight, the riders would tend to ride at high speed and when an accident occurs, the situation will be worse or fatal. Further exploration of our results found that accidents that involved deaths or serious injury occur more often when the weather is clear and road surface condition is dry. Other than that, the motorcycle accident that involved deaths and serious injury is more likely to happen during evening and during weekdays. More data needs to be collected to confirm these findings.

Some research limitations arise in this study since police report data was analyzed. Besides, even though there are a lot of independent variables available in the database, the database contains a lot of missing values, unreliable data and some important variables cannot be further investigated for instances speed of motorcycle, helmet and alcohol use as well as cause of an accident. In future studies, more explanatory variables that might be available from other sources in Malaysia should be considered.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Malaysian Institute of Road Safety Research (MIROS) for their contributions by providing the data. We thank the Malaysia Ministry of Higher Education (MOHE) for the funding of this research. We also highly appreciate the comments from reviewers.

REFERENCES

[1] World Report on Road Traffic Injury Prevention. World Health Organization. (2004). Available: http://whqlibdoc.who.int/publications/2004/9241562609.pdf (Accessed on 12 April 2011)

[2] Global Status Report on Road Safety: Time for Action. World Health Organization. (2009). Available: http://www.un.org/ar/roadsafety/pdf/roadsafetyreport.pdf (Accessed on 12 April 2011)

[3] Status Paper on Road Safety (2009) Malaysia. Available: http://www.unescap.org/ttdw/common/Meetings/TIS/EGM-Roadsafety 2010/CountryStatus2009/13.Malaysia.pdf (Accessed on 12 April 2011)

[4] Fatalities by Mode of Transport. Available: http://www.miros.gov.my/web/guest/road (Accessed on 12 April 2011)

[5] Royal Malaysian Police, Traffic Division, Bukit Aman, Kuala Lumpur (2009). Statistical Report Road Accident Malaysia.

[6] Royal Malaysian Police, Traffic Division, Bukit Aman Kuala Lumpur (2010). Statistical Report Road Accident Malaysia.

[7] Radin, U. R. S., Mackay, M. G. and Hills, B. L., "Modelling of conspicuity-related motorcycle accidents in Seremban and Shah Alam, Malaysia," *Accident Analysis and Prevention*, vol. 28, no. 3, pp. 325-332, 1996.

[8] Pai, C. W. and Saleh, W., "Modelling motorcyclist injury severity by various crash types at T-junctions in the UK," *Safety Science*, vol. 46, no. 8, pp. 1234-1247, 2008.

[9] Martin, J. L., "Relationship between crash rate and hourly traffic flow on interurban motorways," *Accident Analysis and Prevention*, vol. 34, no. 5, pp. 619-629, 2002.

[10] Abdel-Aty, M. A. and Radwan, A. E., "Modeling traffic accident occurrence and involvement," *Accident Analysis and Prevention*, vol. 32, no. 5, pp. 633-642, 2000.

[11] Teoh, E. R. and Campbell, M., "Role of motorcycle type in fatal motorcycle crashes," *Journal of Safety Research*, vol. 41, no. 6, pp. 507-512, 2010.

[12] Indriastuti, A. K. and Sulistio, H., "Motorcycle accident model for severity level and collision type," *International Journal of Academic Research*, vol. 2, no. 5, pp. 210-215, 2010.

[13] Lee, A. H., Stevenson, M. R., Wang, K. and Yau, K. K. W., "Modeling young driver motor vehicle crashes: Data with extra zeros," *Accident Analysis and Prevention*, vol. 34, no. 4, pp. 515-521, 2002.

[14] Lee, J. and Mannering, F, "Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis," *Accident Analysis and Prevention*, vol. 34, no. 2, pp. 149-161, 2002.

[15] Wan Yaacob, W. F., Lazim, M. A. and Yap, B. W., "A Practical Approach in Modelling Count Data," *Proceedings of the Regional Conference on Statistical Sciences, 2010 (RCSS'10)*, pp. 176-183.

[16] Law, T. H., Noland, R. B. and Evans, A. W., "Factors associated with the relationship between motorcycle deaths and economic growth," *Accident Analysis and Prevention*, vol. 41, no. 2, pp. 234-240, 2009.

[17] Wan Yaacob, W. F., Lazim, M. A. and Yap, B. W., "Applying fixed effects panel count model to examine road accident occurrence," *Journal of Applied Sciences*, vol. 11, no. 7, pp. 1185-1191, 2011.

[18] Beshah, T. and Hill, S., "Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia," 2010.

[19] Kuhnert, P. M., Do, K. A. and McClure, R., "Combining non-parametric models with logistic regression: An application to motor vehicle injury data," *Computational Statistics and Data Analysis*, vol. 34, no. 3, pp. 371-386, 2000.

[20] Tesema, T. B., Abraham, A. and Grosan, C., "Rule mining and classification of road traffic accidents using adaptive regression trees," *International Journal of Simulation*, vol. 10, no. 6, pp. 80-94, 2005.

[21] Chang, L. Y. and Chen, W. C., "Data mining of tree-based models to analyze freeway accident frequency," *Journal of Safety Research*, vol. 36, no. 4, pp. 365-375, 2005.

[22] Cameron, A. C. and Trivedi, P. K., *Regression analysis of count data*. Cambridge University Press, 1998.

[23] Greene, W., *Functional forms for the negative binomial model for count data*. Economics Letters 99, 2008, pp. 585-590.

[24] Olson, D. and Yong, S., *Introduction to Business Data Mining*. McGraw Hill International Edition, 2006.

[25] Yap, B. W., Ismail, N.H. and Fong, S., "Predicting Car Purchase Intent Using Data Mining Approach," *IEEE Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) Proceedings*, 2011, pp. 2052-2057.

[26] Rokach, L. and Maimon, O. Z., *Data mining with decision trees: theory and applications*. World Scientific Publishing Co. Pte. Ltd, 2008, pp. 49.

[27] Yan, Xin. and Xiao, Gang. Su., *Linear regression analysis: theory and computing*. World Scientific Publishing Co. Pte. Ltd, 2009, pp. 157.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)