XUECAI XU, Ph.D.
E-mail: xuecai_xu@hust.edu.cn
Huazhong University of Science and Technology
& University of Hong Kong
China
ŽELJKO ŠARIĆ, Ph.D. Candidate
E-mail: zeljko.saric@fpz.hr
Faculty of Transport and Traffic Sciences,
University of Zagreb
Vukelićeva 4, 10000 Zagreb, Croatia
AHMAD KOUHPANEJADE, Ph.D.
E-mail: edkouhpa@yahoo.com
University of Nevada
Dept. of Civil & Environmental Engineering
4505 S. Maryland Parkway,
Las Vegas, NV 89154-4015, USA

Safety and Security of Traffic
Original Scientific Paper
Submitted: May 20, 2013
Approved: May 21, 2014

FREEWAY INCIDENT FREQUENCY ANALYSIS BASED ON CART METHOD

ABSTRACT

Classification and Regression Tree (CART), one of the most widely applied data mining techniques, is based on the classification and regression model produced by binary tree structure. Based on CART method, this paper establishes the relationship between freeway incident frequency and roadway characteristics, traffic variables and environmental factors. The results of CART method indicate that the impact of influencing factors (weather, weekday/weekend, traffic flow and roadway characteristics) of incident frequency is not consistent for different incident types during different time periods. By comparing with Negative Binomial Regression model, CART method is demonstrated to be a good alternative method for analyzing incident frequency. Then the discussion about the relationship between incident frequency and influencing factors is provided, and the future research orientation is pointed out.

KEY WORDS

data mining; Classification and Regression Tree; incident frequency; binary tree

1. INTRODUCTION

According to the statistics of the United States, the impact of incidents accounts for 50~60 percent of total delay on US freeways [1]. Thus, evaluating and analyzing the traffic delay caused by incidents has become more and more significant during the last decades. Although a number of transportation departments and agencies invest in a large amount of human resources, materials and financial supports every year to reduce traffic incidents, various accidents caused by traffic incidents stay at a high level, so it is necessary to find out the effective approaches to identify the influencing factors of traffic incidents.

Regression analysis, such as linear regression models, Poisson regression and Negative Binomial (NB) regression, has been widely used in the traffic safety field. However, most of regression models require the assumptions among the variables, and if these assumptions are violated, or homoscedasticity of the residuals is violated, the erroneous analysis results would be generated [2], e.g. the assumptions of the linear regression model requires that the dependent variable is continuous, the relationship between variables is inherently linear, and the observations are independently and randomly sampled. When any of the requirements are not met, the analysis results may be biased, and remedial actions should be taken.

Classification and Regression Tree (CART), one of the most popular data mining techniques, was introduced by Breiman et al. [3], and has been applied in business administration, medicine, industry, and engineering fields [4]. CART is an interesting and effective non-parameter classification and regression method, in which the binary tree is established to recursively partition the data into smaller and smaller strata so as to improve the fit as best as possible. But so far the application of CART in analyzing traffic safety problems has been rare. Therefore, the objective of this study is to investigate whether CART method can be utilized to analyze the various factors influencing incident frequency. The structure of the paper is as follows: the paper begins with the literature review of accident frequency, and then the methodology is presented. By using the data collected, CART method is analyzed and evaluated, and compared with NB regression models. Finally, the results are obtained; the conclusions are made and further investigation direction is pointed out.

2. LITERATURE REVIEW

Generally speaking, the average or total delay caused by specific incidents depends on incident duration, incident severity, incident frequency, traffic demand before and after incident, and carrying capacity [5], in which incident frequency, incident severity and incident duration are the most significant influencing factors, and incident delay can be considered as the function of these three factors. This study mainly investigates the incident frequency. Due to limited studies on incident frequency, but more on accident frequency, and accidents belong to the incidents with more serious injury severity, the literature review emphasizes the accident frequency.

The study on accident frequency has experienced different perspectives and approaches since years ago. From the methodology perspective, there have been many studies on the accident frequency models including (see Literature [6] for detailed description): Poisson models, negative binomial models, Poisson-lognormal models, zero-inflated count models, Conway-Maxwell-Poisson models, Gamma models, generalized estimating equation models, generalized additive models, random effects models, negative multinomial models, random parameters count models, finite mixture and Markov switching models, and other intelligent algorithms. Most of these studies focused on identifying the influencing factors such as intersection geometric features (i.e., number of through lanes, right-turn lanes, left-turn lanes, etc.), traffic control and operational features (i.e., signal phase, speed limit, etc.) and traffic flow characteristics (saturated and unsaturated), and these factors were found to have significant impact on the accident occurrence.

From the practice perspective, many researchers [7-9] attempted to assess the influencing factors of accident frequency by identifying impact factors, such as roadway geometric design (horizontal and vertical location, median type, or shoulder width), traffic features (hourly volume, average daily volume, vehicle proportion) and environmental conditions (land use, roadway condition, light condition or weather condition), etc. Shanker [10] developed NB regression model to analyze the impact of roadway geometric design and environmental condition on rural accident frequency. The results showed that the number of curves and weather-related factors (rainy and snowfall days) influenced accident frequency significantly, which served as the basis for cost-benefit analysis. Karlaftis and Tarko [11] employed NB regression model to investigate the relationship between crash frequency and influencing factors, such as vehicle miles travelled (VMT), population, and income. The results suggested that the methodology can be of assistance in developing improved models to account for possible difference in the highway sections examined. Ivan et al. [8] explored the impact of traffic density, land use, light condition, and bi-directional road lanes on multiple vehicle crashes, and showed that the ratio of flow/capacity, the proportion of blocked areas, shoulder width, number of intersections and the number of lanes had significant influence on single vehicle crashes, while the time, number of intersections and the number of lanes influenced the multiple vehicle crashes significantly. Carson and Mannering [9] examined the influence of warning signs on snow day accidents, and evaluated three separate models to analyze the accident frequency on inter-state freeway, major arterials and minor arterials. The results showed that the space (e.g. urban areas), roadway characteristics (e.g. shoulder width, roadway class), and traffic features (e.g. average daily traffic, the proportion of vehicles) had significant influence on accident frequency. To sum up, various studies have investigated the influencing factors of accident frequency from the methodological and practical perspective, and have concluded that different factors contribute to accident occurrence under different conditions, and each study has its own applicable requirements, but all of them are parametric models and require model assumptions.

Table 1 - Summary of accident frequency in literature

	Authors	Model & Method
Methodology Perspective	Lord and Mannering [6]	Poisson models, negative binomial models, intelligent algorithms, etc.
Practice Perspective	Ivan et al. [8]	Poisson regression models for single and multi-vehicle crash rates
	Carson and Mannering [9]	Effectiveness of ice warning signs in reducing accident frequency and accident severity
	Shanker [10]	Combine NB regression model with structural equation models
	Karlaftis and Tarko [11]	Combine NB model with cluster analysis

Data mining is a multi-discipline analysis technique, and has been widely used in various fields. Among different data mining techniques, decision tree and regular, non-linear regression and classification method, relative learning models are adopted frequently. However, the application of data mining technique in transportation field is still rare. In traffic safety field, a few studies have analyzed the accident frequency and damage severity based on the tree model, e.g. Kuhnert et al. [12] used logic regression, CART method and MARS (Multivariate Adaptive Regression Spline) to analyze vehicle damage data, and proved that CART method can identify high accident danger population by comparison. Karlaftis and Golias [13] set up the recursion tree model to analyze the impact of roadway geometry and traffic features on accident rates in double and multi lanes, which showed theoretical and practical advantage in accident rate analysis. Chang and Chen [4] analyzed the accident frequency with tree-based model to analyze roadway accidents on the National Freeway in Taiwan and Chang and Wang [14] developed CART method to establish the relationship between accident severity and driver/vehicle characteristics, highway/environmental variables and accident variables, and proved that CART method can be used for dealing with prediction and classification problems of accident frequency and severity, and Kashani and Mohaymany [15] verified this by analyzing the traffic injury severity on two-lane two-way rural roads. Moreover, the study by Pakgohar et al. [16] investigated human factors affecting prediction and classification of accident severity in Iran. The results showed that the driving license and the safety belt attributed to the crash severity and established the relationship between human factors and roadway crashes using CART method. Recently, Yap et al. [17] have compared CART with Poisson regression and negative binomial regression models for motorcycle accident frequency. The results showed that CART performed better than both count models, which gives the prerequisite in our study.

3. METHODOLOGY

3.1 CART Method

CART analysis is vital for prediction problems. When the target variable is a discrete value, the classification tree is formed, while the regression tree is used for a continuous target variable. The data classification and prediction rules established by CART are given in the form of a binary tree, each non-terminal node of the tree having a corresponding inquiry as branch base, and its algorithm includes tree growing, tree pruning and tree size selection. The method starts from the root node including all training data, and finds out the splitting point of the minimum division error through exhaustive search. After the splitting point is produced, the root node is divided into two sub-nodes, and then the same splitting procedure continues to be performed on the two sub-nodes till the classification error of the terminal node is less than the threshold value.

In practice, the decision tree is expected to be simple and compact, only a few nodes, i.e. the best one is the simplest model that is able to explain the data. The first step of CART is tree growing, whose basic principle is to split recursively the target variable so that the impurity of the terminal node is the minimum. The node impurity of the classification tree is defined as the following:

$$i(t) = \phi(p(1 \mid t), p(2 \mid t), \ldots, p(j \mid t)) \tag{1}$$

where $i(t)$ is the impurity measure of node t, $p(j \mid t)$ is node scale (the proportion of the amount of the dependent variable in node t related to class j), and ϕ is the non-negative function. The node impurity by Gini criteria, the default attribute of CART, can be expressed as:

$$i(t) = \sum_{i \neq j} p(i \mid t) p(j \mid t) = 1 - \sum_{i} p(i \mid t)^2 \tag{2}$$

For all input variables, the division is completed by searching of all possible threshold values of splitting points so as to find the maximum threshold value, which changes the impurity of the resultant nodes, that is to say, by selecting the search that provides the fastest reduction of the impurity:

$$\Delta i(s,t) = i(t) - p_R i(t_R) - (1 - p_R) i(t_L) \tag{3}$$

where s is the search, t_R and t_L are the right and left branch nodes, $i(t_R)$ and $i(t_L)$ are the impurity, respectively, p_R is the probability of tree growing from t to t_R when the search is accepted, and the best splitting point is to maximize $\Delta i(s,t)$.

The second step is tree pruning. Pruning is the mechanism of producing a series of simple trees by removing the important nodes. During the whole process of pruning, the smaller trees are created gradually, forming into a pruned tree series. Selecting the optimal pruned tree is to find out the optimal complexity parameter α to maximize equation (4), and the complexity of each sub-tree T $R_\alpha(T)$ can be described as:

$$R_\alpha(T) = R(T) + \alpha \left| \tilde{T} \right| \tag{4}$$

where $\left| \tilde{T} \right|$ is the complexity of the tree, equal to the number of terminal nodes of the sub-tree, α is complexity parameter, $R(T)$ is the misclassification cost of the tree, which can be defined as:

$$R(T) = \sum_{r \in T} r(t) p(t) \tag{5}$$

where $r(T)$ is the node misclassification cost, defined as

$$r(t) = 1 - p(j \mid t) \tag{6}$$

The third step is to select the tree of an appropriate size from the pruned ones. When applied to analyze a new database, an over-sized number would produce higher misclassification. When the sample number is not big enough, all the data are usually expected to be used to establish the tree, and cross validation evaluation method can be used to provide an error rate assessment for each sample as well as to establish the tree. The data are divided into training and learning data; one subset is separated from the training data for tree construction, and the rest for misclassification rate assessment. Then multiple replicated divisions are performed for different subsets, and the obtained misclassification rate is averaged so as to reach the cross validation evaluation of a suitable tree. The tree size of producing the minimal cross validation evaluation method is determined to be the final model. More details about CART analysis and application can be found in Breiman et al. [3], Chang and Chen [4].

3.2 Negative Binomial Regression Model

Statistics modelling techniques have been utilized in analyzing the relationship between accidents and influencing factors years, in which NB regression models are widely used due to the discrete and non-negative attributes of accident frequency. The process of accident occurrence can be viewed as a Bernoulli trial, each with unequal probabilities of independent events. A Bernoulli trial has two potential outcomes: one is considered as a "success" (i.e., accident) and the other is "failure" (i.e., no accident). The number of trials with "success" in a certain time period follows binomial distribution. With the large number of trials, the binomial distribution can be approximated with a Poisson distribution. Poisson regression models applied to this study to relate the expected number of accidents λ to explanatory variables can be expressed as:

$$\ln(\lambda_i) = \beta \cdot X_i + \varepsilon_i \tag{7}$$

where X_i is a vector of explanatory variables, β is a vector of estimable parameters, and $\exp(i)$ is a gamma-distributed error term with mean one and variance α^2. The resulting negative binomial probability distribution is:

$$P(y_i) = \frac{\Gamma((1/\alpha)+y_i)}{\Gamma(1/\alpha)y_i!}\left(\frac{1/\alpha}{(1/\alpha)+\lambda_i}\right)^{1/\alpha}\left(\frac{\lambda_i}{(1/\alpha)+\lambda_i}\right)^{y_i} \tag{8}$$

where $\Gamma(x)$ is a value of the gamma function, y_i is the number of accidents and α is an over-dispersion parameter. More details about NB regression can be found in Washington et al. [2].

4. DATA DESCRIPTION

The incident data, specifically for this study, were collected from December 1995 to February 1996 by New York Department of Transportation, and the incident types for analysis included C0: incident free, C1 property damage, C2 injury and death, and C3 wrecked vehicle. The analogy to crash severity classification, the severity levels increase from C0 to C3, C0 is the slightest, no incident, C1 only involves property damage, C2 includes injury and death, and C3 is the worst, vehicles involved are wrecked. The data are stored as the four types. Although there were also other incident types, those types were not significantly associated with traffic operation.

The incidents in this study contained complete information, involving the date, road name, detection time, clearance time, number of lanes blocked by incidents, units or departments involved by incident response, incident location, incident type, vehicle types involved, number of vehicles involved, weather, and the closest ramp names on both sides of the incident location. In order to compare CART method with the statistics model, the data collected were divided randomly into two types for training and testing, and the total number was 858 (accounting for 75% of the total) and 286, respectively. The reason that the training sample is 3 times the testing data is to train the model better so that the testing results are more accurate. The Chi-square tests were used to establish whether or not an observed frequency distribution differs from the theoretical distribution, which indicated that the incident frequency distribution of the two samples was similar, 13.683 and 14.426, respectively.

In order to investigate the impact of roadway geometry on incident frequency, the roadways are required to be divided into homogenous segments, representing the geometry related variables. One approach to splitting the roadways is to divide the roadways into equal segments (Shankar 1997); the other way is to divide them according to homogenous geometric design and traffic flow because two adjacent ramps can be used as the splitter. The two methods have the advantages and disadvantages, respectively: the weakness of equal segment division is difficult implementation, while the critical problem of the latter one is that the unequal segments intensify the potential heteroskedasticity problem due to homogenous requirements. To overcome the weaknesses of the two methods, the compromised one was adopted. First, on the base of roadway exits all the segment breakpoints were selected, so the segment length was not equal. However, from the exit point the breakpoint was selected and combined with the shorter homogenous segment, so that the segments were almost equal to each other. In this study the segment length is around 3.0 miles, the average length is 2.946 miles, the shortest is 2.407 miles and the longest is 3.463 miles. The standard deviation of the length is 0.3 miles; thus the segment length can be considered as basically equal.

5. INCIDENT FREQUENCY MODEL

5.1 CART Method Estimation

The factors leading to the number of incidents within some segments vary, such as weather, weekday/weekend, roadway conditions, roadway geometry, etc., thus the variables of incident frequency model in this study are composed of four groups as listed in *Table 2*. Group 1 is weather condition, including rain and snow conditions; Group 2 is related to time features. Because the peak period in incident frequency cannot be used as an independent variable, both peak and off-peak periods need to be analyzed, so weekday is considered as one variable; Group 3 is associated with traffic flow features; this variable reflects the congestion level of segment selected, and the congestion level takes into consideration the lane numbers as well as the traffic volume; Group 4 is about the roadway geometric features.

Figure 1 shows the classification tree produced by CART method, and the splitting procedure is as follows: the initial splitting of node 1 is based on the number of congested lanes. If the number of congested lanes is more than 2, CART puts it on the left side, forming terminal node 1, otherwise it puts it on the right side as node 2. For terminal node 1, CART predicts that 10% (1/10) incident free occurs, that is to say, under this condition, the probability of incident occurrence is small. But for node 2 it is still possible to cause congestion when the number of lanes congested is not more than 2, so the variable of splitting incident frequency selects the average traffic volume. If the average traffic volume is more than 2,000, node 3 is produced on the left side, while node 4 is on the right side if the index is not more than 2,000. Because the average traffic volume is different on weekdays and weekends, node 3 and 4 are divided into node 5 and 6, node 7 and 8, respectively. Considering the influence of rain and snow weather on roadway traffic conditions, node 5 and 6, node 7 and 8 are split step by step till the terminal nodes, in this way the whole tree of freeway incident frequency prediction can be obtained.

5.2 Comparison between CART and NB Regression Model

Incident frequency shows different results at different time periods, e.g. the traffic conditions and segment travel time at peak hours (6 to 9 a.m. and 5 to 8 p.m.) and off-peak hours are definitely different, so the two conditions need to be considered separately. *Table 3* summarizes the estimation results of NB regression model, and the significant influencing factors. The estimated coefficients with positive signs represent that those variables may increase the incident possibility significantly, e.g. the coefficient of rainy day is positive, meaning that as the rain volumes increases, the possibility of incident occurrence is raised up whether during peak or off-peak hours. Similarly, the estimated coefficients with negative signs representing the occurrence of incidents are less likely. Compared with NB model, it can be found that CART method relies more on traffic and environmental variables than on geometry when dividing incident frequency, as shown in node splitting, the average ramp distance and waving area have no impact on incident occurrence.

In order to investigate the performance of CART method in analyzing freeway incident frequency, the predictive performance between CART method and the NB regression model should be compared. By using the NB regression model to predict incident frequency, first the average incident frequency (i.e. λ in Equation (7)) can be determined. After having the average incident frequency of each individual freeway section, the probability of incidents can be calculated and

Table 2 - Variable description

Variable	Description	Average	Min.	Max.	Std. Err.
Group 1: Weather Condition					
Rain	1: Rain, 0: No rain	0.26	0	1	0.2
Snow	1: Snow, 0: No snow	0.25	0	1	0.19
Group 2: Time Feature					
Weekday	1: Incident on weekdays, 0: Incident on weekends	0.73	0	1	0.2
Group 3: Traffic Flow Feature					
Average traffic volume	Average traffic volume per lane	1,043	351	1,516	817.84
Group 4: Geographic Feature					
Lane congested	Number of lanes congested within segment	2.95	2	3	0.05
Average ramp distance	Segment length divided by ramp numbers (mile)	0.64	0.28	1.36	0.10
Lane changing	1: Lane changed, 0: No lane changing	0.55	0.00	1.00	0.26
Waving area	Number of waving areas within segment	0.50	0.00	2.00	0.58

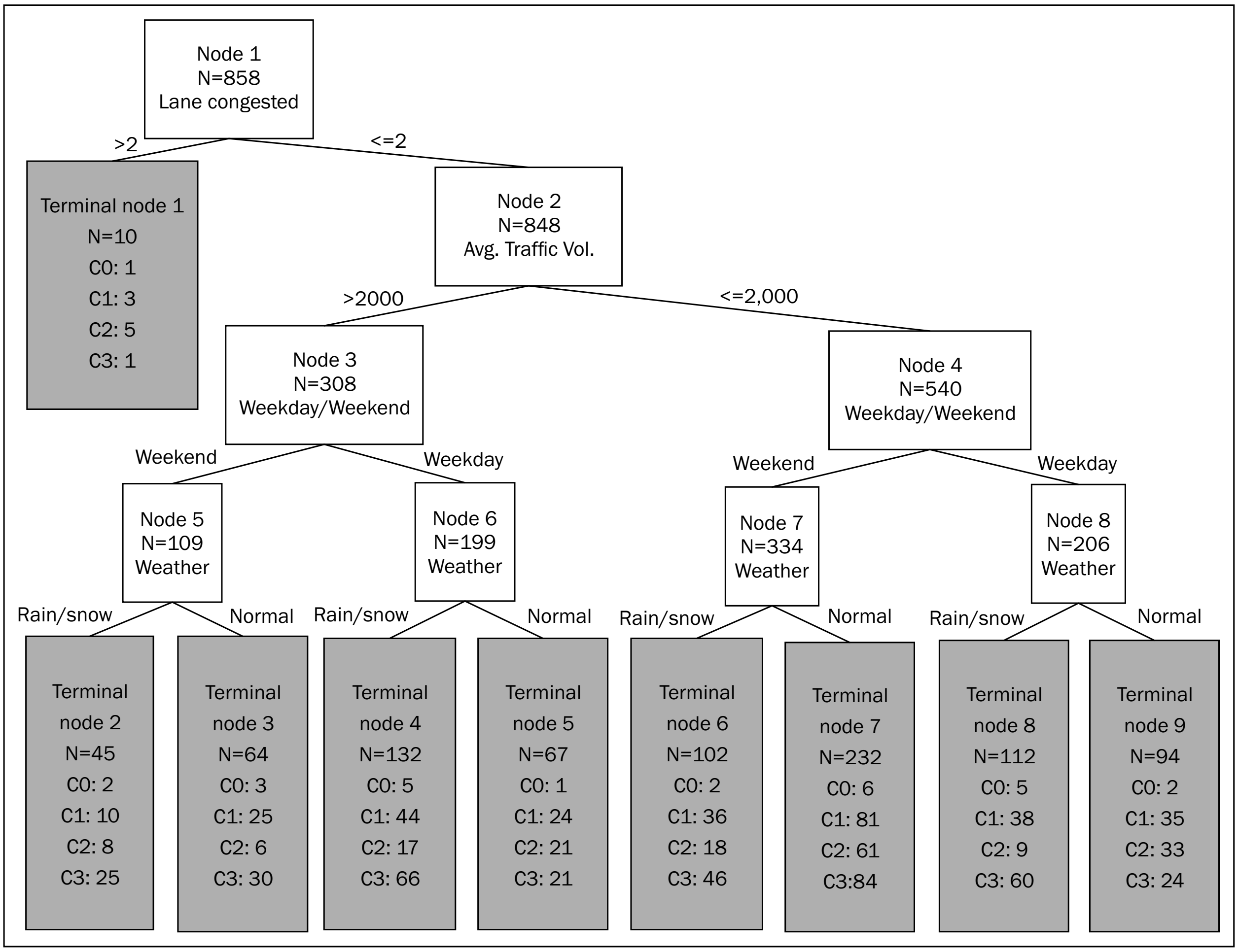

Figure1 - Output of CART Method

Table 3 - Variables of NB Regression Model

Variable	Peak Hour		Off-Peak Hour	
	Est. Coefficient	t-test	Est. Coefficient	t-test
Constant	-7.129	-2.729	-8.663	-4.313
Weekday	0.521	1.572	-0.286	-1.760
Snow	0.175	1.636		
Rain	0.294	2.776	0.489	3.010
Lane congested	0.927	1.103	1.254	2.565
Average ramp distance	-0.350	-1.669		
Lane changing	0.629	2.116	0.545	3.181
Congestion index	1.060E-03	1.412	2.190E-03	3.756
$t_{\alpha NB}$	3.765	2.749	1.554	3.357
Log-likelihood at zero	-439.417		-595.515	
Log-likelihood at convergence	-414.669		-561.759	
Freedom	7		4	
Likelihood ratio testing	49.496		67.508	

the classification can be determined by the frequency category with the largest probability. For instance, for a particular freeway section the incident probabilities with NB model prediction are 40%, 30%, 15%, 10% and 5% for 0, 1, 2, 3, and 4 or more incident frequencies, respectively, and then this freeway section is considered as having one incident frequency.

As for CART method, the incident frequency can be achieved by following each node till the terminal one. *Table 4* and *Table 5* display the prediction results. By

Table 4 - Estimated results of CART method

Training data					
Observed frequency	Predicted frequency				
	C0	C1	C2	C3	Total
C0	10	6	7	4	27
C1	5	154	59	78	296
C2	3	45	85	45	178
C3	2	62	65	228	357
Total	20	267	216	355	858

Total prediction accuracy of training data is 55.6%.

Testing data					
Observed frequency	Predicted frequency				
	C0	C1	C2	C3	Total
C0	3	2	2	1	8
C1	2	51	20	26	99
C2	1	15	28	15	59
C3	1	21	22	76	120
Total	7	89	72	118	286

Total prediction accuracy of testing data is 55.2%.

Table 5 - Estimated results of NB Regression Model

Training data					
Observed frequency	Predicted frequency				
	C0	C1	C2	C3	Total
C0	11	6	4	6	27
C1	5	145	59	87	296
C2	3	44	85	46	178
C3	3	90	80	184	357
Total	22	285	228	323	858

Total prediction accuracy of training data is 49.5%.

Testing data					
Observed frequency	Predicted frequency				
	C0	C1	C2	C3	Total
C0	4	2	1	2	9
C1	2	48	20	29	99
C2	1	15	28	15	59
C3	1	30	27	61	119
Total	8	95	76	107	286

Total prediction accuracy of testing data is 49.3%.

combining the freeway segment with more than two incidents, the Chi-square test of 3 by 3 correlation table was conducted to compare the prediction with observation frequency. Although the predicted and observed results are different for training and testing data, *Table 4* and *Table 5* provide valuable information for the prediction performance of CART method and NB regression model. For CART method, the prediction accuracy of training data is 55.6% while the accuracy of testing data is 55.2%. Similarly, the prediction accuracy for NB regression model is 49.5% and 49.3%, respectively. It can be seen that the accuracy of CART method is higher than that of NB regression model, and the prediction of each incident type is higher than that of NB regression model, so based on the results above, CART method can be an alternative to NB regression model in incident frequency analysis. Although the prediction accuracy is increased only around 5%, it can be proved that CART method is more effective and easier to implement by non-professionals than NB model, even though the difference is not so significant.

5.3 Results analysis and discussion

In this study, the prediction performance provided by CART method and NB regression model with training and testing data was considered to be similar, and it can be seen that CART analysis is an effective method to forecast incident frequency. Now the parameters influencing incident frequency need to be discussed.

The traffic volume within roadway segment is the product of the existing number of lanes and average traffic volume. The existing number of lanes determines the segment capacity, so the number of congested lanes is considered as one splitting criterion. However, the average traffic volume reflects the degree of segment congestion level, i.e. the higher the level of congestion, the higher the probability of incident occurrence. During the peak period, higher congestion level increases the number of incident types, while during the off-peak period the congestion level raises the frequency of C1 property and damage, and C2 injury and death, but C3 wrecked vehicles show no obvious rise.

Weekday is a good examination for traffic volume and trip purpose. The trips on weekdays are likely to be related to work, and the trip volume is high, especially during peak hours, thus more incidents would probably occur. Shown from the estimation results, weekday has positive impact on C1 and C2 during peak period, that is to say, more incidents happen during peak periods on weekdays than on weekends. The result is reasonable because on weekdays higher traffic volume during peak period increases the chances of incident occurrence whereas during off-peak period on weekdays have no obvious influence on incidents, but the incident frequency of C3 is lower than that on weekends because shopping and travelling account for most of the traffic on weekends; moreover, the traffic volume during off-peak periods on weekends is even higher than on weekdays. Additionally, the trip routes on weekends are not so straight as on weekdays, which might cause the drivers to prolong the trips.

The impact due to rain and snow weather on roadway is obvious. The results indicate whenever during peak and off-peak periods the frequency of C3 is kept high, and the main reason is that rain and snow have negative influence on vehicle braking system, which makes the vehicles hard to operate at low speeds. In addition, the visibility caused by rain and snow is reduced, as well as road surface friction, so the possibility of incident occurrence is raised. Rain and snow show certain relation with C1 frequency during peak period mainly due to the higher traffic volume, but has no obvious association with C2 frequency probably because of the driver's caution and low-speed driving. To sum up, C3 incident frequency is increased significantly due to rainy and snowy days, while C1 frequency is raised slightly, but C2 frequency displays a falling trend.

In comparison to the NB regression model or other parameter models, CART analysis provides theoretical and practical advantages. In theory, it is unnecessary for CART analysis to know the model function in advance and attached relationship assumptions among risky factors. Furthermore, CART analysis can deal with co-linearity issue effectively. In practice, CART method is capable of displaying the analysis results clearly and predicting the incidents through binary tree structure distinctly. Moreover, CART method is capable of searching for the optimal splitting point automatically. However, CART method has its own weakness, e.g. it cannot effectively utilize consecutive and serial variables, or provide the probability level of influencing factors and prediction, and it is difficult to conduct the elasticity or sensitivity analysis. Another deficiency is that the data sample selected should be big enough, as indicated in this study, the training data should be much larger than the testing data so as to guarantee the prediction accuracy of the model, otherwise the results inference might be biased or misinterpreted.

6. CONCLUSION

Based on CART method of non-parameter regression, this paper establishes the relationship between freeway incident frequency, roadway characteristics, traffic variables and environmental factors. The results show that the influencing factors (weather, weekday/weekend, traffic flow and roadway features) of incident frequency vary from different incident types during different time periods. The prediction performance proves that CART method is an alternative to analyze incident frequency, although it is a small methodological step in analyzing the incident frequency.

Shown in the results, both CART and NB regression models perform similarly, and CART method is a quite practical and efficient tool for those with a non-statistical background, which is more easily conducted than the traditional NB regression model. On the other side, CART method cannot solve some issues that NB regression model deals with, e.g. the marginal effects and elasticity from NB regression model can provide insights into the analysis process, the heterogeneity of accidents modelling can be addressed by random-parameter NB models [18, 19], and temporal and spatial analysis of accidents, the endogeneity and heterogeneity issues can be interpreted by panel data random-parameter NB model and panel data simultaneous equation models [20, 21]. Therefore, further exploration of CART method might provide a better understanding of the influencing factors of incident frequency for highways and intersections.

For the future work, the study will compare the analysis results between CART method and statistical models. As mentioned in the paper, statistical models such as NB regression model has been used to analyze the influencing factors of accident frequency. By comparing the influencing factors and prediction performance between them, it can provide valuable insights into the relationship between the influencing factors and incident frequency. However, the comparisons between non-parametric and parametric tree-based models should be made carefully, because tree-based models are often unstable [14].

Future studies might focus on how CART method uncovers more potential influencing factors and improves the prediction performance. Finally, more studies should continue with various data mining techniques, e.g. association principle and neural network, to analyze the influencing factors of incident severity and incident duration, and find out the suitable analysis tool, so that the traffic management and engineering stuff can improve roadway traffic conditions gradually and improve the roadway design.

ACKNOWLEDGEMENTS

This study was supported by the Fundamental Research Fund for the Central Universities (HUST: 2013QN031), National Natural Science Foundation of China (NSFC) (Grant No: 51208222 & 51308242) and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry of China. Thanks for the reviewers' valuable suggestions and comments.

REFERENCES

[1] **Lindley JA**. *Urban freeway congestion: quantification of the problem and effectiveness of potential solutions.* ITE J. 1987 Jan;57:27-32.

[2] **Washington SP**, **Karlaftis MG**, **Mannering FL**. *Statistical and econometric methods for transportation data Analysis*, New York: Chapman and Hall/CRC; 2003.

[3] **Breiman L**, **Friedman JH**, **Olshen RA**, **Stone CJ**. *Classification and regression trees*. London: Chapman & Hall/ CRC; 1998.

[4] **Chang LY**, **Chen WC**. *Data mining of tree-based models to analyze freeway accident frequency*. J Safety Res. 2005;36(4):365-375.

[5] **Yang P**, **Wu B**. *Traffic management and control*. Beijing: Renmin Traffic Press; 2004.

[6] **Lord D**, **Mannering FL**. *The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives*. Transp Res Part A Policy Pract. 2010 Jun;44(5):291-305.

[7] **Poch M**, **Mannering FL**. *Negative binomial analysis of intersection-accident frequencies*. J Transp Eng. 1996 Mar;122(2):105-113.

[8] **Ivan JN**, **Wang C**, **Bernardo NR**. *Explaining two-lane highway crash rates using land use and hourly exposure*. Accid Anal Prev. 2000 Nov;32(6):787-795.

[9] **Carson J**, **Mannering FL**. *The effect of ice warning signs on accident frequencies and severities*. Accid Anal Prev. 2001 Jan;33(1):99-109.

[10] **Shankar VN**. *Limited dependent variable and structural equations models: empirical applications to traffic operations and safety*. Dissertation: University of Washington; 1997.

[11] **Karlaftis MG**, **Tarko AP**. *Heterogeneity considerations in accident modeling*. Accid Anal Prev. 1998 Jul;30(4):425-433.

[12] **Kuhnert PM**, **Do K**, **McClure R**. *Combining non-parametric models with logistic regression: an application to motor vehicle injury data*. Comput Stat Data Anal. 2000 Sept;34(3):371-386.

[13] **Karlaftis MG**, **Golias I**. *Effects of road geometry and traffic volumes on rural roadway accident rates*. Accid Anal Prev, 2002 May;34(3):357-365.

[14] **Chang LY**, **Wang HW**. *Analysis of traffic injury severity: an application of non-parametric classification tree techniques*. Accid Anal Prev, 2006 Sept;38(5):1019-1027.

[15] **Kashani AT**, **Mohaymany AS**. *Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models*. Safety Sci. 2011 Dec;49(10):1314-1320.

[16] **Pakgohar A**, **Tabrizi RS**, **Khalili M**, **Esmaeili A**. *The role of human factors in incidence and severity of road crashes based on CART and LR regression: a data mining approach*. Procedia Comput Sci. 2011;3:764-769.

[17] **Yap BW**, **Norashikin N**, **Wong, SV**, **Mohamad AL**. *Decision tree model for count data*. Proceedings of the World Congress on Engineering 2012. Vol I; July 4-6, 2012, London, U.K.

[18] **Anastasopoulos PCh**, **Mannering FL**. *A note on modeling vehicle-accident frequencies with random parameter count models*. Accid Anal Prev. 2009 Jan;41(1):153-159.

[19] **El-Basyouny K**, **Sayed T**. *Accident prediction models with random corridor parameters*. Accid Anal Prev. 2009 Sept;41(5):1118-1123.

[20] **Wang X**, **Abdel-Aty M**. *Temporal and spatial analyses of rear-end crashes at signalized intersections*. Accid Anal Prev. 2006 Nov;38(6):1137-1150.

[21] **Xu X**, **Kwigizile V**, **Teng H**. *Identifying access management factors associated with safety of urban arterials mid-blocks: a panel data simultaneous equation models approach*. Traffic Inj Prev. 2013;14(7):734-742.