Prediction of Freeway Incident Duration based on Classification Tree Analysis

Hsin-Li CHANG [a], Tse-Pin CHANG [b]

[a,b] *Department of Transportation Technology and Management, National Chiao Tung University, 1001, University Road, Hsinchu 30010, Taiwan, ROC*
[a] *E-mail: hlchang@cc.nctu.edu.tw*
[b] *E-mail:tembin.tem88g@nctu.edu.tw*

Abstract: How to provide accurate travel time under incidents is one of the most important premises for highway safety management. Although there was an increase in the studies of predicting incident duration recently, predicting incident duration accurately is currently still a challenging technology due to the quality of incident dataset. Additionally, the previous studies still have some limitations for empirical implication, such as pre-defined function form and strict statistical assumptions. This study employed the huge incident data, recorded accurately from Taiwan freeway systems, to develop predicting incident duration model without the disadvantages of traditional statistical techniques. The analysis results represented that the most two important variables are number of large-sized vehicles and incident type, and provided decision rules to predict incident duration. Therefore, the motorists will avoid traffic jam by changing travel route while the traffic management organization makes decisions timely to mitigate traffic congestion and clear the incident effectively.

Keywords: Incident Duration, highway safety, Classification Tree Analysis

1. INTRODUCTION

Providing accurate and real-time travel time under incidents is one of the most important premises for Intelligent Transportation Systems (ITS) and highway safety management (Garib et al., 1997; Chung, 2010). Under normal traffic conditions, motorists choose the best route based on their travel plans or experience so they always neglect traffic information. In contrast, motorists hope to acquire accurate and real-time traffic information under incidents for deciding whether or not to change their travel routes. In addition, the traffic control center can make effective and real-time decisions of traffic diversion and dispatching emergency resources based on accurate traffic information. In brief, as technology requirements in ITS and highway safety management increase, so does demand for predicting incident duration more precisely and timely.

Although the studies in predicting incident duration increase recently, predicting incident duration is currently still a challenging technology. The accuracy on prediction of incident duration in previous research is not enough to apply to empirical area at present. The main reason is that the incident dataset in previous studies are usually collected from rescuer organizations and they are not representative. Because there were smaller sample size and longer incident duration, in which most of the sample were serious incidents with typically closing two or more lanes of traffic for longer duration and needed some rescuer vehicles. For example, the average incident duration was 162.5 minutes and the number of incidents was 681 cases in Washington State (Nam and Mannering, 2000), the sample size was 700 cases in Virginia State (Ozbay et al., 2006), and the sample size was only 24 cases in Wei's study

(2007). However there were a lot of incidents with short duration, in which there was no rescuer vehicle, recorded in traffic control center and they were not recorded in the rescuer organizations. Additionally, traffic control center is the first receiver of incident occurrence and then notify rescuer organization for specially equipped vehicles so that its incident dataset is more accurate and complete. Therefore, the dataset with most types of incident and being recorded correctly is the premise of predicting incident duration.

In addition to the quality of dataset, another reason affecting the accuracy of predicting incident duration is to employ the appropriate approach. Previous studies have applied many approaches to analyze incident duration ranging from ANOVA to regression analysis (Golob et al., 1987; Giuliano, 1989; Jones et al., 1991; Garib et al., 1997); nonetheless, most of them have strict model assumptions regarding the dataset, such as pre-defined function forms, which are usually difficult to satisfy. Accordingly, the statistical model could result in incorrect estimations of likelihood of target variable if these assumptions are violated. In comparison to the above-mentioned traditional parametric models, some studies tried to apply Data Mining techniques to analyze incident duration, such as Artificial Neural Networks (ANN) (Wei and Lee, 2007) and Bayesian Networks approach (BNs) (Ozbay et al., 2006). Although ANN approach has a reasonable forecast ability, it has inability to identify the relative importance of potential input variable (Piramuthu, 1999). Additionally, even though BNs approach can be used to create dynamic incident duration estimation trees, there are some application limitations, such as determining prior distribution on decision variables and time consuming on constructing the network. In comparison to those limitations of studies above mentioned, Kuhnert et al. (2000) noted that classification and regression tree (CART) is capable of automatically searching for the best input variables with the best threshold values to classify the target variable, and it has been largely applied to many areas, such as engineering, business administration, and industry. Some studies have employed CART model to analyze the relationship between injury severity and input variables associated with crash severity such as Chang et al. (2006) and Sohn et al. (2001, 2003). However, few studies use CART to predict incident duration or explore the important factors affecting incident duration at present.

From an empirical viewpoint, Giuliano (1988) used his statistical model to estimate incident duration and the important finding was significant effect of truck involvement in incident duration. In addition, Garib et al. (1997) noted that "81% of variation in incident duration can be predicted by number of lanes affected, number of vehicles involved truck involvement, time of day, police response time, and weather condition." There was similar result by Chung (2010) that as bigger vehicles involved and the number of the involved vehicles (or the injured) increased, the incident duration increased.

As we indicated in the above studies, it is an important but complex problem in traffic safety area to classify and analyze relationships among a lot of variables. There have been a number of studies, which have explored the relationships between the factors and incident duration, and some data mining techniques have been applied in this area. Nonetheless, there are still many limitations to make a breakthrough that the urgent matter is predicting incident duration based on representative data and identify the important factors. This study used Taiwan official incident records with 4,697 cases and they included most of the incidents on Taiwan freeway systems. In addition, Bevilacqua et al. (2008) has noted that the CART technique provides new case-effect relationships in safety area without the disadvantages of traditional statistical techniques. Ozbay and Kachroo (1999) have suggested that incident duration could be estimated by decision trees. As a result, the goal of this study is to develop a predicting method of freeway incident duration with huge sample size based on classification tree analysis which has no pre-defined function forms between dependent and independent variables. Additionally, another goal of this study is to identify the important factors affecting incident duration. Initially, this paper begins with a brief presentation of the dataset and the method. Next, a presentation of the assessment of the analysis results follows this. Finally, this study concludes with a summary of findings and recommendations for future work.

2. DATA DESCRIPTION

The incident data, which were in computer-ready form, used in this study were obtained from Short Message Service Database of Traffic Incident by Taiwan National Freeway Bureau (TANFB). When the traffic management center received a report of incident occurrence, which might be obtained from cellular telephone reports, extensive closed circuit television (CCTV), police, or patrol vehicles, it would transmit continually cellular phone short message to senior official, region engineering office, and police until the incident was cleared. Accordingly, most of the incidents on Taiwan freeway systems are accurately recorded and the database includes date, notification time, clearance time, number of vehicle involved, kind of vehicle, location, number of lanes affected, incident type, number of person (fatal and Injured).

Several studies (TRB, 1994; Nam et al., 2000; Garib et al., 1997) have noted that the incident delay consists of four phases. As shown in Figure 1, the incident delay can be divided into four parts at five time points that include incident occurrence, incident notification, rescuer arrival, incident cleared, and normal operation. The first three phases illustrate incident duration, and the fourth (recovery time) phase represents the effect persisting of incident after it is cleared. Because the exact time of incident occurrence is not easy to detect precisely, the accuracy of incident detection time is doubtful. In addition, the length of recovery time depends on the traffic demand rate and it is difficult to predict or control. In contrast, the time points of incident notification and incident cleared in this study have official records so that they always have accurate lengths of time.

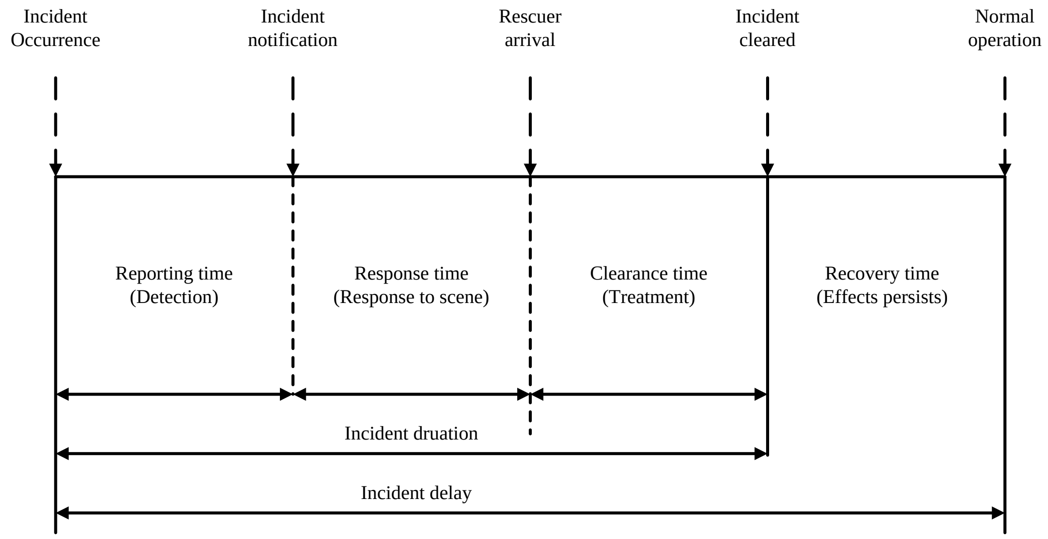


Figure 1. Component of incident delay (Wei et al., 2007; Nam et al., 2000; Chung, 2010)

In this study, the total recorded cases during the year 2008 was 4,908 incidents. Basic checks on the data quality were done and the cases with questionable information were screened out, such as the data of extremely short duration (e.g. under 5 minutes) affecting slightly the traffic or too short to respond. Finally, this study employed the dataset of 4,697 cases after screening out the data with questionable information. As the distribution of incident duration presented in Figure 2, the mean of duration is 36.9 minutes and 90% of the incidents are under 60 minutes. Although standard deviation is 29.9 minutes, it has an extreme right-skewed distribution and the maximum duration is as long as 391 minutes.

As noted in Zeng et al. (2005) and Tan et al. (2006), the data characteristics is not well suited to following analysis, which it may result in appeasing excessively analysis result and inefficiency. Thus the data processing in this study is the discretization of the target variable, which it is transformed from a continuous variable into a categorical variable by cluster analysis.

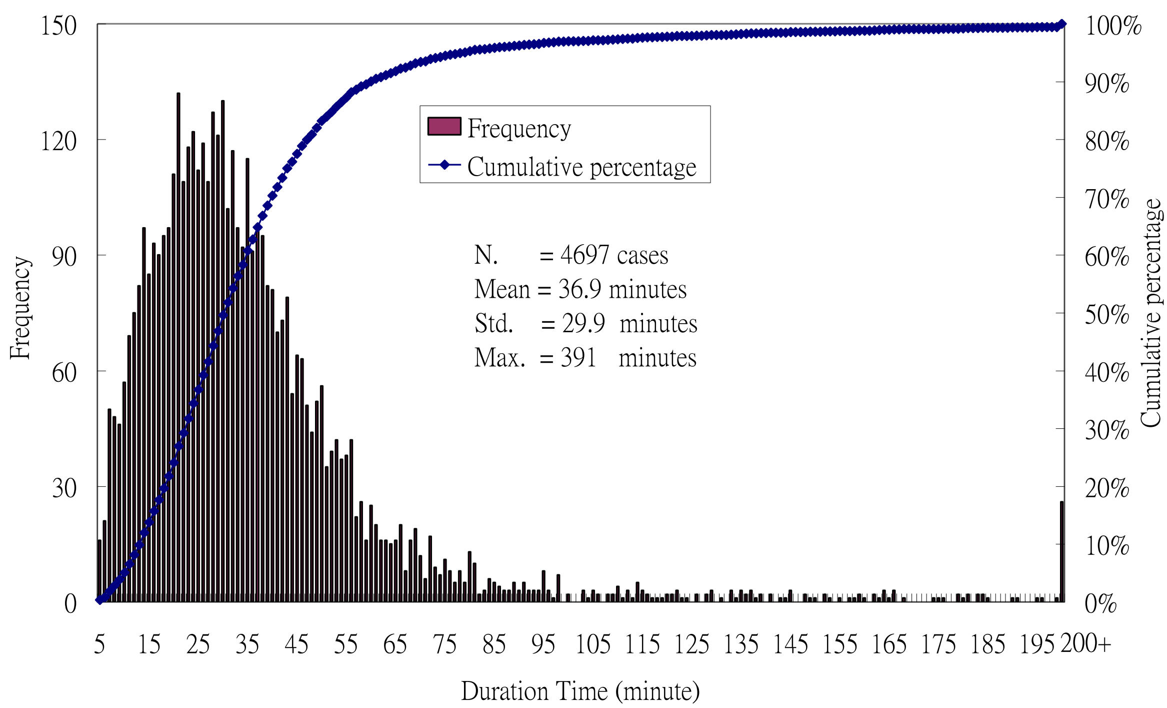


Figure 2. Distribution of incident duration data

According to Sharma (1996), the recommended process is that hierarchical analysis is followed by nonhierarchical clustering (K-means). The data processing by SPSS 10.0 software in this study is summarized bellow. First, the data during January, including 270 cases, were performed hierarchical analysis based on Ward’s method. As illustrated in Figure 3 there is a big change in the values when going from a two-cluster to a five-cluster solution. Next, all the data of 4,697 cases were performed nonhierarchical clustering based on K-mean algorithm. As listed in Table 1., because the duration in first cluster of two-cluster scenario ranges from 5 minutes to 95 minutes, it includes most of cases (i.e. 96.7%) that the two-cluster scenario is not appropriate to differentiate significantly the length of duration in practical application. In addition, the number of incidents of the last one cluster both in four-cluster and five-cluster scenarios is too small to be representative. Consequently, the three-cluster scenario is employed in the following analysis.

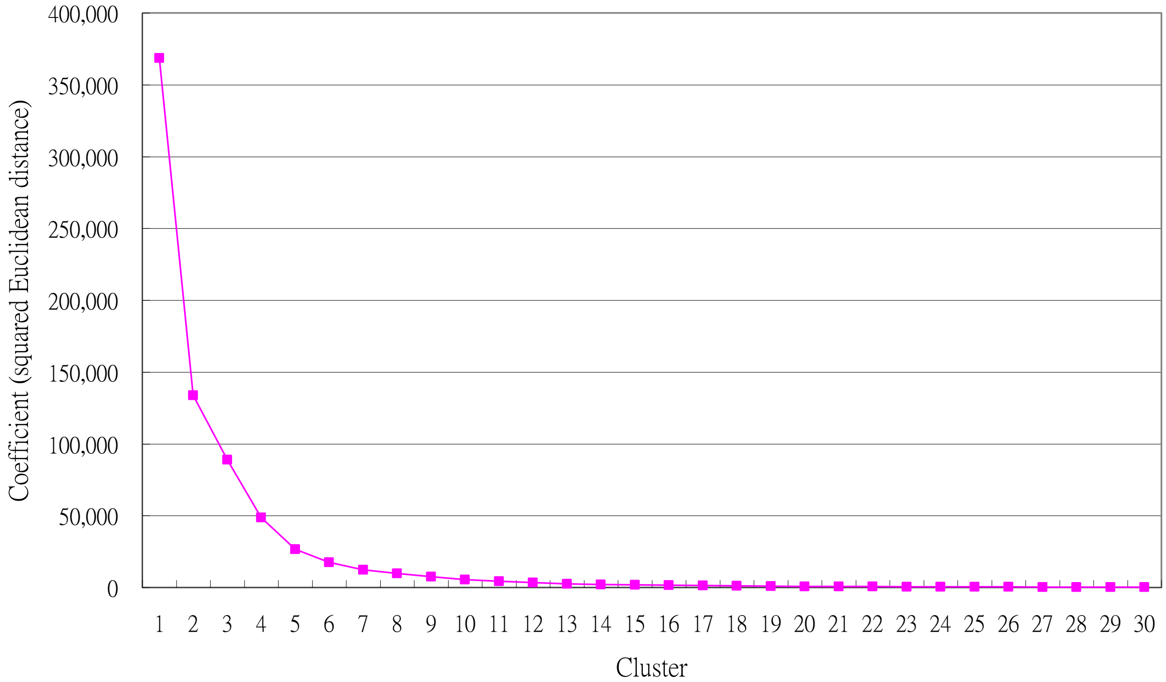


Figure 3. The change of coefficient by the number of cluster

Table 1. Distribution of incident duration by 2-5 clusters

Number of cluster	Statistics	Cluster 1	2	3	4	5	ALL
2	Min. (minutes)	5	96	-	-	-	5
	Max. (minutes)	95	391	-	-	-	391
	Number (cases)	4,542	155	-	-	-	4,697
3	Min. (minutes)	5	42	119	-	-	5
	Max. (minutes)	41	118	391	-	-	391
	Number (cases)	3,371	1,218	108	-	-	4,697
4	Min. (minutes)	5	39	92	200	-	5
	Max. (minutes)	38	91	199	391	-	391
	Number (cases)	3,138	1,387	146	26	-	4,697
5	Min. (minutes)	5	32	59	113	217	5
	Max. (minutes)	31	58	112	216	391	391
	Number (cases)	2,433	1,755	388	101	20	4,697

- Not relevant.

The results of discretization and the distribution of incident duration by all variables are shown in Table 2. The proportion of long and medium duration in the dead of night are higher than the other time of day, and the possible reasons may be that it takes more time to clear the incident under insufficiency of optics or to dispatch some emergency resources (i.e., specially equipped vehicles) in the dead of night. Additionally, there are higher percentages of long and medium duration in the incidents with single-vehicle or more than three vehicles, while the incidents with two or three vehicles have higher percentages of short duration, because they usually belong to serious incidents and require more emergency resources or specially equipped vehicles, e.g. collision with an obstacle, run-off roadway, and large-scale incidents. Other interesting findings depicted in Table 2 reveal that there is higher proportion of long and medium duration in the incidents having truck with trailer involved, large-size vehicle, cargo spill, fatal and injured respectively. In brief, the results of descriptive statistics in Table 2 are in line with the general observations and previous studies in traffic safety analysis.

As proposed by Breiman et al. (1988) and Han and Kamber (2001), the analysis sample has to be divided into training set and testing set, which are drawn from the same distribution. Generally, the proportion of learning set and testing set drawn from the same sample is 2/3 and 1/3 respectively. Because the incident cases in this study were collected sequentially in time, we took one month every season for the testing set (i.e. February, May, August, and November) and the other months were for the training set. There are 3,245 and 1,452 cases in training and testing sets respectively, which the percentages are 69.1% and 30.9% of the total cases respectively. The results of χ^2–test (χ=0.442, P=0.802) reveal that there were no statistically significantly difference between the training and testing sets.

3. METHOD

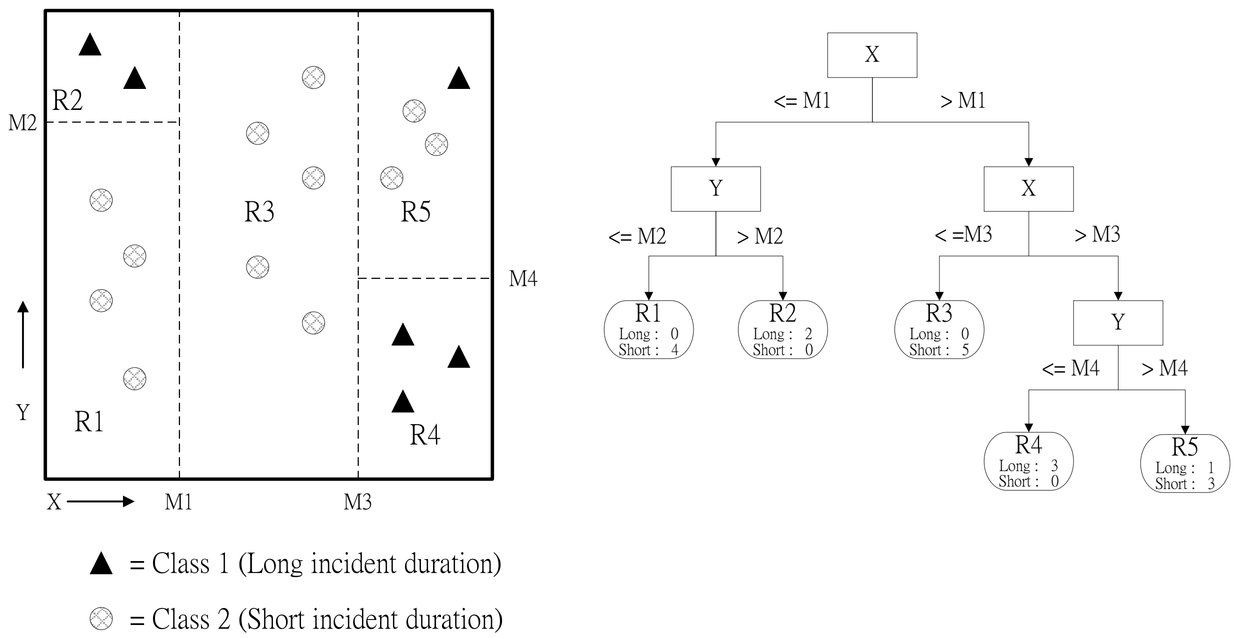
As frequently pointed out by Chang and Wang (2006), CART is one of the most commonly applied data mining techniques in many fields, such as business administration, agriculture, medicine, industry, and engineering. When the value of the target variable is continuous, a regression tree is developed, while a classification tree is built for the discrete target variable. Because the target variable in this study has huge variations, we use cluster analysis to transform the continuous target variable into a categorical variable (i.e., short-duration, medium-duration, and long-duration). Then, a classification tree is developed in this study.

There are great detailed explanation of the theory and analytic process for classification tree analysis in previous research or books (e.g. Breiman et al., 1984; Zeng et al., 2005; Chang and Wang, 2006; Tan et al., 2006) so this section illustrates briefly the concept in the following terms. Figure 4 shows an example of a classification tree and its decision boundaries, and depicts the concept of tree-growing procedure with two variables including x and y. First, the classification tree spilt the cases into two groups by x variable according to whether or not the x variable is more than M1 (one with x >M1 and one with x ≦M1). Second, the left group not more than M1 is further split into two groups by y variable (one with y >M2 and one with y ≦M2), whereas the right group more than M1 is further split into two groups by x variable (one with M1< x ≦M3 and one with x >M3). Third, the group (x >M3) is further split into two groups by y variable (one with M1< y ≦M4 and one with y >M4). Finally, a classification tree splits the cases recursively into five regions that are as homogenous as possible. Additionally, in order to avoid model overfitting, it is generally followed by a tree-pruning step based on the relationship between the misclassification cost and tree complexity (i.e. the number of terminal nodes), which is analogous to stepwise backward regression. Figure 5 illustrates the analysis process in this study.

Table 2. Distribution of magnitude of incident duration by all variables

Independent variable	Short (5-41 minutes)	Medium (42 – 118 minutes)	Long (119-391 minutes)
	3371 (72%)	1218 (26%)	108 (2%)
Notification Time			
0	42 (71%)	14 (24%)	3 (5%)
1	25 (56%)	15 (33%)	5 (11%)
2	20 (63%)	7 (22%)	5 (16%)
3	9 (39%)	7 (30%)	7 (30%)
4	13 (48%)	11 (41%)	3 (11%)
5	22 (41%)	24 (44%)	8 (15%)
6	53 (70%)	19 (25%)	4 (5%)
7	179 (72%)	68 (27%)	3 (1%)
8	201 (72%)	70 (25%)	7 (3%)
9	208 (75%)	64 (23%)	4 (1%)
10	250 (75%)	81 (24%)	4 (1%)
11	185 (72%)	64 (25%)	7 (3%)
12	138 (61%)	84 (37%)	4 (2%)
13	143 (70%)	56 (27%)	5 (2%)
14	228 (74%)	76 (25%)	3 (1%)
15	231 (74%)	76 (24%)	4 (1%)
16	241 (70%)	97 (28%)	5 (1%)
17	272 (75%)	85 (23%)	5 (1%)
18	373 (79%)	90 (19%)	9 (2%)
19	236 (76%)	70 (23%)	3 (1%)
20	114 (69%)	50 (30%)	2 (1%)
21	95 (71%)	37 (28%)	2 (1%)
22	66 (69%)	28 (29%)	2 (2%)
23	27 (48%)	25 (45%)	4 (7%)
Number of vehicles involved			
1	1121 (64%)	544 (31%)	80 (5%)
2	1520 (79%)	386 (20%)	18 (1%)
3	512 (75%)	168 (24%)	7 (1%)
4	144 (69%)	64 (30%)	2 (1%)
5 +	74 (56%)	56 (43%)	1 (1%)
Truck with trailer involved			
No	3244 (74%)	1099 (25%)	52 (1%)
Yes	127 (42%)	119 (39%)	56 (19%)
Number of large-size vehicles (Except small vehicle)			
0	2993 (76%)	924 (23%)	16 (0%)
1	321 (52%)	223 (36%)	78 (13%)
2	52 (42%)	60 (49%)	11 (9%)
3	5 (29%)	9 (53%)	3 (18%)
4	0 (0%)	2 (100%)	0 (0%)
Peak Time			
1. Peak	1505 (69%)	612 (28%)	71 (3%)
2. Off-Peak	1866 (74%)	606 (24%)	37 (1%)
Cargo Spilled			
No	3349 (72%)	1194 (26%)	91 (2%)
Yes	22 (35%)	24 (38%)	17 (27%)

Independent variable	Short (5-41 minutes)	Medium (42 – 118 minutes)	Long (119-391 minutes)
	3371 (72%)	1218 (26%)	108 (2%)
Number of lanes affected			
0	182 (65%)	74 (26%)	26 (9%)
1	2842 (75%)	896 (24%)	54 (1%)
2	342 (56%)	245 (40%)	27 (4%)
3	5 (63%)	2 (25%)	1 (13%)
4	0 (0%)	1 (100%)	0 (0%)
Location			
1.Travel Lane & shoulder	3189 (72%)	1144 (26%)	82 (2%)
2.Ramp	155 (69%)	54 (24%)	15 (7%)
3.Outside of Road	27 (47%)	20 (34%)	11 (19%)
Incident Type			
1.Rear-end	2200 (76%)	651 (23%)	30 (1%)
2.Overturn	409 (57%)	267 (37%)	37 (5%)
3.Disabled vehicle	5 (42%)	4 (33%)	3 (25%)
4.Vehicle Fire or smoke	85 (63%)	45 (33%)	6 (4%)
5.Collision with Tollbooth	5 (42%)	6 (50%)	1 (8%)
6.Run-off-roadway	12 (48%)	10 (40%)	3 (12%)
7.sideswipe	95 (75%)	26 (21%)	5 (4%)
8.Collision with barrier	425 (75%)	135 (24%)	10 (2%)
9.Collision (opposing direction)	0 (0%)	4 (80%)	1 (20%)
10. others	135 (62%)	70 (32%)	12 (6%)
Number of persons (fatal or injured)			
0	3178 (75%)	1007 (24%)	74 (2%)
1	162 (50%)	144 (44%)	21 (6%)
2	21 (31%)	35 (52%)	11 (16%)
3 +	10 (23%)	32 (73%)	2 (5%)
Route			
1. N1H	43 (74%)	15 (26%)	0 (0%)
2. N1	2065 (71%)	771 (27%)	68 (2%)
3. N2	68 (66%)	35 (34%)	0 (0%)
4. N3	1067 (74%)	333 (23%)	35 (2%)
5. N4	11 (50%)	9 (41%)	2 (9%)
6. N5	5 (21%)	17 (71%)	2 (8%)
7. N6	1 (33%)	2 (67%)	0 (0%)
8. N8	26 (93%)	2 (7%)	0 (0%)
9. N3A	20 (95%)	1 (5%)	0 (0%)
10. N10	65 (66%)	33 (33%)	1 (1%)
Open Shoulder For Temporary Driving			
No	3189 (72%)	1142 (26%)	107 (2%)
Yes	182 (70%)	76 (29%)	1 (0%)
Region			
1. North	1447 (76%)	425 (22%)	36 (2%)
2. Central	814 (67%)	364 (30%)	41 (3%)
3. South	1110 (71%)	429 (27%)	31 (2%)



(a) Decision boundaries **(b) A decision tree**

Figure 4. A classification tree and its decision boundaries for a two-dimension dataset

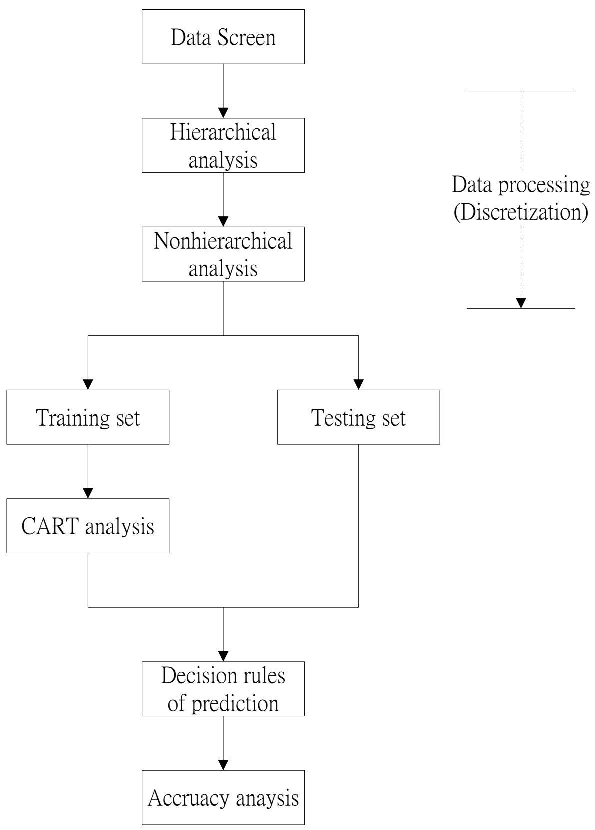


Figure 5. The anaysis process of this study

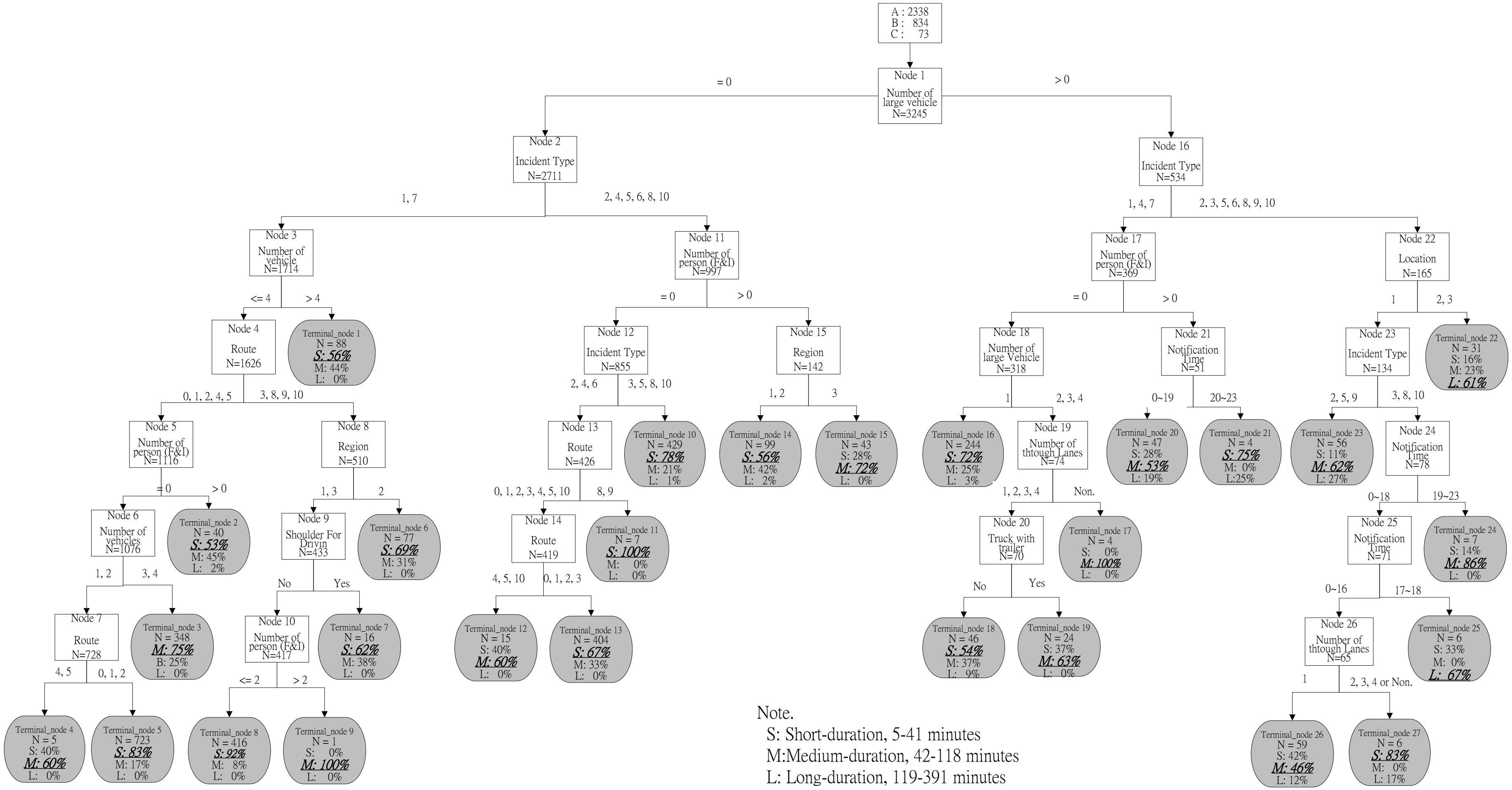
4. ANALYSIS RESULTS

This study performed classification analysis using SAS v.10 software (Enterprise Miner) by CART's default, such as Gini splitting criterion. The results of the classification tree model with thirteen predictor variables and a three-level target variable (i.e. length of incident duration) were shown in Figure 6. Twenty-seven terminal nodes provide twenty-seven decision rules for predicting the level of incident duration, including fourteen short-duration rules, eleven medium-duration rules, and two long medium-duration rules. The primary splitters denote the important variables in predicting incident duration, which the importance of the variables is in order from top to down in the tree. The most important variable (i.e. number of large-size vehicles) is the initial split at Node 1, the second important variable is incident type and so forth.

From the decision tree in Figure 6, the interpretation of decision rule is straightforward and this can be seen in the following examples. As shown by terminal node 1, if the number of large-sized vehicles is zero, incident type is rear-end or sideswipe and the number of vehicles is more than four, then the level of incident duration is most likely to be short (56%). At terminal node 15, if the number of large-sized vehicles is zero, incident type is overturn, disabled vehicle, vehicle fire or smoke, collision with tollbooth, run-off-roadway, collision with barrier or others, at least one person is fatal or injured, and the region is south, then the level of incident duration is most likely to be medium (72%). At terminal node 22, if there is at least one large-size vehicle, incident type is overturn, disabled vehicle, collision with tollbooth, run-off-roadway, collision with barrier, collision (opposing direction) or others, and the location is ramp or outside of road, then the level of incident duration is most likely to be long (61%).

In addition, the most two important variables are number of large-sized vehicles and incident type but the third important variables in different subtrees are not the same (i.e. number of vehicles, number of fatal and injured persons, location). In order to provide parsimonious representation and practical implication, this study used the most two important variables to divide the tree into four subtrees. The results of four subtrees divided from right to left in the full tree (i.e. Figure 6) are listed in Table 3. From the first subtree to fourth subtree, the proportion of short-duration decreased gradually, whereas the total proportion of medium-duration and long-duration increased. The first subtree (i.e. terminal node 1-9), which the number of large-sized vehicles is zero and incident type is rear-end or sideswipe, has the largest proportion of short-duration and the smallest total proportion of medium-duration and long-duration. On the other hand, the fourth subtree (i.e. terminal node 22-27), which there is at least one large-sized vehicle and incident type is overturn, disabled vehicle, collision with tollbooth, run-off-roadway, collision with barrier, collision (opposing direction) or others, has the largest total proportion of medium-duration and long-duration and the smallest proportion of short-duration.

The performance of the classification tree model is shown in Table 4. The overall accuracy of classification for training data is about 75.1%, while it is about 73.6% for testing data. Although the model has high accurate classification for short-duration (more than 96% both for the training and testing sets), the accuracy of medium-duration and long-duration are relatively lower. As noted by Chang and Wang (2006) exploring traffic injury severity, the possible reason is the imbalance dataset (the proportion of short-duration dataset is as high as 72%) and then the largest percentage of cases may give rise to a higher proportion of accurate classification. Under the limitations above, it should be careful for further implication and assessment of performance on medium-duration and long-duration.



Note.
S: Short-duration, 5-41 minutes
M:Medium-duration, 42-118 minutes
L: Long-duration, 119-391 minutes

Figure 6. Structure of the output of decision tree

Table 3. The subtrees by the most two important variables

Sub -tree	Terminal node	The number of terminal nodes			The number of cases (row percentage)							
		Short	Medium	Long	Short		Medium		Long		Total	
1	1-9	6	3	0	1,380	(81%)	333	(19%)	1	(0%)	1,714	(100%)
2	10-15	4	2	0	689	(69%)	304	(30%)	4	(0%)	997	(100%)
3	16-21	3	3	0	225	(61%)	123	(33%)	21	(6%)	369	(100%)
4	22-27	1	3	2	144	(54%)	75	(28%)	46	(17%)	265	(100%)

Table 4. Analysis results of the classification tree model

	Training data (N =3,245)				Testing data (N = 1,452)			
	Observed	Predicted	Correctly predicted		Observed	Predicted	correctly predicted	
Short-duration (5-41 minutes)	2,338	2,947	2,257	(96.5%)	1,033	1,326	999	(96.7%)
Medium-duration (42 – 118 minutes)	834	261	156	(18.7%)	384	117	66	(17.2%)
Long -duration (119-391 minutes)	73	37	23	(31.5%)	35	9	4	(11.4%)
Total	3,245	3,245	2,436	(75.1%)	1,452	1,452	1,069	(73.6%)

5. DISCUSSION AND CONCLUSION

This study used large data, which was recorded correctly, from Taiwan freeway systems to develop predicting model of incident duration based on classification tree analysis. The model with reasonable ability in this study indicates that it is an appropriate method for predicting the level of incident duration. Despite the large data and non-traditional statistical model compared to previous studies, the present study is not without limitations. From an overview of the analysis in this study, the following discussion and future recommendations can be drawn:

1. Although the classification tree model has reasonable overall predicting ability, the accuracy of predicting the level of medium and long is obviously lower. Besides the effect of imbalance dataset, another possible reason is that the variables on hand can't completely explain the variety of the target variable. For example, the weather condition might affect the efficiency of clearing incident, the upstream traffic could delay the coming rescuers, and response time would be affected by the distance between the incident and the standby rescuer. In the future, the traffic management center should record other factors influencing incident duration and further research might consider some of the above variables into the analysis. Besides, future studies may try two potential methods to improve problem above, e.g. equal-sampling across the incident duration levels or weighting for the prediction accuracy of the incident duration levels.
2. Compared to other studies predicting the incident duration with continuous variable, this study perform data processing of discretization before developing model because this large dataset with an extreme right-skewed distribution is not well suited to following analysis, while other studies with relatively smaller dataset don't have the similar problem. Because the three-level target variable has some limitations for implication, further exploration can combine other predicting methods (e.g. ANN) or multistage predicting process for continuous target variable. For example, the first stage is to employ

the method of this study and the second stage is to develop different ANN models with the continuous target variable from the three level datasets respectively. Therefore, the classification tree model can regard as complement to other predicting methods.

Compared to other studies using traditional statistical model with some strict assumptions and finite dataset, this study presents a predicting model of freeway incident duration using classification tree analysis, which the correctly recorded and large dataset from traffic management center consists of most of the incidents on freeway systems in Taiwan. From a methodological standpoint, this is an attempt to analyze the complex subject about incident duration by a non-traditional statistical model. Although there is acceptable reasonableness on overall predicting accuracy, the model still has some limitations such as lower predicting accuracy on medium-duration and long duration incidents. Thus it is a potential topic for further research. In addition, the classification tree model has another advantage, which is the disadvantage of the traditional regression model, is to handle multi-collinearity problems (e.g., number of large-size vehicles involved and cargo spill). Although the incidents with at least one large-size vehicle or cargo spilled has longer duration, the analysis result show that the variable of cargo spilled is not important variables in predicting model. Therefore, the multi-collinearity variables in classification tree analysis would not be a great concern.

From an empirical viewpoint, the analysis result of classification tree could regard as the decision rules or incorporate into an expert system, and then the traffic operator (engineer) can predict more accurately the incident duration for following implication. For advanced traveler information systems (ATIS), the motorists can change their travel route immediately based on the real-time and more accurate incident information, while in the past they just perceive the downstream incident information without recognizing the effect. For advanced traffic management systems (ATMS), when traffic management center are notified about the incident, they can acquire the level of incident duration by entering the input variables (predictors) into the predicting model, and then do the following steps on traffic management and incident management. For example, according to the prediction of incident duration, the traffic management center can provide more accurately incident duration for motorists, make real-time and rational traffic diversion strategies, and dispatch effectively rescuers. Ultimately, motorists would avoid a traffic jam and traffic operator could mitigate traffic congestion due to incidents. In addition, another interesting finding is that the special incidents have longer incident duration, including at least one large-size vehicle involved and incident type being overturn, disabled vehicle, collision with tollbooth, run-off-roadway, collision with barrier, collision (opposing direction) or others. Therefore, the incident management engineers could plan appropriate rescuers (e.g. specially equipped vehicles) near the road sections with frequent similar incident beforehand; moreover, the traffic management operators could dispatch suitable rescuers immediately while the special incident occurred. Thus the duration of the special incident can be reduced and the traffic congestion can also be lessened.

In conclusion, although the classification tree model can not supplant to previous predicting methods of incident duration, it can present the important variables influencing incident duration and predict the level of duration. Hence, the motorists and traffic management organization can perceive accurately the effect of incident and make decisions timely to mitigate the traffic congestion. In the future, further research should consider more variables and combine other predicting methods for providing more accurate prediction.

ACKNOWEDGMENTS

This data used in this study are provided by Taiwan area National Freeway Bureau.

REFERENCES

Bevilacqua, M., Ciarapica, F.E., Giacchetta, G. (2008) Industrial and Occupational ergonomics in the petrochemical process industry: A regression tress approach. *Accident Analysis & Prevention*, 40 (4), 1068-1479.

Breiman, L., Friedman, J.H., Olshen, A.R., Stone, C.J. (1984) *Classification and Regression Tree*. Chapman and Hall, New York.

Chang, L.Y., Wang, H.W. (2006) Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, 38 (5), 1019-1027.

Chung, Y. (2010) Development of an accident duration prediction model on the Korean Freeway Systems. *Accident Analysis & Prevention*, 42 (1), 542-555.

Garib, A., Radwan, A.E. Al-Deek, H., (1997) Estimating magnitude and duration of incident delays. *Journal of Transportation Engineering*, 123 (6), 459-466.

Giuliano, G. (1988) Incident characteristics, frequency, and duration on a high volume urban freeway. *Institute of Transportation Studies*, University of California, Irvine, Calif..

Giuliano, G. (1989) Incident characteristics, frequency, and duration on a high volume urban freeway. *Transportation Research 23A (5)*, 387-396.

Golob, T., Recher, W., Leonard, J. (1987) An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis & Prevention*, 19 (5), 375-395.

Han, J., Kamber, M. (2001) *Data Mining- Concepts and Techniques*. Morgan Kaufmann Publishers, New York.

Jones, B., Janseen, L., Mannering, F. (1991) Analysis of the frequency and duration of freeway accidents in Seattle. *Accident Analysis & Prevention*, 23 (4), 239-255.

Kuhnert, P.M., Do, K.A., McClure, R. (2000) Combining non-parametric models with logistic regression: an application to motor vehicle injury data. *Computational Statistics & Data Analysis*, 34 (3), 371-386.

Nam, D., Mannering, F. (2000) An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice* 34 (2), 85-102.

Ozbay, K., Kachroo, P. (1999) *Incident Management in Intelligent Transportation systems*. Artech House, Boston, MA.

Ozbay, K., Noyan, Nebahat. (2006) Estimation of incident clearance times using Bayesian Networks approach. *Accident Analysis & Prevention*, 38 (3), 542-555.

Piramuthu, S. (1999) Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of the Operational Research* 112, 310-321.

Simith, B.L., Smith, K. (2001) Incident Duration Forecasting. Forecasting the clearance Time of Freeway Accidents Research Report No. UVACTS-15-0-35. USDOT University Transportation Center.

Sharma, S. (1996) *Applied multivariate techniques*. Wiley. New York.

Sohn, S.Y., Shin, H. (2001) Pattern recognition for road traffic accident severity in Korea. *Ergonomics* 44 (1), 107-117.

Sohn, S.Y., Lee, S.H. (2003) Data fusion, ensemble and clustering to improve the classification accuracy for severity of road traffic accidents in Korea. *Safety Science*. 41 (1), 1-44.

Tan, P.-N., Steinbach, M., Kumar, V. (2006) *Introduction to Data Mining*. Pearson Education and Addison Wesley, New York.

TRB, (1994) Highway capacity manual: special report 209. Transportation Research Board, National Research Council. Washington, D.C.

Wei. C.H., Lee, Y. (2007) Sequential forecast of incident duration using Artificial Neural Network models. *Accident Analysis & Prevention*, 23 (2), 239-255.

Zeng, S.S., Cai, S.M., Su, D.S., Zent, C.R., Wang, C.Y. (2005) *Data Mining*. Flag Publishing Co., Ltd. Teipei.