Analyses of Rear-End Crashes Based on Classification Tree Models

Xuedong Yan\textsuperscript{a}; Essam Radwan\textsuperscript{b}

\textsuperscript{a} Department of Civil & Environmental Engineering, University of Central Florida, Orlando, Florida, USA \textsuperscript{b} Center for Advanced Transportation Systems Simulation, University of Central Florida, Orlando, Florida, USA

To cite this Article Yan, Xuedong and Radwan, Essam(2006) 'Analyses of Rear-End Crashes Based on Classification Tree Models', Traffic Injury Prevention, 7: 3, 276 — 282

To link to this Article DOI: 10.1080/15389580600660062
URL: http://dx.doi.org/10.1080/15389580600660062
Analyses of Rear-End Crashes Based on Classification Tree Models

XUEDONG YAN
Department of Civil & Environmental Engineering, University of Central Florida, Orlando, Florida, USA

ESSAM RADWAN
Center for Advanced Transportation Systems Simulation, University of Central Florida, Orlando, Florida, USA

Objective. Signalized intersections are accident-prone areas especially for rear-end crashes due to the fact that the diversity of the braking behaviors of drivers increases during the signal change. The objective of this article is to improve knowledge of the relationship between rear-end crashes occurring at signalized intersections and a series of potential traffic risk factors classified by driver characteristics, environments, and vehicle types.

Methods. Based on the 2001 Florida crash database, the classification tree method and Quasi-induced exposure concept were used to perform the statistical analysis. Two binary classification tree models were developed in this study. One was used for the crash comparison between rear-end and non-rear-end to identify those specific trends of the rear-end crashes. The other was constructed for the comparison between striking vehicles/drivers (at-fault) and struck vehicles/drivers (not-at-fault) to find more complex crash pattern associated with the traffic attributes of driver, vehicle, and environment.

Results. The modeling results showed that the rear-end crashes are over-presented in the higher speed limits (45–55 mph); the rear-end crash propensity for daytime is apparently larger than nighttime; and the reduction of braking capacity due to wet and slippery road surface conditions would definitely contribute to rear-end crashes, especially at intersections with higher speed limits. The tree model segmented drivers into four homogeneous age groups: <21 years, 21–31 years, 32–75 years, and >75 years. The youngest driver group shows the largest crash propensity; in the 21–31 age group, the male drivers are over-involved in rear-end crashes under adverse weather conditions and the 32–75 years drivers driving large size vehicles have a larger crash propensity compared to those driving passenger vehicles.

Conclusions. Combined with the quasi-induced exposure concept, the classification tree method is a proper statistical tool for traffic-safety analysis to investigate crash propensity. Compared to the logistic regression models, tree models have advantages for handling continuous independent variables and easily explaining the complex interaction effect with more than two independent variables. This research recommended that at signalized intersections with higher speed limits, reducing the speed limit to 40 mph efficiently contribute to a lower accident rate. Drivers involved in alcohol use may increase not only rear-end crash risk but also the driver injury severity. Education and enforcement countermeasures should focus on the driver group younger than 21 years. Further studies are suggested to compare crash risk distributions of the driver age for other main crash types to seek corresponding traffic countermeasures.

Keywords Rear-End Crashes; Signalized Intersections; Crash Database; Quasi-Induced Exposure; Classification Tree; Entropy Algorithm

Rear-end crashes are one of the frequently occurring types of crashes, accounting for almost one-third of all reported crashes in the US and 11.8 percent of multivehicle fatal crashes (NTSB, 2001). Moreover, signalized intersections are accident-prone areas especially for rear-end crashes due to the diversity of the braking behaviors of drivers increases due to signal change. In order to develop effective rear-end crash countermeasures it is important to use the crash database and appropriate statistical model to explore the overall characteristics of rear-end crashes.

Generally, traffic crashes are comprehensively associated with numerous risk factors classified by driver characteristics, vehicle types, and environmental features. The driver age and gender were considered as main human factors that might be associated with a rear-end crash. Slower reaction times for older versus younger drivers contribute to a disproportionately heightened degree of risk especially when older drivers are faced with two or more choices of action (Staplin et al., 2001). However, the younger drivers are more likely to drive aggressively and...
inattentively. A previous study on rear-end crashes indicated that drivers younger than 18 years were most vulnerable to roadway crashes, followed by 18- to 24-year-old drivers, and the older drivers showed a higher crash involvement propensity as compared to the drivers who were 25 to 69 years-old (Santokh, 2003). For different type of vehicles, steering and braking performance are critical in the avoidance of crashes. Strandberg (1998) pointed out that differences between vehicles in braking performance are responsible for many rear-end crashes. Furthermore, the critical environmental conditions may play a significant role in rear-end crashes, such as lighting conditions, the roadway surface conditions, highway characteristics, the weather conditions, and so on. For example, in icy and snowy road surface conditions vehicle deceleration capacity may decrease by more than 90% compared to dry condition (Strandberg, 1998).

To explore the traffic accident database, the Quasi-induced exposure technique (Carr, 1970; Haight, 1973) is a proper method used for analyses of crash risk factors. In quasi-induced exposure concept drivers in the crash database are identified and classified into the at-fault group and not-at-fault group, respectively. The at-fault drivers are those who were mostly responsible for the accident occurrence and the not-at-fault drivers are those victims in the accidents. Further, it was assumed that the distribution of not-at-fault drivers closely represents the distribution of all drivers exposed to accident hazards. Thus, the comparison between the attributes of those at-fault drivers and not-at-fault drivers can be used to measure the crash propensities related to those traffic risk factors, such as driver characteristics, vehicle types, and environmental features. To identify the statistical significance of the risk factors, the binary logistic regression was applied by previous studies (Stamatiadis and Deacon, 1995; Hing et al., 2003) because the dependent variable (at-fault or not-at fault drivers in crashes) is a typical dichotomy variable. Although the logistic regression is a proper method for the quasi-induced exposure analyses, it is hard to investigate the complex interaction effect between more than two independent variables.

Therefore, in this study classification tree (or decision tree) models were used to investigate characteristics of rear-end crashes occurring at signalized intersections. The classification tree is a nonparametric model that does not require a functional form to be specified and any additive assumption of independent variables. The advantage of classification trees over many of the other methods used here is to construct classification of those crashes by dividing the data set into smaller and more homogeneous groups. The tree models are helpful to find the complex crash pattern based on traffic attributes of driver, vehicle, and environment. However, the applications of classification trees to analyze traffic-safety problems have been relatively few, especially on the crash risk analyses based on the quasi-induced exposure concept. The objective of this study is to improve knowledge of the relationship between risk factors and rear accidents based on the quasi-induced exposure concept and classification tree models.

METHOD

Crash Database

The 2001 Florida crash database was used in this study. The database includes seven files and each file deals with a specific of a traffic crash. Files may be linked to each other as needed to combine the information contained. Three files used in the analysis presented here were the event (containing the characteristics and environment of the accident), drivers (containing the drivers’ characteristics), and vehicles (information about the vehicles’ characteristics and vehicles actions in the traffic crash) files. Rear-end accident scenarios may involve three or more vehicles. To simplify the assignment of driver culpability and easily identify accident roles of vehicle (striking or struck) in the crashes the analysis in this study was restricted to two-vehicle collisions in which both struck and striking vehicles proceeded straight at signalized intersections.

Furthermore, the vehicles/drivers involved in rear-end crashes can be split into striking group and struck group respectively. In a rear-end crash the struck driver/vehicle more likely had no improper action. On the other hand, the striking vehicle more likely took responsibility for the traffic accident and the main contributing factors are careless driving and following too closely. Thus, the attributes of the struck drivers/vehicles can be used as exposure information to compare the attributes of the striking drivers/vehicles, so as to find the propensity of drivers/vehicles resulting in rear-end crashes. This analysis method is based on Quasi-induced exposure concept. Here, the striking vehicles/drivers can be considered as the at-fault ones and the struck vehicles/drivers are not-at-fault ones.

Classification Tree

For a classification tree, the target variable takes its values from a discrete domain and for each leaf node the classification tree associates a probability (and in some cases a value) for each class (i.e., value of the target variable). The tree-based classification algorithms have tree structures consisting of nodes (or leaves), branches, etc. In tree-structured representations, a set of data is represented by a node and the entire dataset is represented as a root node. When a split is made, two child nodes are formed that correspond to partitioned data subsets. If a node is not to be split any further, it is called a terminal node that is associated with a group membership; otherwise, it is an internal node. The tree structure is constructed following a set of decision rules applied sequentially. Each decision rule is used to form branches (i.e., splitting) connecting the root node to the terminal node at a certain level of the tree.

In this study, the Enterprise Miner 4.1 in SAS 8.2 (SAS Institute, Cary, NC) was used to develop the classification tree based on the reduction of entropy \((i_t)\) as a measure of split criteria. The entropy function to measure node impurity is shown in Eq. (1):

\[
    i_t = -p_t \log(p_t) - (1 - p_t) \log(1 - p_t)
\]  

where \(p_t\) is the proportion of at-fault drivers in a specified node \(t\). This function is at its lowest level when \(p_t = 0\) or 1. In other
Table I  Input variables into two tree models

<table>
<thead>
<tr>
<th>Variables</th>
<th>Measurement</th>
<th>Variables</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target (rear-end or not)</td>
<td>Binary</td>
<td>Target (Striking or struck)</td>
<td>Binary</td>
</tr>
<tr>
<td>Divided/undivided highway</td>
<td>Binary</td>
<td>Alcohol/drug use (driver)</td>
<td>Nominal</td>
</tr>
<tr>
<td>Lighting condition</td>
<td>Binary</td>
<td>Physical defect</td>
<td>Nominal</td>
</tr>
<tr>
<td>Day of week</td>
<td>Binary</td>
<td>Residence code</td>
<td>Nominal</td>
</tr>
<tr>
<td>Rural/urban</td>
<td>Binary</td>
<td>Driver gender</td>
<td>Binary</td>
</tr>
<tr>
<td>Location type</td>
<td>Nominal</td>
<td>Driver age</td>
<td>Interval</td>
</tr>
<tr>
<td>Road surface condition</td>
<td>Nominal</td>
<td>Vehicle type</td>
<td>Nominal</td>
</tr>
<tr>
<td>Weather condition</td>
<td>Nominal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Highway character</td>
<td>Nominal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of shoulder</td>
<td>Nominal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speed limit</td>
<td>Interval</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of lanes</td>
<td>Ordinal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crash injury severity</td>
<td>Ordinal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alcohol/drug involved</td>
<td>Nominal</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Besides above variables in the model #2, those environmental variables in the model #1 are also input into the model #2 to check if there is interactive effect between traffic environments and driver/vehicle characteristics.

words, there is the least impurity when the node is perfect. On the other hand, it reaches the maximum when \( p_t = 1/2 \), that is, the node is equally mixed with target variable’s components.

The entropy algorithm begins with the entire set of data to split the data into two subsets based on the attribute with the lowest entropy value pointing to the most informative attribute among a set of attributes and then repeatedly splits until the size of each subset reaches an appropriate level. Two previous articles (Strnad et al., 1998; Vorko & Jovic, 2000) clearly presented the entropy method and related details for decision trees.

The generation of a tree model involves partitioning the model data set into at least two parts: two-thirds of the available data as the training sample and one-third of the data randomly selected as the validation sample. Misclassification rates (default in SAS Miner) are used as the model assessment measure to choose the tree with the smallest misclassification rate, which is the total number of misclassified points divided by the total number of data points. There are two major phases of the tree generation process: the growth phase and the pruning phase. In the growth phase, misclassification rates based on training sample are decreasing monotonically as the number of nodes increase until the purity of a subgroup cannot be further improved. This procedure results in a very large tree that will usually perform poorly if a new observation is classified. To avoid over-fitting, the pruning phase aims to generalize the tree model developed in the growth phase. In the pruning phase, the tree model is evaluated against the validation sample in order to generate a subtree of the large tree generated in the growth phase. As the node number increases, the misclassification rates based on the validation sample will decrease first and then increase when the tree model is over-fitting. The minimum misclassification rate is corresponding to the best size tree model. After the best tree size is identified, the tree structure model can be used to explain and analyze the relationship between the target variable and those important factors. More detailed descriptions of these decision tree algorithms are beyond the scope of the present study. For further discussions of tree methodology, the reader is referred to Breiman and colleagues (1984).

**TREE MODELS AND ANALYSES**

Two binary classification tree models were developed in this study. The input variables for the two tree models are listed in Table I. In the tree model #1 the dataset includes all two-vehicle crashes occurring at signalized intersections that can be categorized into two groups: rear-end crashes and non-rear-end crashes. The target variable is whether the crash is rear-end \( (Y = 1) \) or not \( (Y = 0) \). Thus, the crash comparison between rear-end and non-rear-end can be used to identify those specific trends of the rear-end crashes different from other crashes. In the tree model #2 the dataset includes only those rear-end crashes and the target variable is whether the role of the vehicle in a rear-end crash is striking \( (Y = 1) \) or struck \( (Y = 0) \). So, the comparison between striking vehicles/drivers (at-fault) and struck vehicles/drivers (not-at-fault) can be used to find more complex crash patterns associated with traffic attributes of driver, vehicle, and environment.

**Tree Model #1**

The dataset input to the Enterprise Miner consisted of 23,400 observations of two-vehicle crashes at signalized intersections, which include 7,666 rear-end crashes and 15,734 non-rear-end crashes. 15,678 of them were included in the training sample and 7,722 were randomly selected and set aside as the validation sample. The misclassification rates for both training sample and validation sample corresponding to the increasing tree size are shown in Figure 1. Based on the training sample the maximum tree was splitting up to yield the sequence of 42 terminal nodes and the misclassification rate tends to become stable as the number of terminal nodes is larger than 10. On the other hand, the
misclassification rate based on the validation data shows that to reach a minimum value (0.298) with 10 terminal nodes. For the best tree size the number of terminal nodes in the decision tree model #1 is determined to be 10.

Figure 2 shows the decision tree diagram for the crash comparison between rear-end and non-rear-end crashes based on the training data. In the node boxes, the top two lines give the node number and the total number of two-vehicle crashes entering the node; the third line shows the proportion of rear-end crashes in the number of two-vehicle crashes, which is also the conditional probability that rear-end crashes happen at the intersections. The diamond-shaped boxes indicate the most important variables used to split data, including speed limit, alcohol use, crash injury severity, road surface condition, and lighting condition.

From Figure 2, the classification tree firstly segments the data into two groups: lower speed limits (25–40 mph) and higher speed limits (45–55 mph). It shows that the risk of the rear-end crashes are overpresented in the higher speed limits since the conditional probability (0.43) in Node 3 is higher than the lower speed limits (0.257) in Node 2. At signalized intersections with the higher speed limits, generally, drivers are more likely to fall into the dilemma zone, where they possibly can neither execute the intersection crossing nor execute the stopping maneuver safely and comfortably at the onset of yellow. Such a negative situation may result in rear-end crashes due to relatively higher operation speeds (Zimmerman & Bonneson, 2004). Of those crashes occurring at higher speed limits, rear-end crashes are more likely to cause no injury (level 1) or possible injury (level 2) as shown in Node 6, compared to nonincapacitating evident injury (level 3), incapacitating injury (level 4), and fatal injury (level 5) as shown in Node 7. In the cases of no injury and possible injury the rear-end risk is over-presented when drivers were involved in alcohol use; as shown in node 11, the conditional probability could be as large as 0.726. For the drivers without alcohol use, the rear-end risk of possible injury \(P = 0.518\) is relatively higher than that of no injury \(P = 0.348\), as shown in Nodes 14 and 15. Further, an interesting finding is that the rear-end risk for the night condition \(P = 0.398\) in Node 19 is apparently lower than that for the daytime condition \(P = 0.555\) in Node 18. The presumable reasons are that the traffic volume...
at daytime is higher than that at night; the morning peak and afternoon peak may affect driving attitude and contribute to the crashes. In another track starting from Node 7, rear-end crashes are less likely contribute to incapacitating injury and fatal injury compared to the non-incapacitating evident injury, as shown in Node 13. If the road surface is wet (level 2) or slippery (level 3) in Node 17, the rear-end risk could be 1.54 times higher and cause more incapacitating injury than the dry road surface (level 1) in Node 16.

Of those crashes occurring at lower speed limits in Node 2 alcohol use could also contribute to a higher probability of rear-end crashes \( (P = 0.5) \) in Node 4, but they are more likely to result in no injury or possible injury. However, in lower speed limits the probability \( (P = 0.368) \) of evident injury, incapacitating injury, and fatal injury caused by rear-end crashes related to alcohol use are even slightly higher than that \( (P = 0.326) \) in higher speed limits. Generally, at the lower operation speed, the following drivers can more easily change lanes or brake to avoid striking the leading vehicle.

### Tree Model #2

For the tree model #2, the dataset input to Enterprise Miner consisted of 15,322 observations only from the rear-end crashes, which include 7,666 striking vehicles/drivers (coded as 1) and the same number of corresponding struck vehicles/drivers (coded as 0). Of those input dataset 10,272 cases were used for the training sample and the other data was selected randomly as the validation sample. As shown in Figure 3, the maximal tree has 29 terminal nodes and the misclassification rates based on the validation data show to reach a minimum value (0.362) that identified the best tree size with nine terminal nodes. Figure 4 illustrates the decision tree diagram for drivers/vehicles analyses based on the training data. In the node boxes, the second line gives total number of observations entering the node; the third line shows the proportion of striking vehicles/drivers, which can also be considered as the conditional probability of at-fault drivers as striking roles resulting in rear-end crashes.

Based on the first data split, alcohol use has the most significant affect on the drivers’ striking role. The drivers involved in alcohol use are substantially over-presented in rear-end crashes and the conditional probability is 0.96 in Node 3, which could be two times higher than that for the normal drivers \( (P = 0.454) \) in Node 2. Considering the age factor, the classification tree basically segments the data into four age groups: <21 years, 21–31 years, 32–75 years, and >75 years. The drivers younger than 21 years in Node 6 have the largest propensity of rear-end crashes \( (P = 0.645) \), presumably because of risk-taking and attitudinal factors. Although the crash frequency is smaller \( (N = 303) \) in Node 9), the older drivers (>75 years) presented the relatively
higher risk involving rear-end crashes. Their conditional crash probability (0.617) is far higher than the middle-age groups, presumably because of age-related deterioration of their physical and cognitive abilities. The crash risk of the 21–31 years drivers (P = 0.491 in Node 7) is lower than the older group, but higher than the middle age group (P = 0.387 in Node 8). The result is consistent with the previous analysis (Santokh, 2003): the propensity of drivers involved in rear-end crashes showed a decreasing trend with increasing age until the age of 56–65, after which the drivers increase to a higher crash involvement propensity for the age group older than 75. It should be noted that it is straightforward to conclude that the two nodes splitting off from a parent node are statistically different, but it is risky to compare nodes from different sections of the tree. Therefore, the comparisons between age groups in different subtrees still need to provide further statistical evidence to testify the conclusions. Based on Chi-square test for the crash risk comparison between the 21–31-year-old group (Node 7) and the older group, (Node 9) the p-value is less than 0.001 (χ² = 72.588). In the 21–31 age group, the results also show significant gender difference in rear-end crashes. According to the model, the crash probability of male drivers is 0.535 in Node 10, slightly higher than that of female drivers in Node 11 (P = 0.432). Furthermore, those male drivers are more likely involved in rear-end crashes under cloudy or rainy weather conditions compared to clear weather. In the middle-age group (32–75 years), the physical defect of driver may significantly contribute to rear-end crashes. The crash database recorded possible physical defects of drivers involved in crashes, which are coded as 1—no defect, 2—eyesight defective, 3—fatigue/sleep, 4—hearing defect, 5—illness, 6—seizure/epilepsy/blackout, and 7—other physical defect. Since the data sample sizes in some levels are very small, so the levels 2–7 are combined together in this study. It is not surprising that drivers with some kind of physical defects have a very high rear-end crash risk (P = 0.674), as shown in Node 12.

In addition vehicle type is also found to significantly affect accident propensity. There are a total of 13 types of vehicles classified by accident vehicles in the database. Four types of vehicles are focused on in the study including 1—passenger car, 2—passenger van, 3—pickup/light truck, and 4—large size vehicle. Large size vehicle is combined with medium truck, heavy truck, truck-tractor, motor home, and bus because the sample size of these types of vehicles is relatively small. Results show that conditional rear-end probabilities for passenger cars, passenger vans, and pickup/light trucks are very similar and relatively lower (P = 0.374 in Node 16), while large size vehicles are over-involved in rear-end crashes (P = 0.556 in Node 17). This result is consistent with the conclusion drawn by the Federal Motor Carrier Safety Administration (2005) that trucks strike other vehicles in the rear much more often than they are struck by other vehicles. Large-size vehicles are heavier than automobiles in the traffic stream. Relatively, they are less maneuverable and take longer to stop. Bus or truck drivers sit up much higher than passenger vehicle drivers and can see much further down the road, but they may have difficulty responding to brake lights of the leading car with a small headway.

**DISCUSSIONS AND CONCLUSIONS**

Using the 2001 Florida traffic crash database, this study investigated the overall characteristics of rear-end crashes at signalized intersections based on the classification tree method and Quasi-induced exposure concept. The classification tree is an appropriate technique for statistical modeling and significance testing in induced exposure analysis. The tree models identified a series of important factors significantly associated with the rear-end risk, including speed limit, alcohol use, crash injury severity, road surface condition, lighting condition, weather condition, driver age, vehicle type, driver physical defect, and gender.

With consideration of traffic environmental conditions, the analysis showed that the rear-end crashes are over-presented in the higher speed limits; the rear-end crash propensity for daytime is apparently larger than nighttime; and the reduction of braking capacity due to wet and slippery road surface conditions would definitely contribute to rear-end crashes, especially at intersections with higher speed limits. Compared to the other two-vehicle crashes, rear-end crashes are more likely to cause no injury or possible injury. Generally, angle, head-on, and left-turn collisions are overinvolved in serious driver injuries and vehicle damages. However, drivers involved in alcohol use may contribute to not only rear-end crash risk, but also a higher chance of nonincapacitating evident injury, incapacitating injury, and fatal injury.

Either the logistic regression method or the classification tree method can handle continuous independent variables in statistical models. However, for a classification purpose to segment or classify the continuous variables into homogeneous groups the logistic models have to assign the continuous variables to pre-defined groups in a single step, whereas classification trees have the flexibility of assigning the objects in one or more steps based on the similarity of the observation. Moreover, the classification tree as a nonparametric model doesn’t need to make assumptions about the nature of the data. Logistic regression makes the assumption that a link function can be used to relate the probabilities of group membership to a linear function of the predictor variables. It is also assumed that the observations are independent (Worth & Cronin, 2003). Previous traffic safety studies (Stamatiadis & Deacon, 1995, 1997) in the Quasi-induced exposure analysis grouped driver age into the five- or 10-year intervals to fit the logistic regression models. However, the researcher’s subjective age classification may not accurately reflect the feature of crash risk change associated with driver age. The tree models can objectively construct driver age classification by dividing the data set into smaller and more homogeneous groups that can be made further analyses. For the rear-end crash analysis in this study, it segmented the rear-end drivers into
four homogeneous age groups: <21 years, 21–31 years, 32–75 years, and >75 years. Younger drivers are more likely to be associated with offences and a variety of accident types involving speeding (Clarke et al., 2005). The largest propensity indicates that the drivers younger than 21 years are the most vulnerable group for the rear-end crash risk. Some countermeasures, such as particular education programs for students, are needed to improve young driver accident involvement by focusing on issues of skill-based learning and hazard perception. Another interesting result from the tree model is that a very large population ranged from 31 to 75 show a similarity in the rear-end crash risk, which is different from the general driver age risk trend that, after 65 years old, drivers are overinvolved in traffic crashes than all other drivers including the younger (Stamatiadis & Deacon, 1995, 1997). Therefore, further studies are suggested to compare crash risk distributions for the driver age between main crash types, such as angle, rear-end, head-on, and sideswipe, which would display more specific driver age difference and contribute to seeking corresponding traffic countermeasures.

Another advantage of decision tree models is to identify and easily explain the complex patterns associated with crash risk. However, in a multiple logistic regression model it is hard to analyze interaction effect with more than two independent variables. In this study it was found that the driver age, gender, and weather condition are correlated to each other; in the 21–31 age group, the male drivers are especially overinvolved in rear-end crashes under the adverse weather conditions. The tree model also showed that the 32–75-year-old drivers driving large size vehicles have a larger propensity for rear-end crashes compared to those driving passenger vehicles. Therefore, for urban highway systems safety issues related to compatibility of large size vehicles and other common passenger cars need to be concerned further. On the other hand, because the drivers for large size vehicles are professional drivers, driver training programs in buses and truck transportation companies may effectively improve the crash involvement by large-size vehicles.

In summary, easily-explained tree models are a proper statistical tool for traffic safety analysis to investigate the crash propensity related to traffic conditions of drivers, vehicles, and environments. The tree classification analyses in this article may provide a better understanding of the rear-end risks and more information to seek effective crash countermeasures.

REFERENCES


