Development of a Hierarchical Tree Based Regression Model For Rural Traffic Crashes

Terrance M. RENGARASU
Ph.D., Senior Lecturer
Faculty of Engineering,
University of Ruhuna,
Hapugala, Galle, 80,000, Sri Lanka.
Tel: +94912245762, Fax: +94912245762
E-mail: rengarasu@cee.ruh.ac.lk

Toru HAGIWARA
Ph.D., Professor
Graduate School of Engineering,
Hokkaido University,
N13, W8, Kita-Ku, Sapporo, 060-8628, Japan.
Tel/Fax: +81117066214
E-mail: hagiwara@eng.hokudai.ac.jp

Masayuki HIRASAWA
Ph.D., Senior Researcher
Civil Engineering Research Institute for Cold Region, Japan
Hiragishi 1-3-1-34, Toyohira-Ku, Sapporo, 062-8602, Japan.
Tel: +81118411738 Fax: +81118419747
Email hirasawa@ceri.go.jp

Abstract: Development of a hierarchical tree based regression model (HTBRM) to find the causal factors of traffic crashes at rural sections of the road is presented in this paper. CHAID, rarely used type of HTBRM, was used to model traffic crashes in six national roads of Hokkaido, Japan from 1997 to 2001. CHAID is a non binary HTBRM therefore able to represent the complex relationship between the traffic crashes and the dependent variables. Four CHAID models 1 km non-winter, 1 km winter, 3 km non-winter, and 3 km winter were developed. Models with 3 km road segments were found to be better performing. Tunnels, bridges, and snow sheds were found to be significantly increasing the traffic crashes in winter months of the year. Models were able to represent the complex relationship between traffic crashes and the independent variables. Further several combinations of variables which cause high number of traffic crashes were identified.

Key Words: rural traffic crashes, hierarchical tree based regression model, CHAID

1. INTRODUCTION

Number of uses of Hierarchical Tree Based Regression Models (HTBRM), a non parametric approach, in road safety studies is in the rise. Many attractive features of HTBRMs have been noted. From a theoretical perspective HTBRM is a non-parametric model without any pre-defined underlying relationship between the target (dependent) variable and the predictors (independent variables). HTBRM posses the capability to automatically search for the best predictors and the best threshold values for all predictors to classify the target variable. Another advantage is that HTBRMs can effectively handle multi-co-linearity problems. In addition, in HTBRMs, data outliers are isolated into a node and result in no effect on splitting. From a practical perspective, the first advantage of the HTBRM is the graphical representation of the model. This makes the results more easily understood across disciplines, and also by non-statisticians. The model is structured as a sequence of “if-then” questions which can be further incorporated into an expert system. By answering these questions and
tracing a path down the tree to a terminal node, traffic engineers can easily predict the injury likelihood of a traffic crash. In addition, HTBRM effectively deals with large data sets containing a large number of independent variables and can produce useful results by using only a few selected important independent variables. For situations where there are many possible independent variables including all possible interactions, variable selection for the regression analysis would be an issue. Although automated selection methods such as “best” subsets algorithms and stepwise selection techniques are available, those lead to the identification of one particular subset of variables.

Previous studies have used different types of HTBRMs and for variety of purposes. Use of HTBRM to stratify the study data can be seen in a study by Park and Saccomanno (2006). Recursive partitioning method (RPART) was used to stratify the study data set into homogeneous sub-classes. Thereafter, using the “homogeneous sub-classes”, a Poisson regression model was developed to model traffic crashes at highway-railway grade crossings. Use of HTBRM to identify candidate variables for another model can be seen in a study by Hakkert et al. (1996). Hakkert et al. explored the direct and interactive effect of geometric features and traffic characteristics on non-intersection traffic crashes on interurban road links in Israel. Classification and regression tree (CART) was used in this study as a preliminary tool to identify meaningful candidate variables and interactive terms for the generalized linear modeling (GLIM) analysis.

Use of HTBRM, as the main analysis tool, can be seen in many studies. A study by Scheetz et al. (2009) used HTBR to identify severe and moderate vehicular injuries in young and middle-aged adults. This study used 74,626 records of adults (age 18-64 years). Two binary decision tree models were developed using the CART algorithm; one for the patients with severe injuries and the other for the patients with moderate injuries. A study by Harb et al. (2009) used HTBR to explore the pre-crash maneuvers of drivers involved in rear-end, head-on, and at angle traffic crashes. Using the same set of independent variables and changing the crash type (i.e. dependent variable) three CART models were developed. In addition, this study used a method called “decision forests” to find importance factors of the independent variables. A study by Hallmark (2002) used HTBR to identify on-road geometric and operational characteristics that influenced the fractions of vehicle activity spent in specific modes. In this study several CART models were developed. Results indicated that queue position, grade, downstream and upstream per-lane hourly volume, distance to the nearest downstream signalized intersection, percent heavy vehicles, and posted link speed limit were the most statistically significant variables. A journal paper by Washington (2000) proposed a methodology to combine OLS with HTBR. The journal paper presented an example for trip generation. A study by Council and Stewart (1996) developed severity indexes for roadside fixed objects using HTBR. Illinois (1985-1991) and North Carolina (1980-1992) data was used in the study. CART algorithm was used in the study to create several decision tree models.

Of importance to this study, a 2010 study by Yan et al. (2010) used HTBR to explore the train–vehicle crashes at passive highway-rail grade crossings. This study used data covering 27 years from 1980 to 2006 and across USA. In this study using an algorithm called CHAID (Chi-squared Automatic Interaction Detection) four non-binary decision tree models were developed. Compared with use of CART, CHAID has very few uses in traffic safety literature. However, use of CHAID has several advantages over CART. First, the CHAID model can generate non-binary decision trees, meaning that some splits have more than two branches. CHAID algorithm therefore tends to create a wider tree than binary decision tree growing algorithms. Second, CHAID algorithm can handle several types of variables;
especially it can handle non-categorical dependent variables. The CHAID algorithm was selected as the main analysis or modeling tool of this study.

As for the type of traffic crashes, non intersectional and non pedestrian traffic crashes occurred in rural national roads were selected. Several reasons affected the selection of rural roads. First, less frequent public transportation services, and farming related activities force the rural residents of Hokkaido, the study area, to use private transportation. Essentially, transportation is a life line for the rural areas. Thus traffic crashes in rural area have high effect on life of people than the urban traffic crashes. Second, in 2007, rural roads accounted for half of the traffic crashes occurred in Japan (International Association of Traffic and Safety Sciences 2006; International Association of Traffic and Safety Sciences 2008). Further, rural traffic crashes have high fatal crash ratios than the urban traffic crashes.

Aims of this study are threefold; develop CHAID models to predict traffic crashes occurred in rural roads using same data with different aggregation road segment lengths, compare the developed CHAID models for the prediction accuracy and select suitable road segment length, and to identify the combination effects of road and seasonal variables on traffic crashes occurred on rural roads.

2. DATA

2.1 Traffic Crash Data
Crash data obtained from Traffic Accident Analysis System, which was created by the CERI, was used in this study (Hirasawa and Asano, 2003). Using the inbuilt filters in TAAS, crash data for the 5 years from 1997 to 2001 was extracted. Unavailability of road geometry database except for the year 1999, and the availability of AADT data for 1999 influenced the selection of traffic crash data from year 1997 to 2001. As for the crash type, only the non intersectional traffic crashes occurred at rural national roads were considered. Due to locations based variables (such as intersection width, traffic control type etc.) involved, which fall out side of the aims of this study, intersectional traffic crashes are not considered in this study. Traffic crashes were grouped into winter and non-winter traffic crashes according to the month of occurrence. Traffic crash data was limited to six national roads; 5, 38, 39, 40, 44, and 274. For the simpicity and ease of result interpretation only two types of road segments, 1 km and 3 km road segments, are used in this study. Traffic crash data for 1 km road segments and 3 km road segments was collected separately. In total 1,219 one kilometer long road segments and 390 three kilometer long road segments were used. In the case of 1 km road segments, there were 1,273 non-winter traffic crashes (at 1.044 traffic crashes/1 km/5 years), and 1,165 winter traffic crashes (at 0.956 traffic crashes/1 km/5 years). Basic statistics of traffic crashes used in this study with 1 km road segments are shown in Table 1. In the case of 3 km road segments, there were 1,177 non-winter traffic crashes (at 3.080 traffic crashes/3 km/5 years), and 1,060 winter traffic crashes (at 2.718 traffic crashes/3 km/5 years). Basic statistics of traffic crashes used in this study with 3 km road segments are shown in Table 2. Traffic crash distributions with 1 km and 3 km road segments for non-winter and winter months of the year are shown in Figure 1.

2.2 Independent Variables
Data for 11 independent variables was collected from the MICHI database (Ministry of Land, Infrastructure and Transport Japan, 1999). MICHI is a road database compiled by the
The independent variables were of four classes; road geometry, road-cross section, road structure, and traffic flow. For road geometry class, data for hilliness, maximum grade, horizontal radius and bendiness was collected. The variable ‘Hilliness’ was defined as the absolute sum of the vertical ascents and descents of a vehicle in a road segment. The variable ‘Maximum Grade’ was defined as the maximum value of vertical grade (%) within a given road segment. Equations (1) and (2) show the calculations of the variables ‘Hilliness’ and ‘Maximum Grade’ respectively.

\[ \text{Hilliness} = \sum_{i=1}^{n} h_i \]  
\[ \text{Maximum grade} = \max_{i} S_i \]

In Equations (1) and (2), \( h_i \) is the absolute value of ascent or descent of the road in meters, \( S_i \) is the absolute value of the vertical grades and \( n \) is the number of different vertical grades within a given road segment. The variable ‘Horizontal radius’ was defined as the average horizontal radius. Calculation of the variable horizontal radius is shown in Equation (3). The variable ‘Bendiness’ was defined as the sum of all deflection angles within that road segment. The calculation of the variable bendiness is shown in Equation (4).

\[ \text{Horizontal radius} = \frac{\sum_{i=1}^{m} \frac{L_i}{R_i}}{\sum_{i=1}^{m} \left( \frac{L_i}{R_i} \right)} \]  
\[ \text{Bendiness} = \sum_{i=1}^{m} \left( \frac{L_i}{R_i} \times \frac{180}{\pi} \right) = \sum_{i=1}^{m} \left( \theta_i \times \frac{180}{\pi} \right) \]

### Table 1 Basic statistics of traffic crashes used in this study with 1 km road segments

<table>
<thead>
<tr>
<th>Road</th>
<th>Number of road segments</th>
<th>Non- winter traffic crashes</th>
<th>Winter traffic crashes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Mean</td>
<td>S. dev</td>
</tr>
<tr>
<td>5</td>
<td>233</td>
<td>470</td>
<td>2.017</td>
</tr>
<tr>
<td>38</td>
<td>250</td>
<td>228</td>
<td>0.912</td>
</tr>
<tr>
<td>39</td>
<td>188</td>
<td>167</td>
<td>0.888</td>
</tr>
<tr>
<td>40</td>
<td>229</td>
<td>91</td>
<td>0.397</td>
</tr>
<tr>
<td>44</td>
<td>115</td>
<td>89</td>
<td>0.774</td>
</tr>
<tr>
<td>274</td>
<td>204</td>
<td>228</td>
<td>1.118</td>
</tr>
<tr>
<td>Total</td>
<td>1,219</td>
<td>1,273</td>
<td>1.044</td>
</tr>
</tbody>
</table>
In Equation (3) and (4), $R_i$ is the horizontal curve radius, $L_i$ is the horizontal curve length, $\theta_i$ is the horizontal curve deflection angle, and $m$ is the number of horizontal curves within the given road segment.

For the road structure class, location data was collected for tunnels, bridges and snow sheds. Three separate dummy variables were used to identify the road segments with these structures. The dummy variable ‘Tunnel’ took a value of 1 when there is at least one tunnel in a road segment and a value of 0 otherwise. Similarly, the dummy variable ‘Bridge’ took a value of 1 when there is at least one bridge in the road segment and 0 otherwise. A dummy variable was used to indicate whether there is at least one snow shed (a structure with roof to prevent snow slides covering the road) in the road segment. The dummy variable ‘Snow shed’ took a value of 1 when there is at least one snow shed in the road segment, and value of 0 otherwise.

For cross-section class, data was collected for maximum shoulder width, average lane width, and truck lane. The variable ‘Maximum Shoulder Width’ was defined as the maximum width of the shoulder in a segment. To indicate the segments with a truck lane (climbing lane for truck on grades), a dummy variable was used. The dummy variable ‘Truck Lane’ took a value of 1 in the segments with a truck lane, and 0 otherwise. The variable ‘Average Lane Width’ was defined as the weighted average of lane width over the length. Calculation of the variable average lane width is shown in Equation (5).

$$\text{Average lane width} = \frac{\sum_{i=1}^{p} (LW_i \times l_i)}{\sum_{i=1}^{p} l_i} \quad (5)$$

In Equation (5) $LW_i$ is the lane width in meters, $l_i$ is the length of each section in meters and $p$ is the number of sections corresponding to the different lane widths in a given segment.
For traffic flow class, AADT data were collected. Since there is a reduction of traffic flow in the harsh winter, the need arose to collect AADT data separately for winter months and non-winter months. The AADT data were obtained from census measurements carried out by the Ministry of Construction, Japan. For the purpose of AADT measurement each road is divided into smaller ‘census divisions’. The Census divisions are decided so that the flow rate within a division remains fairly equal. For the selected six national roads, AADT was measured at 169 census divisions with an average length of 7.675 km. However, the measurements were limited to the non-winter months (April to October). The winter (January to March and November to December) AADT was estimated using the winter to summer AADT ratio obtained from the AADT count of the year 1999 measured by the Hokkaido Development Bureau.

3. METHODOLOGY

There were 4 databases; 1 km non-winter, 1 km winter, 3 km non-winter, and 3 km winter. For each of the 4 databases, a HTBR model was developed. Each of the 4 databases were divided into two; learning data and testing data. Records were randomly distributed among

Figure 1 Traffic crash distributions with 1 km and 3 km road segments for non-winter and winter months of the year
the learning data and testing data, 67% of the whole database was set as the learning data and rest was set as the testing data. After dividing the data into learning data and testing data, F-test was performed. Results of the F-test indicated that, the expected values of the number of traffic crashes in the learning data and testing data were not to be significantly different.

In creating the HTBR models, CHAID algorithm was used. The CHAID algorithm uses chi-squared statistics to identify optimal split points. CHAID algorithm runs in loops. In a loop, first, the most significant independent variable is selected using a chi-squared independence test. Second, using chi-squared statistics, the optimal points for dividing the selected variable are calculated. Third, the database including the entire input variables is divided into sub-databases along the optimal points. This ends one loop in CHAID algorithm. In the next loop CHAID algorithm moves onto the resulted sub databases and process is repeated for each of this sub databases. CHAID algorithm stops when one of the stopping criterions is met.

In CHAID algorithm only the learning data was used to create the two models and the test data was used to test the accuracy of the model. In the CHAID algorithm smallest possible node membership was limited to 5 road segments, and the number of levels below the starting node was set to four. In addition, a p-value of 0.05 was used as criteria for splitting data. As for the target variable, number of traffic crashes in 5 years in a road segment was used.

Evaluation of the model is done in two levels, one in the model level and the other in the variable level. The model level evaluation is used to compare the models and possibly select the best model. The variable level evaluation is used to compare the variables within a model. For the model level evaluation, mean residual error was calculated. Mean residual error is always greater than or equal to zero and the model with higher prediction accuracy will have a smaller mean residual error. Calculation of the mean residual error is shown in Equation (6).

\[
Mean \ residual \ error = \frac{\sum_{i=1}^{n}|P_i - O_i|}{n}
\]  

(6)

In Equation (6), \(P_i\) is the predicted number of traffic crashes in the \(i^{th}\) road segment, \(O_i\) is the observed number of traffic crashes in the \(i^{th}\) road segment, and \(n\) is the number of testing data. For the variable level evaluation importance factor value was calculated. Importance factor value was defined as the percentage of explained error by each variable. The calculation of the importance factors is shown in Equation (7).

\[
Importance \ factor = \frac{Reduction \ in \ residual \ error \ by \ the \ variable}{Reduction \ in \ residual \ error \ by \ the \ whole \ tree} \times 100\%
\]  

(7)

4. RESULTS

4.1 CHAID Model Results (1 km – non-winter)
CHAID decision tree model developed for non intersectional traffic crashes occurred at rural national roads in non-winter months using 1 km road segments is shown in Figure 2. Total of four variables; AADT, average slope, dummy variable curve, and dummy variable tunnel were found to be significant variables. CHAID model has 3 levels below the root node.
Proceedings of the Eastern Asia Society for Transportation Studies, Vol.8, 2011

CHAID model has 15 nodes out of which 10 are leaf nodes. Importance factors values of the variables are shown in Table 3.

AADT (Level 0 to 1 Split 1 & Level 1 to 2 Split 1): The variable AADT has an importance factor value of 83%. The number of traffic crashes increase with the increase of AADT.

Tunnel (Level 1 to 2, Splits 2 & 3): Tunnel has an importance factor value of 9%. Variable tunnel appears twice in the CHAID model. In both cases tunnel increases the number of traffic crashes.

Average slope (Level 2 to 3, Split 1): When the AADT is less than 3500; number of traffic crashes is influenced by the average slope. The variable average slope has an importance factor value of 1%. Relationship between number of traffic and average slope is not linear. When the slope is less than 1% then the predicted number of traffic crashes is 0.111, when the slope is between 1% and 1.8% then the predicted number of traffic increases to 0.625, and when the slope is greater than 1.8% then the predicted number of traffic crashes decreases to 0.063.

Curve (Level 2 to 3, Split 2): If the AADT is greater than 14,800 and the road segment does not contain a tunnel, the number of traffic crashes is influenced by variable curve. The variable curve has an importance factor value of 7%. Curved road segments have a higher number of traffic crashes than the straight road segments.

4.2 CHAID Model Results (1 km – winter)
CHAID decision tree model developed for non intersectional traffic crashes occurred at rural national roads in non-winter months using 1 km road segments is shown in Figure 3. Total of seven variables; AADT, dummy variable for bridge, maximum shoulder width, dummy variable for tunnel, average horizontal radius, average slope, and bendiness were found to be significant variables. CHAID model has 4 levels below the root node. CHAID model has 23 nodes out of which 15 are leaf nodes. Importance factors for both non-winter and winter CHAID models are shown in Table 3.
AADT (Level 0 to 1 Split 1 & Level 2 to 3 Split 2): The variable AADT scores an importance factor value of 85%. The variable AADT appears twice in the model; in the first level and in the third level. In both instances, number of traffic crashes increases with the increase in the AADT.

Bridge (Level 1 to 2 Split 1): The variable bridge scores an importance factor value of 0%. When the AADT is between 5900 and 8700, presence of a bridge would increase the number of traffic crashes.

Max shoulder width (Level 1 to 2 Split 2 & Level 2 to 3 Split 1): The variable max shoulder width appears twice; in the second level and in the third level. The variable scores an importance factor value of 0%. In the second level, when the AADT is between 8,700 and 11,500, then the number of traffic crashes is increased from 1.108 to 3.5 traffic crashes/5 years. In the third level, when the AADT is between 5,900 and 8,700 and without a bridge, then the number if traffic crashes has non linear relationship with maximum shoulder width. Traffic crashes peak when max shoulder width is between 1m and 1.7 m.

Tunnel (Level 1 to 2 Split 3): The variable tunnel appears once in the model in level two. Tunnel has an importance factor value of 4%. If the AADT is over 11,500, then the presence of a tunnel increases the number of traffic crashes from 2.368 to 6.8 traffic crashes/5 years.

Average Horizontal Radius (Level 2 to 3 Split 3): The variable average horizontal radius appears once in the model in level 3. Average horizontal radius has an importance factor value of 5%. If the AADT is more than 11,500, without a tunnel, and the radius increases above 300 m; then, the number of traffic crashes decrease from 4.615 to 2.106 traffic crashes/5 years.
Bendiness (Level 3 to 4 Split 2): The variable bendiness appears once in the model in level 4. Bendiness has an importance factor value of 1%. If the AADT is between 5,900 and 7,600, and with a bridge, then the number of crashes increases with the bendiness.

Average Slope (Level 3 to 4 Split 1): The variable average slope appears once in the model in level 4. Average slope has an importance factor value of 5%. When the AADT is between 5,900 and 8,700, and without a bridge, and a maximum shoulder width 1-1.7m, then the number of traffic crashes increases with the bendiness.

### 4.3 CHAID Model Results (3 km – non-winter)

CHAID decision tree model developed for non intersectional traffic crashes occurred at rural national roads in non-winter months using 3 km road segments is shown in Figure 4. Total of 5 variables were found to be significant in the model, those are AADT, horizontal curve radius, curve, average slope, and average shoulder width. CHAID model has 3 levels below the root node. CHAID model has 12 nodes out of which 8 are leaf nodes. Importance factors for both non-winter and winter CHAID models are shown in Table 4.

<table>
<thead>
<tr>
<th>No</th>
<th>Independent variable name</th>
<th>Importance factor value %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AADT</td>
<td>83</td>
</tr>
<tr>
<td>2</td>
<td>Average horizontal radius</td>
<td>Insignificant</td>
</tr>
<tr>
<td>3</td>
<td>Average slope</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Bendiness</td>
<td>Insignificant</td>
</tr>
<tr>
<td>5</td>
<td>Bridge</td>
<td>Insignificant</td>
</tr>
<tr>
<td>6</td>
<td>Curve</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>Maximum shoulder width</td>
<td>Insignificant</td>
</tr>
<tr>
<td>8</td>
<td>Tunnel</td>
<td>9</td>
</tr>
</tbody>
</table>

Bendiness (Level 3 to 4 Split 2): The variable bendiness appears once in the model in level 4. Bendiness has an importance factor value of 1%. If the AADT is between 5,900 and 7,600, and with a bridge, then the number of crashes increases with the bendiness.

Average Slope (Level 3 to 4 Split 1): The variable average slope appears once in the model in level 4. Average slope has an importance factor value of 5%. When the AADT is between 5,900 and 8,700, and without a bridge, and a maximum shoulder width 1-1.7m, then the number of traffic crashes increases with the bendiness.

AADT (Level 0 to 1 Split 1): The variable AATD appears once in the model in the first level. Root node is divided into three nodes with respect to AADT. AADT has an importance factor value of 93%. Predicted number of traffic crashes increases with AADT.

Average horizontal radius (Level 1 to 2 Split 1): The variable average horizontal radius appears once in the model in level 2. Average horizontal radius has an importance factor value of 0%. When the AADT is less than 6,100 and if the variable average horizontal radius is increased to more than 370, then the number of traffic crashes decreases from 1.1462 to 0.623 traffic crashes/5 years.

Curve (Level 1 to 2 Split 2): The variable curve appears once in the model in level 2. Curve has an importance factor value of 2%. If the AADT is between 6,100 and 15,300, then straight
road segments have 1.455 traffic crashes /5 years and curved segments have 3.115 traffic crashes/5 years.

Average slope (Level 2 to 3 Split 1): The variable average slope appears once in the model in level 3. Average slope has an importance factor value of 3%. If AADT is less than 6,100, curve radius is greater than 370 m, and average slope is more than 0.4% then the number of traffic crashes increased to 0.767 from 0.278 traffic crashes/5 years.

Average shoulder width (Level 2 to 3 Split 2): The variable average shoulder width appears once in the model in level 3. Average shoulder width has an importance factor value of 2%. When the AADT is between 6,100 and 15,300, and segment is curved then the number of traffic crashes is influenced by average shoulder width. Relationship between average shoulder width and the number of traffic crashes is of non-linear nature. This relationship peaks when average shoulder width is between 1.2 m and 1.4 m.

4.3.1 CHAID Model Results (3 km – winter)
CHAID decision tree developed for non intersectional traffic crashes occurred at rural two-lane national roads using 3 km road segments in non-winter months has 15 nodes out of which 10 end nodes. Six variables were found to be significant; those are AADT, Bridge, tunnel, average slope, snow shed, and total curved length. Importance factors for both non-winter and winter HTBR model are shown in Table 4.

AADT (Level 0 to 1 Split 1): The variable AADT appears once in the model in level 1. AADT has an importance factor value of 58%. The root node is divided into three according to AADT. Predicted number of traffic crashes increases with AADT.

Bridge (Level 1 to 2 Split 1): The variable bridge appears once in the model in level 2. Bridge has an importance factor value of 2%. If AADT is less than 9,200 then the presence of a bridge will increase the number of traffic crashes from 0.885 to value of 1.448 traffic crashes/3 km/5 years.
Proceedings of the Eastern Asia Society for Transportation Studies, Vol.8, 2011

Tunnel (Level 1 to 2 Split 2): The variable tunnel appears once in the model in level 2. Tunnel has an importance factor value of 6%. If AADT is between 9,200 and 15,300 then the presence of a tunnel will increase the number of crashes from 3.203 to 7.600 traffic crashes/3 km/5 years.

Average Slope (Level 1 to 2 Split 3): The variable average slope appears once in the model in level 2. Average slope has an importance factor value of 28%. If AADT is greater than 15,300 and the average slope is increased over 2.4% then the number of traffic crashes would increase from 5.905 to 20.600 traffic crashes/3 km/5 years.

Snow shed (Level 2 to 3 Split 1): The variable snow shed appears once in the model in level

Table 4 Importance factors of the variables in CHAID models with 3 km road segments

<table>
<thead>
<tr>
<th>No</th>
<th>Independent variable name</th>
<th>Importance factor value %</th>
<th>Non-winter</th>
<th>Winter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AADT</td>
<td>93</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Average horizontal radius</td>
<td>0</td>
<td>Insignificant</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Average shoulder width</td>
<td>3</td>
<td>Insignificant</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Average slope</td>
<td>2</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Bridge</td>
<td>Insignificant</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Curve</td>
<td>2</td>
<td>Insignificant</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Snow Shed</td>
<td>Insignificant</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Total curved length</td>
<td>Insignificant</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Tunnel</td>
<td>Insignificant</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
3. Snow shed has an importance factor value of 3%. If AADT is less than 9,200 and has a bridge then the presence of a snow shed will increase the number of traffic crashes from 1.343 to 3.167 traffic crashes/3 km/5 years.

Total curved length (Level 2 to 3 Split 2): The variable total curved length appears once in the model in level 3. Total curved length has an importance factor value of 3%. When AADT is between 9,200 and 15,300 and without a tunnel then the number of crashes is influenced by total curve length. The relationship between total curved length and the traffic crashes is not linear.

5. CONCLUSIONS

Comparison of the observed and predicted traffic crashes in the four CHAID models are shown in Figure 6. When comparing the correlation between observed and predicted values in the four models, it is clear that the 3 km CHAID models are better performing than the 1 km CHAID models. Further the mean residual error of the models was also used to compare the accuracy of the models. Mean residual errors in the four CHAID models are shown in Table 5. When compared, models with 3 km road segments have smaller error in predicting traffic crashes per kilo meter than the 1 km CHAID models. In addition to the accuracy, the developed CHAID model should preferably be less complex. From examining the models it is clear that the 3 km CHAID models have less number of nodes than the 1 km CHAID models. Both winter and non-winter 3 km CHAID models are less complex models with smaller mean residual errors. Therefore, winter 3 km CHAID and non-winter 3 km CHAID models were selected to draw conclusions.

- Results from 3 km CHAID models showed that the nature of the road segment (curved or straight) has significant effect on the number of traffic crashes in non-winter months. Curved road segments were more traffic crash prone than the straight road segments. In winter months, results showed that total curved length of a road segment is significantly changing the number of traffic crashes. Tree suggests a complicated relationship between the number of traffic crashes and the total curved length. However in general, it can be concluded that the higher total curve lengths are safer than the smaller total curved length road segments. Average horizontal radius has effect on traffic crashes in non-winter months. However it has very small importance factor. This indicates that average horizontal radius’ effect on the number of traffic crashes is very limited.

- Average slope had a significant effect on traffic crashes in both winter and non-winter models. However its importance factor value in winter is higher indicating that the effect of average slope is greater in winter. In both winter and non-winter models with the increase of slope number of traffic crashes was increased. Average shoulder width has a significant effect on number of traffic crashes in non-winter months. Shoulder does not have a significant effect in winter. This may be due to the fact that in winter months shoulders are covered with snow thus preventing the function of shoulder.

- Bridges were found to be significantly increasing the number of traffic crashes in winter months while the bridges had no significant effect on the number of traffic crashes in non-winter months. In winter months the bridge surfaces tend to get icy often than the other road surfaces. A general road surface is exposed to weather on the top only. However, in the bridges the road surface is exposed to weather in all sides. Therefore, bridges tend to
Similar to bridges, tunnels were found to be significantly increasing the number of traffic crashes in winter months. However, no significant effect was found in the non-winter months. Several reasons might be sighted for the increase in the number of traffic crashes near tunnels in winter months of the year. First, drivers entering or exiting the tunnels have to deal with sudden change in the light conditions. This change in the light conditions is greater in winter months than the non-winter months. In winter months snow covered considerably increases the brightness outside the tunnels, hence the high contrast. Second, sunless and windless conditions in the tunnels may lead to formation of dark ice patches. Third, drivers have to deal with the change in road surface when entering or exiting the tunnels in winter months. In winter months a layer of ice is formed near the ends of tunnel by snow dragged on by the vehicles. This may lead to contrasting driving surfaces near the ends of a tunnel.

Results of this study showed that the presence of a snow shed increases the number of traffic crashes in winter months. However, the presence of a snow sheds did not show any significant effect in non-winter months. In terms of light conditions and road surface,
snow sheds present similar conditions to that of a tunnel. Therefore, one would expect a snow shed to have a similar tendency for traffic crashes as a tunnel.

- In the 3 km CHAID models, AADT was found to be a significant variable in winter and non-winter months. In every instance, the number of traffic crashes increases with the AADT. This is a trivial and a common result.

CHAID model was able to identify and represent the complex relationship between the number of traffic crashes and the dependent variables. In addition, using the CHAID models developed in this study, several combinations of the dependent variables which are responsible for a high number of traffic crashes were identified.

Results of this study have shown that the CHAID model can be adapted to traffic safety studies. However, there is a need to compare the results of CHAID models with other methodologies such as negative binomial model and CART model. CHAID models developed in this study have identified several causal factors affecting the number of traffic crashes. Although, above mentioned variables are affecting the number of traffic crashes, the variables cannot be controlled or manipulated in order to reduce traffic crashes. Therefore, further studies are required to find countermeasures with respect to the identified causal factors to reduce the number of traffic crashes.

**REFERENCES**


