Variable Selection Issues in Tree-Based Regression Models

Xiao Qin and Junhee Han

Recently, there has been increasing interest in the use of classification and regression tree (CART) analysis. A tree-based regression model can be constructed by recursively partitioning the data with such criteria as to yield the maximum reduction in the variability of the response. Unfortunately, the exhaustive search may yield a bias in variable selection, and it tends to choose a categorical variable as a splitter that has many distinct values. In this study, an unbiased tree-based regression generalized unbiased interaction detection and estimation (GUIDE) model is introduced for its robustness against the variable selection bias. Not only are the underlying theoretical differences behind CART and GUIDE in variable selection presented, but also the outcomes of the two different tree-based regression models are compared and analyzed by utilizing intersection inventory and crash data. The results underscore GUIDE's strength in selecting variables equally. A simulation shed additional light on the resulting negative impact when an algorithm was inappropriately applied to the data. This paper concludes by addressing the strengths and weaknesses of—and, more important, the differences between—the two hierarchical tree-based regression models, CART and GUIDE, and advises on the appropriate application. It is anticipated that the GUIDE model will provide a new perspective for users of tree-based models and will offer an advantage over existing methods. Users in transportation should choose the appropriate method and utilize it to their advantage.

Advancements in statistical science and the development of statistical software packages have led to tremendous development in highway safety studies. Significant efforts have been taken to improve model usability and comprehension across the diverse population of safety practitioners. Within the last decade, there has been increasing interest in the use of classification and regression tree analysis.

Stewart illustrated the application of the classification tree to determine the subset of a collection of predictors that yielded the most differentiated likelihood of drivers being either killed or seriously injured when they hit guardrails (*1*). Karlaftis and Golias modeled the relationship between crash rates and rural roadway geometries by using a hierarchical tree-based regression (HTBR) model (*2*). A similar approach was taken by Abdel-Aty et al. to analyze signalized intersection safety performance, which highlighted individual crash prediction models by crash types instead of total number of crashes (*3*). Washington et al. promoted the application of HTBR by extending these models to other transportation-related areas such as trip generation and motor vehicle emissions (*4–6*). Coincidentally, Washington and Stewart (*1, 4–6*) summarized and recommended the potential promise of the regression-tree-based method as a preliminary analysis tool for identifying important variables and suggesting approximate functional forms for parametric models. The recommendation was adopted in a study conducted by Park and Saccomanno, in which highway–railway grade crossing collision models were sequentially developed on the basis of a tree-based data stratification method (*7*).

In general, the outcomes of tree-based models are relatively simple for a nonstatistician to interpret, which is of primary interest to practitioners and engineers. The use of the tree-based model has expanded quickly because of its availability as a function in more and more statistical packages, including the classification and regression tree (CART) or its counterparts S-plus (TREE, RPART), SYSTAT (TREES), and R (TREE, RPART). A thorough understanding of the underlying theories of the tree-based method seems to be lacking in previous studies, however, which may prevent practitioners from producing an optimal model and manipulating data appropriately.

Although not new to readers who are aware of the tree-based method, a tree-based regression model can be constructed by recursively partitioning data with such criteria as the total sum of the squared error (SSE). In other words, the values of all the variables in the model, either discrete or continuous, are selected to yield the maximum reduction in the variability of the response. The algorithm exhaustively searches all the variables as well as all the values for each selected variable to obtain the optimal results. Unfortunately, the exhaustive search can yield a bias in variable selection, and it tends to choose a categorical variable as a splitter that has many distinct values.

For instance, if a categorical variable has n distinct values, there are $2^{n-1} - 1$ possible binary splits and they increase exponentially with n. In this study, an unbiased tree-based regression model called generalized unbiased interaction detection and estimation (GUIDE) is introduced, which is known for the robustness of its variable selection bias (*8*). Compared with the dominant use of the CART model, GUIDE is relative new, but it holds promise for transportation studies. The limited use of GUIDE is partially attributed to the lack of its availability within statistical packages, but it has been successfully applied in developing winter snowstorm cost functions and winter maintenance activities (*9, 10*). It is anticipated that the GUIDE model will provide a new perspective for users of tree-based models and will offer an advantage over existing methods. Users in transportation should choose the appropriate method and utilize it to their advantage.

Starting with a brief description of CART and GUIDE, this study elaborates the theories behind the types of models with a focus on the difference between the two algorithms in variable selection. Following a case study of intersection crash data, the outcomes of the two different tree-based regression models are compared and analyzed. The results of a simulation model are included to illustrate the negative impact resulting from inappropriate application of an

Traffic Operations and Safety (TOPS) Laboratory, 1415 Engineering Drive, University of Wisconsin, Madison, WI 53706. Corresponding author: X. Qin, xqin@engr.wisc.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2061, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 30–38.
DOI: 10.3141/2061-04

algorithm to the data. In conclusion, the strengths and weaknesses of—and, more important, the differences between—the two HTBR models (CART and GUIDE) are addressed and advice is given on the appropriate application.

POISSON REGRESSION TREE THEORY

Crash data are nonnegative, discrete count data that can be modeled by the Poisson distribution when the equality of the mean and variance is not violated or by negative binomial distribution when data overdispersion is present. In this study, the number of crashes, the response variable Y_i, is assumed to follow a Poisson distribution such as

$$Y_i \sim \text{Poisson}(\mu_i) \qquad i = 1, 2, \ldots, n \tag{1}$$

where Y_i is number of crashes at location i and μ_i is the expected number of crashes.

The expected number of crashes μ_i can be expressed as the product of traffic exposure and the exponential function of the potential crash-contributing factors or other explanatory variables:

$$\mu_i = V * \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \tag{2}$$

where

V = traffic exposure [for an intersection, million entering vehicles (MEV); for a segment, million vehicle miles traveled],
$\mathbf{X}_i$ = $(x_{i1}, \ldots, x_{ik})$ = vector of predictor variables for location i, and
$\boldsymbol{\beta}$ = $(\beta_0, \ldots, \beta_k)$ = vector of unknown parameters.

When tree-based regression models or a generalized linear model is fitted, MEV can be handled differently, either as an offset or a predictor. If treated as an offset, as in this study, it is not used to categorize an intersection or split tree; otherwise it will be considered as a splitter and the regression model structure will be changed accordingly.

The linear relationship between the expected number of crashes and the corresponding vector of predictors is obtained from the logarithm transformation:

$$\log(\mu_i) = \log(V) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{3}$$

The maximum likelihood estimation of $\beta_0, \beta_1, \ldots, \beta_k$ can be obtained by maximizing the likelihood function $L(\mu; \mathbf{y}) = \prod_{i=1}^{n} e^{-\mu_i} \mu_i^{y_i} / y_i!$ or, equivalently, its log likelihood function (*11*):

$$l(\mu; \mathbf{y}) = \sum_{i=1}^{n} y_i \log(\mu_i) - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} y_i!$$

Recursive Partition Method

The same notation to express the tree algorithm was adopted from Breiman et al. (Table 1) (*12*). Both CART and GUIDE use the same algorithm to partition data. When a data set is partitioned recursively, it is necessary to select an appropriate splitter. If the predictor variable x_i is numerically ordered, the data set is partitioned by $x_i \le c$, and if x_i is categorical, the data set is partitioned by $x_i \in A$, where c is a constant and A is a fixed subset of possible values of x_i. The method proceeds by an iterative search for the variable as well as its specific value from all of the variables within all the possible levels or values in the model that result in the maximum reduction in variability of the dependent variable. The best splitter, s^*, is determined by deviance D or by squared error, where a squared error for a node t is defined as follows:

TABLE 1 Tree Notation

Notation	Meaning
t	Node
T	Tree
$\tilde{T}$	Set of terminal nodes of T
$\lvert\tilde{T}\rvert$	Number of terminal nodes of T
T_t	Subtree of T with root node t
$\{t\}$	Subtree of T_t containing only the root node t

$$D(t) = \sum_{x \in t} (y_n - \hat{\mu})^2 \tag{4}$$

where $\hat{\mu}$ is an estimation of mean or a sample mean $\bar{y}$. For generalized linear models, the deviance is also called the log likelihood (ratio) statistic, defined by $D = 2 \times (l(\mu_{max}; y) - l(\hat{\mu}; y))$, where μ_{max} is the maximum likelihood estimate. In the Poisson case (*13*), deviance can be simplified as

$$D = 2 \times \left\{ \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}} \right) - \sum_{i=1}^{n} (y_i - \hat{\mu}) \right\}$$

If the deviance for a node t is denoted $D(t)$, the deviance for a tree T is

$$D(T) = \sum_{t \in \tilde{T}} D(t) = \sum_{t \in \tilde{T}} \sum_{x \in t} (y_n - \bar{y}(t))^2 \tag{5}$$

For a binary partitioning by a splitter s, a difference by s is defined as

$$\Delta D(s, t) = D(t) - D(t_L) - D(t_R) \tag{6}$$

where t_L and t_R are the left and right child nodes of t, respectively.

Finally, the best splitter, s^*, is obtained by maximizing the difference:

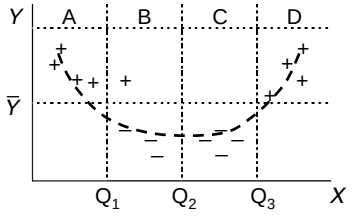
$$\Delta D(s^*, t) = \max_{s \in S} \Delta D(s, t) \tag{7}$$

where S is the set of all possible splitters.

The maximum reduction occurs at some s, a specific value of a selected variable. When the data are split at s into two samples, these remaining samples have much smaller variance in Y than does the original data set. Thus, the reduction at node t is the greatest when the deviances at nodes t_L and t_R are smallest.

GUIDE Variable Selection

The main difference between GUIDE and CART lies in how the splitter variables are selected. CART has adopted one of the most common approaches for making this selection. It searches all possible axis-orthogonal partitions and selects the split that decreases a statistic (an important measure or an impurity measure is used) the most. However, an exhaustive search such as this is biased toward variables with more levels or split points (*14, 15*). The Bonferroni-adjusted test is



(a)

Residuals	A	B	C	D
+	4	1	0	4
-	0	3	4	0

(b)

FIGURE 1 Curvature test of residuals in GUIDE: (*a*) four groups (A, B, C, and D) with the quartiles (Q25%, Q50%, Q75%) and (*b*) contingency table of residuals.

suggested to avoid this bias problem, but since it can be too conservative, it tends to select variables with fewer levels or split points (*16*). The selection methods used in CART and GUIDE are described briefly in this section.

CART uses the following process for selecting a variable. First, let x_k be the kth predictor ($k = 1, \ldots, n$). The best split of x_k at node t is selected as explained in the previous section; that is, $\Delta x_k(t) = \Delta D(s^*, t) = \max_{s \in S} \Delta D(s, t)$. $\Delta D(s, t)$ is called an impurity measure. CART then chooses the predictor with maximum $\Delta x_k(t)$ at node t. The steps are applied to the child nodes recursively.

GUIDE has a different approach to variable selection. As its name indicates, GUIDE is designed to be unbiased through its use of residual analysis (chi-square tests). There are two main tests for selecting a variable in GUIDE; one is a curvature test and another is an interaction test. Figure 1 presents a hypothetical example. After a Poisson model is fit for a numerically ordered variable, the residuals are divided into four groups at the quartiles and cross-tabulated with signs of residuals as rows and groups as columns. As shown in Figure 1*a*, the quartile and sign divide the space into eight grid cells. In Figure 1*b*, the p-value from the χ^2-test can be calculated through the contingency table, in which the number in each cell is the count of the residuals in each grid cell of Figure 1*a*. The same steps can be applied to categorical variables when categories are used as columns, the same application as that using quartiles for continuous variables. It is called a curvature test; the name comes from the shape of the distribution of residuals in the plot.

The correlations among predictors are common and cannot be overlooked. As the name suggests, the interaction test is responsible for testing the intersection between variables. The residuals of each pair of numerical variables are assigned to each of the four quadrants divided by the sample medians of the paired variables as shown in Figure 2*a*. As with the curvature test, the contingency table is generated in which the p-value from the χ^2-test can be calculated on the basis of how many observations with positive (or negative) residuals are located in each cell in Figure 2*b*. The same test applies to categorical variables and the mixture of categorical and numerical variables. If the smallest p-value is from a curvature test, the associated predictor is selected. Otherwise, if the smallest p-value is from an interaction test and they are categorical, the one with the smaller curvature p-value is chosen, or if they are numerical, the one with the smaller total SSE is chosen.

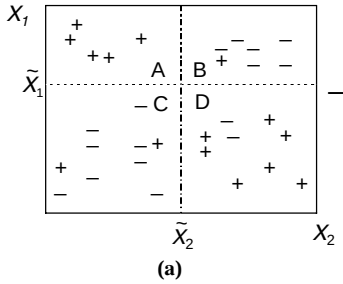
The widely different variable selection approaches provided in CART and GUIDE yield different results, especially for a variable with a large number of distinct variables. As explained earlier, the exhaustive search in CART can yield a bias in variable selection and tends to choose a categorical variable with many distinct values, whereas GUIDE is designed to be more robust and unbiased in variable selection, regardless of how many distinct values the variable may have. The following case study and a simulation shed light on how the two tree-based algorithms performed.

INTERSECTION CRASH CASE STUDY

A common practice of many highway safety research studies is to develop a reliable classification criterion that can be used to categorize similar sites into groups sharing similar attributes and consequences. Traditional statistical methods are cumbersome to use or are of limited utility in addressing these types of classification problems. A tree-based regression method significantly improves the model efficiency.

Explanatory Variable Analysis

Three years of Wisconsin intersection crash data (2001–2003) were used in this study. For each intersection, the crash counts were categorized by crash severity such as fatal, injury, and property



(a)

Residuals	A	B	C	D
+	5	1	2	7
-	0	6	8	2

(b)

FIGURE 2 Interaction test of residuals in GUIDE: (*a*) divide (X_1, X_2) space into 4 quadrants (A, B, C, D) at the sample medians ($\tilde{X}_1$, $\tilde{X}_2$) and (*b*) cross tabulate data with signs of residuals.

damage only (PDO). The intersection file contained a lot of additional information that might be useful in explaining crash causes and in estimating crash frequency by number of daily entering vehicles, area type (rural or urban), and more important, geometric features of the intersection such as number of legs, number of lanes, existence of dividend or left lane, and so forth. These geometric characteristics were categorized as one variable called GEOTYPE in Figure 3 (*17*). The variables used in this study are shown in Table 2 and their descriptive statistics are summarized in Table 3.

The count of fatal crashes is low compared with injury or PDO crashes. For example, there were only 47 fatal counts from all 3,202 intersections, whereas there were 5,026 injury crashes and 7,721 PDO crashes in 2001; 2002 and 2003 data had similar statistics. The dominant crash types were either angle (44.6%) or rear-end (32.4%) types. Small sample size posed an extra challenge to the validity of the conclusions drawn from statistical analysis and inference, as discussed regarding other models such as the Poisson–gamma model (*18*). Extra effort is needed in the

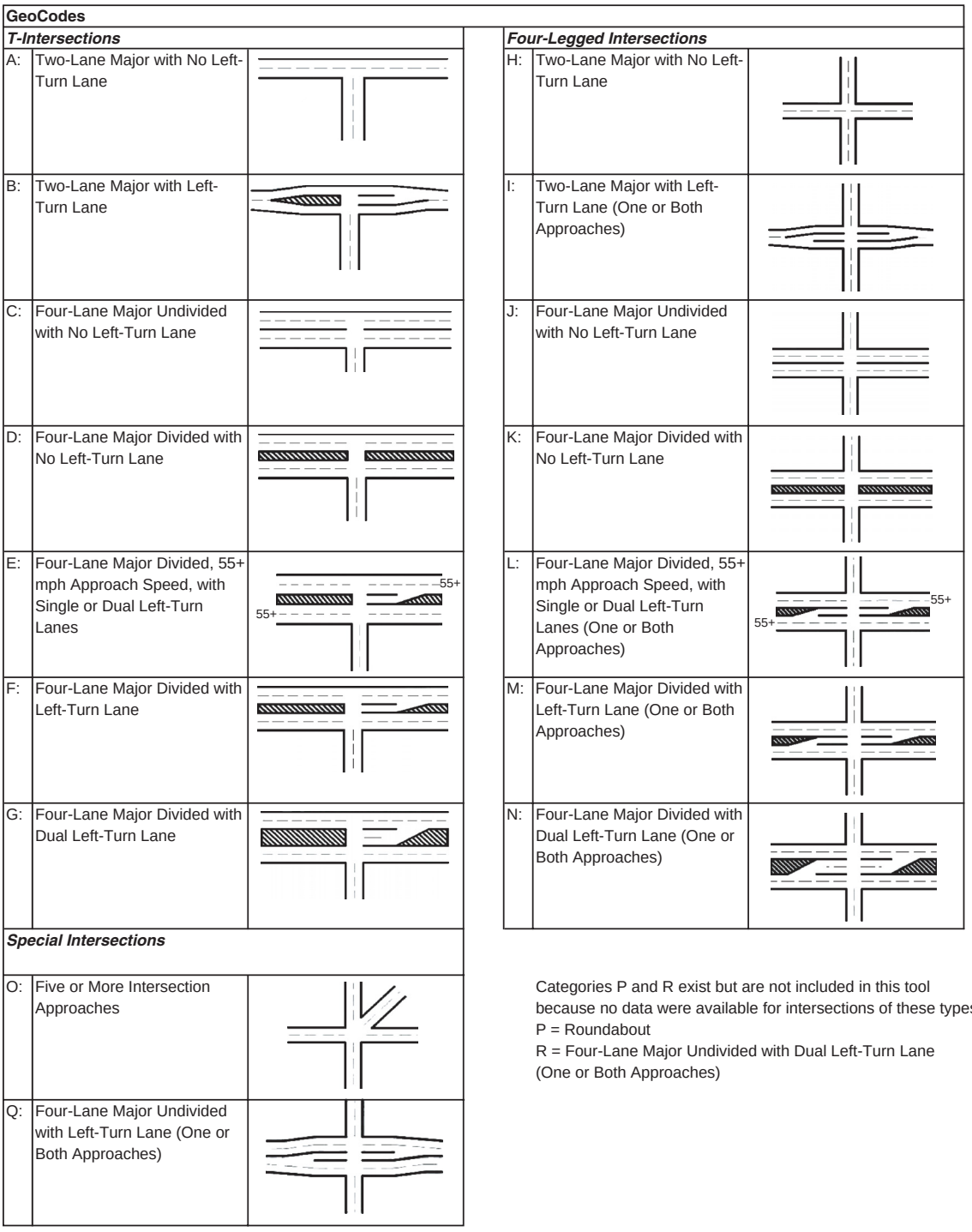


FIGURE 3 Geometric category definitions (GEOTYPE) (*17*).

TABLE 2 Description of Variables

Variable	Description
TOT0103	Total number of crashes between 01–03
GEOTYPE	Geometric type in Figure 3
AREATYPE	Types of area (rural or urban)
ENTVEH	Millions of annual entering vehicles (MEV)
TRFCNTL	Types of traffic controls (3-way, 4-way, yield, flash, side, signal)
NUMLEG	Number of intersection approach legs (3, 4, or unknown)
NUMLANE	Number of major roadway lanes (2, 4, or unknown)
DIVIDED	Existence of major roadway median
LEFTTURN	Existence of left-turn lane(s)

TABLE 3 Summary Statistics of Variables

TOT0103	
Min.	1
Mean	12
Max.	134
SD	14.17

ENTVEH (MEV)	
Min.	0.53
Mean	14.73
Max.	89.87
SD	12.9

NUMLANE	Sites
2	697
4	1,076
Unknown	1,429

AREATYPE	Sites
Urban	1,152
Rural	2,050

DIVIDED	Sites
Yes	788
No	985
Unknown	1,429

LEFTTURN	Sites
Yes	1,086
No	687
Unknown	1,429

NUMLEG	Sites
3	361
4	1,349
Unknown	1,492

TRFCNTL	Sites
3-way	4
4-way	41
Flash	2
Side	2,147
Signal	1,005
Yield	3
Total	3,202

future to investigate the sensitivity of small sample size on tree-based methods.

Since it is difficult to get a meaningful classification result with an extremely low count, the study focused on the total number of crashes for all years in the study. Another outstanding issue in traffic safety studies is the missing information and how it is handled. A general discussion of statistical analysis with missing data has been made by Little and Rubin (*19*), and Feelders discusses missing data in tree-based methods (*20*). GUIDE and CART have different approaches to handling missing data: GUIDE employs mean imputation for a missing numerical variable and creates a new categorical variable for a missing category (*21*), whereas CART uses surrogate splits (*12*, *20*).

The average annual crash frequency was calculated for each individual intersection assigned a geometric category (GEOTYPE). GEOTYPE represents a large variety of intersections whose safety performance is determined by crash frequencies. Figure 4 shows that intersection crashes vary drastically from a high point at GEOTYPE N with average crashes of 50 to Types F, G, J, K, L, M, O, Q with above-average numbers of crashes, to other types with fewer crashes. An alternative approach is to replace GEOTYPE with four primary geometrics that include the number of intersection approach legs, number of major roadway lanes, whether the major roadway had a median or not, and the existence of a left-turn lane or lanes. The latter grouping method is slightly undermined by not incorporating other details such as speed limit and number of left-turn lanes, but it simplifies intersection categories by not counting as many as 16 distinct values in the GEOTYPE variable. The comparison between using GEOTYPE and its substitute is provided in the next section.

Stratification Results

Since the exhaustive search adopted by CART is biased toward variables with more levels or split points (*15*) and GUIDE is more robust in variable selection by design, the variable GEOTYPE, which has 16 distinct values excluding unknowns, may perform differently in CART and GUIDE. It is of primary interest to investigate how the

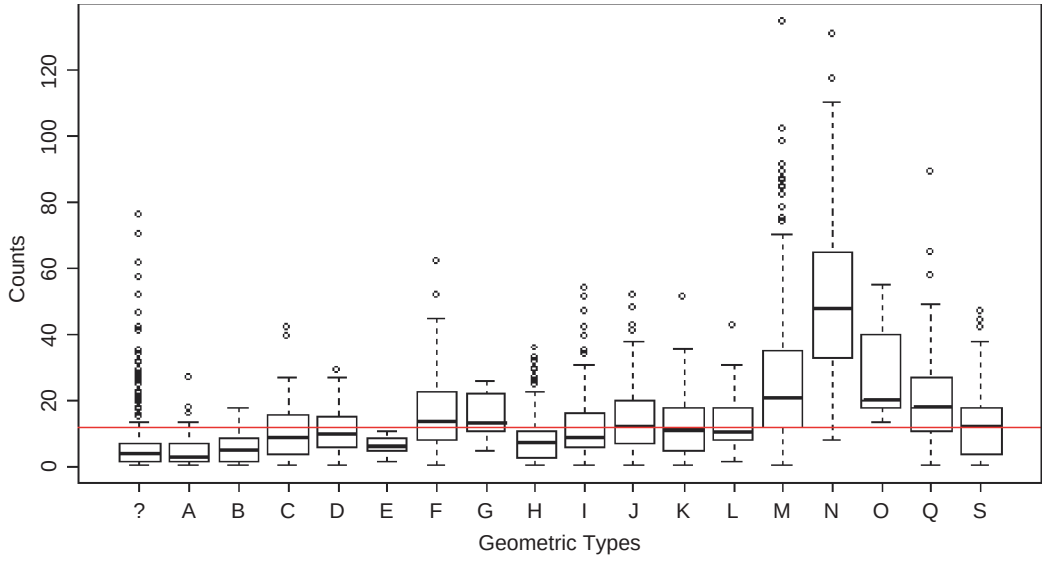


FIGURE 4 Crashes by GEOTYPE (red solid line indicates average of crashes and question mark stands for missing or uncategorized data).

variable GEOTYPE gets selected in terms of both sequence and splitting value in both algorithms.

Figure 5 shows that GEOTYPE is used as the first variable to stratify intersection by crash rate (crash per MEV) in both GUIDE Version 5.2 (*21*) and CART via rpart in the R package (*22*). The splitting values, however, are different. In GUIDE, GEOTYPE is divided by Group {A, B, C, D, F, G, K} and Group {E, H, I, J, L, M, N, O, Q, S}, whereas CART splits the intersections by Group {A, B, F, G} and Group {C, D, E, H, I, J, L, K, M, N, O, Q, S}. The two algorithms largely disagree on Types C and D—a four-lane major undivided intersection with no left-turn lane and a four-lane major divided intersection with no left-turn lane. From an engineering perspective, GUIDE is more reasonable because it groups intersections of Types A, B, C, D, F, and G (which are T-intersections with at least 9 conflict points) and of Types H, I, J, L, M, and N (which are four-leg intersections with at least 32 conflict points). Another finding is that GUIDE continues to stratify intersections by traffic control for both child nodes from the first split, whereas CART only stratifies the node (Node [2]) with a larger number of distinct values, suggesting a variable selection bias. For the node partitioned by TRFCNTL,

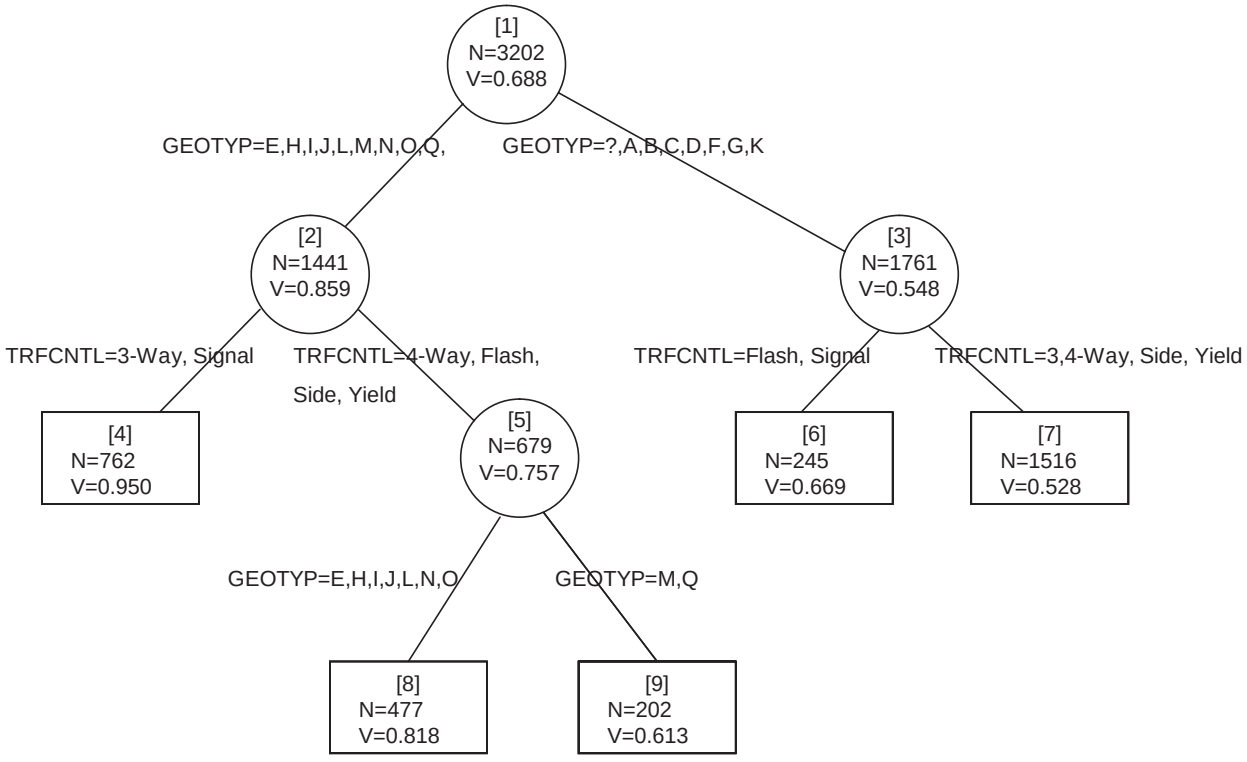


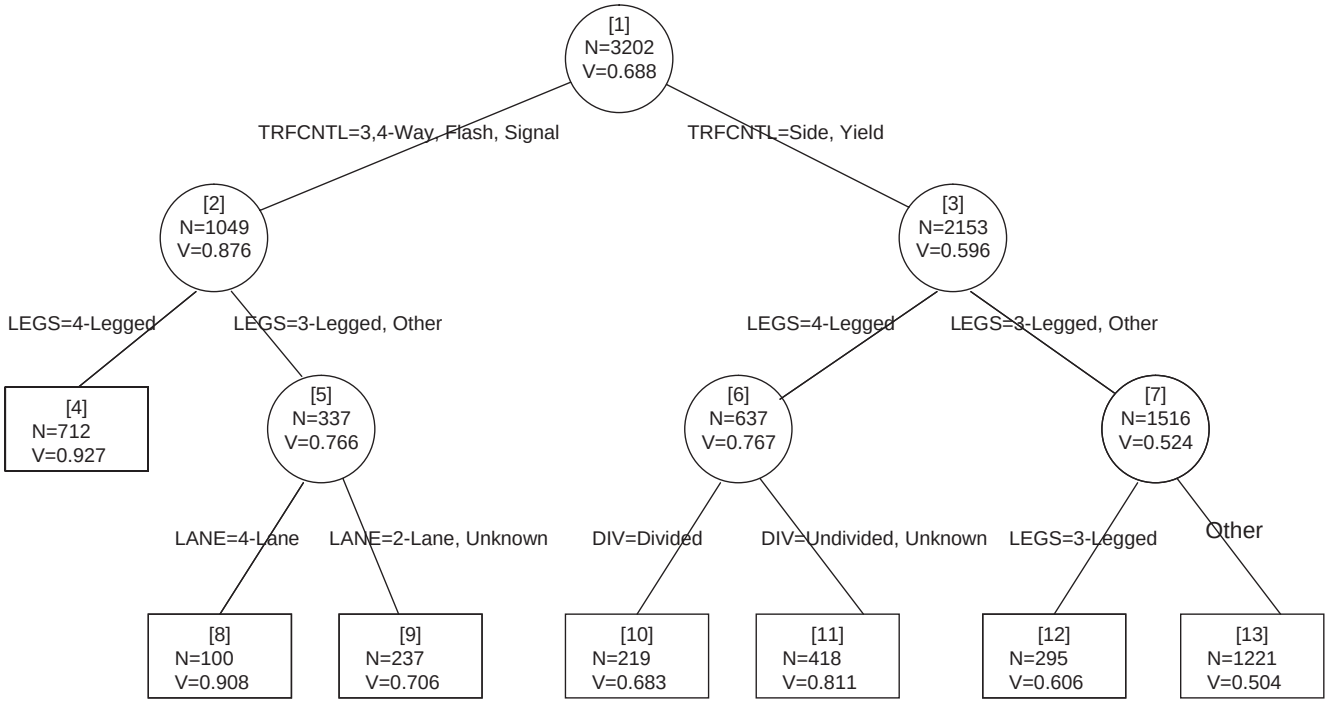
FIGURE 5 (*a*) GUIDE results and (*b*) CART results, including GEOTYPE as a predictor. (*N* = number of sites in group; *V* = average crash rate in group.)

both GUIDE and CART partition the intersection by Group {3-Way, Signal} and Group (4-Way, Flash, Side). GUIDE continues to stratify the node using GEOTYPE, but CART ceases to do so.

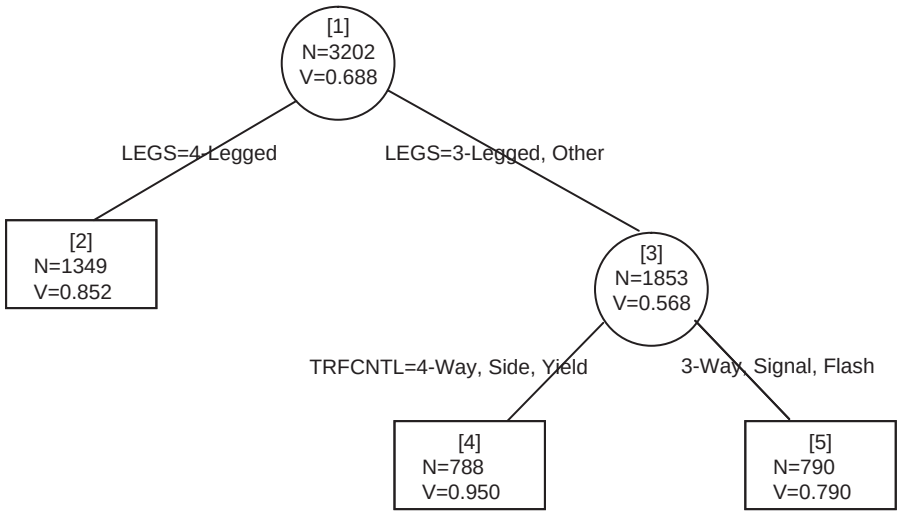
As suggested in the section on the explanatory variable analysis, replacing the variable GEOTYPE (which contains 16 distinct values with four primary geometric features) may mitigate the negative influence of variable selection bias. It would be of interest to investigate whether the two algorithms perform consistently when decomposing a complex variable with simple variables to identify if they are exchangeable. Figure 6 illustrates the following observations:

- GUIDE and CART split the tree quite differently: GUIDE uses TRFCNTL, whereas CART uses NUMLEG as the first splitter;
- As for the second splitter, the two algorithms switched the sequence of variables with the use of NUMLEG in GUIDE and the use of TRFCNTL in CART; and
- GUIDE continues to split the tree using DIVIDED, but CART stops at the traffic control.

The sequence of variables introduced into each model and their relative importance can be produced via the tree-based method. Variables found to be significant were kept in the tree, and insignificant variables were rejected from the tree. This procedure is considered as the means to rate the importance of input factors (*3*). In GUIDE, the variable TRFCNTL is more important in reducing data impurity than is NUMLEG, but the sequence is reversed in CART. Moreover,



(a)



(b)

FIGURE 6 Results for (*a*) GUIDE and (*b*) CART, including GEOTYPE as a predictor. (*N* = number of sites in group; *V* = average crash rate in group.)

the tree-based stratification method is regarded as a valid way to account for control factors in collision prediction by selecting the ones that reduce the impurities in the data (*7*). The variable DIVIDED would be included in the model according to GUIDE but would be rejected if the model were based on CART. From an engineering perspective, even though TRFCNTL and NUMLEG may be equally important to categorize intersections, traffic control type (signalized and unsignalized intersections) is probably more popularly used than the number of legs because the former directly reduces the number of conflicts produced by the latter. Median presence is also an important safety indicator and should be included in the prediction model because it directly affects the sight distance at intersections, the storage space for crossing vehicles, and the way vehicle cross each other's path. In summary, GUIDE's results are preferable to engineers' judgment and more consistent with their expectations.

Besides all the above-mentioned differences, the final outcomes of GUIDE and CART in Table 4 manifest that the terminal nodes generated by two different tree algorithms are arguably similar. Generally speaking, GUIDE produces more precise categories than CART with the typical nodes such as Nodes [3] and [5] with GEOTYPE and Nodes [2] and [5] without GEOTYPE in CART. For other categories, significant similarities were observed as well. Therefore, the relatively new GUIDE algorithm can be cross-validated by the popularly used CART, and the two tree algorithms can produce consistent results.

Comparisons Between GUIDE and CART

The stratification results were compared and analyzed from the engineering perspective in the previous section, where GUIDE is more favorable than CART. From a statistical point of view, the prediction error, calculated as

$$D_1 = \sum_{t\in\tilde{T}} \sum_{x\in T} \left(y_x - \bar{y}(t)\right)^2$$

is one of the most important measures of model prediction performance. Despite the fact that GUIDE produced more terminal nodes than CART (5 by GUIDE and 3 by CART with GEOTYPE; 7 by GUIDE and 3 by CART without GEOTYPE), the prediction errors from GUIDE, with GEOTYPE (1011.13) or without GEOTYPE (1016.23), are slightly smaller than those of CART (1011.93 with GEOTYPE and 1031.9 without GEOTYPE). Therefore, GUIDE predicted more accurately than CART with more stratification.

Even though the evidence from the intersection case demonstrates that GUIDE outperforms CART from an engineering perspective and a statistical standpoint, the final outcomes (terminal nodes) generated by both tree algorithms are arguably similar and the prediction error margins are relatively small. The following simulation study reveals a clearer and more convincing comparison between GUIDE and CART, especially in the variable selection.

TABLE 4 Comparisons of Terminal Nodes

Cart	Guide
With GEOTYPE	
[4] Geotype {C,D,E,H,I,J,K,L,M,N,O,Q} signal or 3-way	[4] Geotype {E,H,I,J,L,M,N,O,Q} signal or 3-way
[5] Geotype {C,D,E,H,I,J,K,L,M,N,O,Q} unsignalized or flash	[8] Geotype {E,H,I,J,L,N,O} unsignalized or flash [9] Geotype {M,Q} unsignalized or flash
[3] Geotype {A,B,F,G,?}	[6] Geotype {A,B,C,D,F,G,K,?} Signalized [7] Geotype {A,B,C,D,F,G,K,?} Unsignalized
Without GEOTYPE	
[2] 4-legged	[4] 4-legged 3 or 4-way, flash, signal [10] 4-legged side or yield divided [11] 4-legged side or yield undivided
[4] 3 legged unsignalized	[12] 3-legged unsignalized
[5] 3-legged signalized	[8] 3-legged 3 or 4-way, flash, signal, 4-lane [9] 3-legged 3 or 4-way, flash, signal, 2-lane [13] Other leg type unsignalized

SIMULATION CASE STUDY

In the intersection case study, the results from GUIDE are different from those from CART when the variable GEOTYPE (with 16 distinct values excluding unknowns) is treated as one of the splitters (predictors). The outputs from both algorithms, however, are arguably still similar, as is shown in Table 4. In the real world, many examples use significantly more complex variables (such as mixed types) than what is demonstrated here, and GUIDE exhibits superior performance in variable selection under more complicated situations.

A simulated case was constructed to demonstrate the strength of GUIDE's fairness in variable selection. To create a similar situation as the intersection crash data while following the same approach used by Loh (*8*), the response variable *Y* was randomly sampled as counting data from a uniform distribution and the independent variables were generated as categorical data from multinomial distributions. There were three variables in the first simulation case: *X*1 (with 5 levels), *X*2 (with 10 levels), and *X*3 (with 20 levels). There were two variables in the second simulation case: *X*1 (with 3 levels) and *X*2 (with 20 levels). The difference between Simulations 1 and 2 was that the number of observations for each level of categories was the same in Simulation 1, whereas they were different in Simulation 2 (which may be closer to more realistic situations).

For example, *X*1 in Simulation 1 had 1,000 observations from five distinct levels, where the number of observations from each level was approximately 200 ($5 \times 200 = 1{,}000$). Similarly with *X*2, the number of observations from each level was approximately 100 ($10 \times 100 = 1{,}000$), and so on. In Simulation 2, *X*1 had 1,000 observations from three distinct levels, but this time each level had a different number of observations (approximately 200, 300, and 500 for each level). Each simulation case was iterated 100 times and the number of times that the variables were selected is shown for CART and GUIDE. Numbers in parentheses indicate the number of levels.

Simulation 1	*No Partition*	*X1* (*5*)	*X2* (*10*)	*X3* (*20*)	*Total*
CART	36	2	3	59	100
GUIDE	None	36	30	34	100

Simulation 2	*No Partition*	*X1* (*3*)	*X2* (*20*)	*Total*
CART	62	1	37	100
GUIDE	None	55	45	100

From the simulation study, CART seems to be biased toward choosing the variable with more levels as its first splitter, whereas GUIDE tends to choose all the variables with likely probabilities. For instance, in the result from Simulation 2, CART selects $X3$ as its first splitter dominantly over $X1$ and $X2$, whereas GUIDE does not show this bias toward $X3$. Moreover, CART fails to partition variables 36 times out of 100 iterations in the first simulation case, and in the second one, the result is even worse: CART fails to grow the tree 62 times out of 100 iterations. This simulation study shows that GUIDE might be preferable when there are several categorical variables with different numbers of levels.

CONCLUSIONS

In the last decade, there has been increasing interest in the use of classification and regression tree analysis. A tree-based regression model can be constructed by recursively partitioning data by using such criteria as to yield the maximum reduction in the variability of the response. Unfortunately, such an exhaustive search may yield a bias in variable selection and will tend to choose a categorical variable as a splitter that has many distinct values. For instance, if a categorical variable has n distinct values, there are $2^{n-1} - 1$ possible binary splits, which increases exponentially with n. In this paper, a new tree-based regression model, GUIDE, was introduced for its robustness in variable selection bias.

Several important findings were discovered from the intersection crash data:

- When the variable GEOTYPE (which has 16 distinct values) is included in the tree-based regression model, both CART and GUIDE chose it as the first splitter but disagreed on the splitting value. From an engineering perspective, GUIDE is more reasonable because it separates most unsignalized intersection types from signalized intersections.
- Variable selection bias can be detected in CART at the second level, where CART is prone to select the node (Node [2]) with a larger number of distinct values.
- When GEOTYPE is replaced with four simple geometric features, each of which includes fewer distinct values, the sequence of variables entering the model was different in CART and GUIDE. Traffic control is chosen first in GUIDE and the number of legs is chosen first in CART.
- Terminal nodes, the final categories generated by GUIDE and CART, manifest that the outcomes of the two different tree-based algorithms are arguably similar and consistent.

In addition to the intersection field, a simulation study demonstrated that CART seems to be biased toward choosing the variable with more levels as its first splitter, whereas GUIDE tends to pick all the variables with likely probabilities. Furthermore, CART cannot successfully grow or prune a tree in many of the simulation iterations. The simulation study showed that GUIDE may be preferable when there are several categorical variables with differing numbers of levels. In summary, variable selection is the key to the tree-based regression model because it not only determines the level of importance for each variable but also chooses the variable based on statistical significance. As demonstrated through the intersection crash data and a simulation study, the variable selection bias can be overcome by GUIDE. CART, one of the most popular tree-based regression algorithms, should be used with caution when the model includes variables with many distinct values.

REFERENCES

1. Stewart, J. R. Applications of Classification and Regression Tree Methods in Roadway Safety Studies. In *Transportation Research Record 1542*, TRB, National Research Council, Washington, D.C., 1996, pp. 1–5.
2. Karlaftis, M., and I. Golias. Effect of Road Geometry and Traffic Volumes on Rural Roadway Accident Rates. *Accident Analysis and Prevention*, Vol. 34, No. 3, 2002, pp. 357–365.
3. Abdel-Aty, M., J. Keller, and P. A. Brady. Analysis of Types of Crashes at Signalized Intersections by Using Complete Crash Data and Tree-Based Regression. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1908*, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 37–45.
4. Washington, S., and J. Wolf. Hierarchical Tree-Based Versus Ordinary Least Squares Linear Regression Models: Theory and Example Applied to Trip Generation. In *Transportation Research Record 1581*, TRB, National Research Council, Washington, D.C., 1997, pp. 82–88.
5. Washington, S., J. Wolf, and R. Guensler. Binary Recursive Partitioning Method for Modeling Hot-Stabilized Emissions from Motor Vehicles. In *Transportation Research Record 1587*, TRB, National Research Council, Washington, D.C., 1997, pp. 96–105.
6. Washington, S. Iteratively Specified Tree-Based Regression: Theory and Trip Generation Example. *Journal of Transportation Engineering*, ASCE, Vol. 116, No. 6, 2000, pp. 481–491.
7. Park, Y.-J., and F. F. Saccomanno. Collision Frequency Analysis Using Tree-Based Stratification. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1908*, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 121–129.
8. Loh, W. Y. Regression Trees with Unbiased Variable Selection and Interaction Detection. *Statistica Sinica*, Vol. 12, 2002, pp. 361–386.
9. Adams, T. M., E. Juni, M. Sproul, and L. Xu. Regression Tree Models to Predict Winter Storm Costs. In *Transportation Research Record 1948*, TRB, National Research Council, Washington, D.C., 2006, pp. 117–124.
10. Lee, C. Y., B. Ran, and X. Qin. Analysis of Winter Maintenance Logs Using Regression Tree Algorithm. Presented at 85th Annual Meeting of the Transportation Research Board, Washington, D.C., 2006.
11. McCullagh, P., and J. Nelder. *Generalized Linear Models, Second Edition.* Chapman & Hall/CRC, 1989.
12. Breiman L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, Inc., Monterey, Calif., 1984.
13. Dobson, A. J. *An Introduction to Generalized Linear Models, Second Edition.* Chapman & Hall/CRC, 2002.
14. Shih, Y.-S., and H.-W. Tsai. Variable Selection Bias in Regression Trees with Constant Fits. *Computational Statistics and Data Analysis*, Vol. 45, 2004, pp. 595–607.
15. Doyle, P. The Use of Automatic Interaction Detector and Similar Search Procedures. *Operational Research Quarterly*, Vol. 24, 1973, pp. 465–467.
16. Hawkins, D. M. *FIRM: Formal Inference-Based Recursive Modeling, PC version, Release 2.1.* Technical Report 546. School of Statistics, University of Minnesota, 1997.
17. Knapp, K. K., J. Campbell, and C. Kienert. *Intersection Geometry and Crash Characteristics: Intersection-Related Crash Data 2001 to 2003.* Traffic Operations and Safety Laboratory, Department of Civil and Environmental Engineering, University of Wisconsin, Madison, 2005.
18. Lord, D. Modeling Motor Vehicle Crashes Using Poisson-Gamma Models: Examining the Effects of Low Sample Mean Values and Small Size on the Estimation of the Fixed Dispersion Parameter. *Accident Analysis and Prevention*, Vol. 38, No. 4, 2006, pp. 751–766.
19. Little, R. J., and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, New York, 1987.
20. Feelders, A. Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation? In *Principles of Data Mining and Knowledge Discovery*, Springer, Berlin/Heidelberg, 1999, pp. 329–334.
21. Loh, W. Y. *GUIDE (Version 5.2) User Manual.* Department of Statistics, University of Wisconsin, Madison, Nov. 2007. http://www.stat.wisc.edu/~loh/treeprogs/guide/guideman.pdf. Accessed June 2007.
22. Therneau, T. M., and B. Atkinson. *Rpart: Recursive Partitioning.* http://cran.r-project.org/web/packages/rpart/index.html. Accessed June 2007.

The Statistical Methodology and Statistical Computer Software in Transportation Research Committee sponsored publication of this paper.