FACTOR COMPLEXITY OF ACCIDENT OCCURRENCE: AN EMPIRICAL DEMONSTRATION USING BOOSTED REGRESSION TREES

Yi-Shih Chung
Assistant Professor of Logistics and Shipping Management, School of Transportation and Tourism, Kainan University
Taoyuan, Taiwan, R.O.C., e-mail: yishih.chung@gmail.com

Submitted to the 3rd International Conference on Road Safety and Simulation, September 14–16, 2011, Indianapolis, USA

ABSTRACT

Factor complexity is regarded as a typical characteristic of traffic accidents. This paper proposes a novel method, named boosted regression trees (BRTs), which is particularly appropriate for investigating complicated and nonlinear relationships in high-variance traffic accident data. The Taiwan 2004–2005 single-motorcycle accident data are adopted to demonstrate the usefulness of BRTs. Traditional logistic regression and classification and regression tree (CART) models are also developed to compare their estimation results and predictive performance. Both the in-sample cross-validation and out-of-sample validation results show that the increase of tree complexity provides better but declining improvement on the predictive performance, indicating a limited factor complexity of single-motorcycle accidents. While a certain portion of fatal accidents can be explained by the main effects of crucial variables including geographical, time, and socio-demographic factors, the relatively unique fatal accidents are better approximated by interactive terms, especially the combinations of behavioral factors. The BRTs models generally provide better transferability than logistic and CART models. The implications of analysis results for devising safety policies are also provided.

Keywords: boosted regression trees, crash prediction, motorcycle accidents, machine learning.

INTRODUCTION

Complexity is regarded as a typical feature in the occurrence of traffic accidents. Many studies have addressed the importance of controlling confounding factors when modeling traffic accidents, especially in cross-sectional studies where causes are not known a priori (Hauer 2006). The relationship between the response variable and the predictors may be nonlinear, which further increases complexity. For example, the relationship between accident severity and the driver's age is nonlinear. Young and old drivers are more likely to be involved in a fatal accident than middle-aged drivers, typically because young drivers tend to drive fast and old drivers have relatively fragile bodies (Rutter and Quine 1996, Lin *et al*. 2003a, Chang and Yeh

2007). Another example is the relationship between accident occurrence and traffic flow, which is regarded as a concave curve, since a relatively small number of accidents can be observed when traffic flow is extremely low (too few exposures) or high (too congested), and more accidents can be observed for traffic flow volumes in between the two extremes (Qin *et al.* 2004). The interactions between explanatory variables could also be complicated. This effect can be seen from the recent applications of support vector machine (SVM) methods which model factor interactions in a high-order factor space (Li *et al.* 2008b).

Data mining methods are a typical choice to investigate the aforementioned factor complexities. In a series of studies, Wong and Chung (2007, 2008b, a) used rough sets to explore the circumstances that distinguish accident severity. They used 25 variables, including driver characteristics, trip characteristics, behavioral conditions, and road environment, to describe typical circumstances. Their studies indicated that some circumstances, i.e., combinations of factors, are frequently repeated while some circumstances are sparse and unique. In other words, factor complexity did exist for part of the observed accidents; these accidents did not occur merely due to randomness. Chang and Wang (2006) examined the injury severity of traffic accidents in Taiwan using classification and regression tree (CART) models. Their results demonstrated how CART models can provide a satisfactory predictive performance when numerous predictors with multicollinearity concerns are considered. Li *et al*. (2008b) developed SVM models for accident frequencies on rural frontage roads in Texas. Their results suggested that the SVM models have a better predictive performance than the negative binomial models. A nonlinear relationship between average daily traffic (ADT) and crash frequencies was found using sensitivity analysis. To analyze the influential factors on pre-crash maneuvers, Harb *et al*. (2009) combined the techniques of classification trees and random forests; the tree technique was applied to explore the relationship between accident outcomes and selected factors, while the forest technique was adopted to rank the importance of the selected variables. Abdel-Aty and Haleem (2011) analyzed the occurrence of angle crashes at unsignalized intersections using multivariate adaptive regression splines (MARS), a method that can include a great number of variables, nonlinearity, multicollinearity, and a high degree of interaction among predictors. Their results exhibited a nonlinear relationship between annual average daily traffic (AADT) and angle crash frequency.

These studies clearly indicate the complexity of factor effects for traffic accidents; the affecting factors are numerous, possibly related nonlinearly to the response variable, and may be multicollinear with each other. Such features have led to the attempts of using non-parametric modeling techniques, such as rough sets, CART, and SVM, which allow no pre-specification of function form. However, some difficulties remain: which factors should be incorporated in the model, how complicated of the interactions are, and how the results could be interpreted are still a challenge[1].

To shed light on the factor complexity of accident occurrence, this study adopts a novel method, named boosted regression trees (BRTs). The BRTs method is a tree-based data mining method, and thus has advantages such as no need to pre-specify function forms, and the ability to consider numerous predictors and their possible nonlinear relationship with the response variable.

[1] For example, a huge decision tree could be obtained if a loose pruning strategy is applied. Or, the model-training process is a black-box, and little information can be interpreted for accident causality.

Meanwhile, by incorporating statistical techniques such as bagging, boosting, and shrinkage, the BRTs method can simultaneously reduce the variance and bias of prediction errors and gradually focus on the difficult cases (i.e., relatively unique traffic accidents). This advantage is particularly crucial to accident modeling because traffic accidents are typically unique and highly imbalanced (e.g., fatal accidents only account for a small portion of the total). Due to these statistical techniques, the BRTs method also provides interpretable results. Details of the BRTs models will be introduced in the following section.

To demonstrate the usefulness of BRTs, an empirical dataset of single-motorcycle accidents is adopted and accident severity (fatal vs. non-fatal) is analyzed. As vehicles, motorcycles offer consumers the advantages of low initial cost and, for some models, good fuel efficiency. High fuel prices in recent years have led to an increasing number of registered motorcycles in some countries. In the United States, there are more than 6.2 million registered motorcycles. More than five thousand motorcyclists were killed in 2009, accounting for 12 percent of all highway fatalities (NHTSA 2009). The situation is even worse in developing countries, where powered two-wheelers are a primary mode of transportation in urban areas. For example, motorcycles account for two-thirds of all registered vehicles in Taiwan, and 45 percent of traffic accidents involve motorcyclists (MTC 2007). Single-motorcycle accidents are those that involve only one vehicle (motorcycle). Although single-motorcycle accidents account for a relatively small portion of accidents, they are usually serious. In addition, the occurrence of single-vehicle accidents is expected to be simpler to study than that of multi-vehicle accidents, which is appropriate for this preliminary study to investigate factor complexity of accident occurrence.

Theoretically, the BRTs models can provide satisfactory performance. Yet, to demonstrate the transferability of BRT models with empirical data, logistic regression and CART models are also developed and compared. These two basic models are chosen instead of advanced econometric models (e.g. ordered probit/logit or mixed logit models (Kockelman and Kweon 2002, Milton *et al.* 2008)) or other data mining and soft computing models (e.g. rough sets, SVM, random forests, or MAR) based on two rationales. First, the effectiveness of these two models has been demonstrated in past studies, especially the logistic regression models (Al-Ghamdi 2002, Bedard *et al.* 2002, Valent *et al.* 2002). Second, using advanced econometric models requires delicate model specification and, sometimes, more assumptions on function forms and parameters. The aforementioned complexity of accident occurrence poses challenges of using such models. On the other hand, logistic and CART models provide a good start to compare with the BRTs models.

The remaining parts of this paper are organized as follows. The following section introduces the methodology including a brief introduction of boosted regression tree models, the data and variables, and the analysis procedure. This paper then presents the analysis results, followed by discussions. The concluding remarks are presented in the final section.

METHODOLOGY

Boosted Regression Trees

Boosted regression trees can be characterized by two terms: regression trees and boosting. A BRTs model grows a number of trees by bootstrapping the training data, i.e., randomly selecting a certain proportion of observations from the training data with replacement. Each tree grows as developing a CART, a form of binary recursive partitioning. The term "binary" implies that each group of traffic accidents, represented by a "node" in a decision tree, can only be split into two groups (i.e., a parent node can only have two child nodes). The term "recursive" refers to the fact that the binary partitioning process can be applied over and over again. Lastly, the term "partitioning" refers to the fact that the dataset is split into sections or partitioned. Splitting functions, which measure the purity (or impurity) of a tree, are applied to determine which variable should be included to split the tree; common functions include Gini, Twoing, and Entropy. To prevent overfitting data, trees are typically pruned to cut off the nodes (or branches) resulting in high classification costs (Chang and Wang 2006). A complexity parameter, usually defined as a cost function of misclassification of data, is used to determine which node to prune. Finally, the best tree can be selected using cross-validation or out-of-sample validation.

Despite the advantages of CART models, a single tree is sometimes a weak classifier, especially for high-variance data such as the data for traffic accidents. To deal with this issue, the BRTs model introduces a technique termed as bagging. To control the effects of confounding factors, numerous predictors are usually included in modeling classification and regression trees, which typically results in a model with high variance and low bias. Bagging is a technique for reducing the high variance and involves the following steps: 1) take a bootstrap sample from the training dataset; 2) fit the tree to this bootstrapped dataset; 3) repeat the previous two steps a certain number of times (typically 50–1000); and 4) make predictions for new data using each of the fitted models and average the predictions. The principle behind the bagging technique is used in the random forests method. Random forests develop each tree by taking a bootstrap sample and selecting a random subset of predictors (Harb *et al.* 2009). The randomly selected predictors reduce the correlations between predictors and thus reduce the variance component of prediction error.

In addition to bagging, the BRTs model applies a special mechanism to bootstrap samples, named boosting. Boosting uses the same principle of bagging that a given weak algorithm is repeatedly run, and the computed classifiers are combined in the final estimation or prediction. In other words, boosting, like bagging, can effectively reduce the variance. Yet, while conventional bagging focuses on randomly selecting observations from the original data with replacement, boosting further considers the hardness of the training cases; when repeatedly selecting sub-datasets, boosting tends to generate distributions that concentrate on the harder training cases (Freund and Schapire 1996). This feature is crucial in accident studies because fatal accidents typically account for only a small portion of all accidents.

The algorithm for developing BRTs models is as follows. Suppose we want to build a function $f(x)$ to approximate a response y where x is a vector of predictors. To estimate the function, a loss function is typically specified; for example, a squared-error loss function,

$L(y, f(x)) = (y - f(x))^2$, is mainly used to estimate a linear regression with function form $f(x) = x\beta$ where β is a matrix of parameters. For CART models, additive models (Hastie *et al.* 2009) express $f(x)$ as a sum of basis function $b(x; \gamma_m)$ as follows:

$$f(x) = \sum_m f_m(x) = \sum_m \beta_m b(x; \gamma_m).$$

For boosted trees, the function $b(x; \gamma_m)$ represents individual trees, with γ_m defining the split variables, their values at each node, and the predicted values. The β_m values represent weights given to the nodes of each tree in the collection and determine how predictions from the individual trees are combined (De'ath 2007).

To estimate parameters, the gradient boosting technique is applied (Friedman 2001). Its procedure can be summarized as follows (De'ath 2007):

1) Initialize $f_0(x) = 0$.
2) For $m = 1$ to n:
    a. Calculate the residuals, $r = -([\partial L(y, f(x))]/[\partial f(x)])_{f(x)=f_{m-1}(x)}$.
    b. Fit a least-squares regression tree to r to get the estimate of γ_m of $\beta b(x; \gamma)$.
    c. Get the estimate β_m by minimizing $L(y, f_{m-1}(x) + \beta b(x; \gamma_m))$.
    d. Update $f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$.
3) Calculate $f(x) = \sum_m f_m(x)$.

Step 2a calculates the residuals as the negative of the first derivative of the loss function evaluated for the current value of $f(x)$. Step 2b uses a least-squares regression tree to estimate γ_m. Least-squares trees are used irrespective of the chosen loss function and are computationally very efficient (De'ath 2007). Step 2c then estimates the values β_m assigned to the nodes of the tree to minimize the overall loss.

To reduce the effect of overfitting, the boosted regression tree further applies a shrinkage strategy. A learning rate, ϵ, is introduced at step 2d when the algorithm updates the estimated function:

$$f_m(x) = f_{m-1}(x) + \epsilon\beta_m b(x; \gamma_m) \qquad \text{where } 0 < \epsilon \leq 1.$$

A smaller learning rate requires more iterations (i.e., trees) in the boosting sequence. Studies indicated that a 10-fold reduction in learning rate requires an approximately 10-fold increase in iterations (De'ath 2007), and at least 1,000 trees are recommended (Elith *et al.* 2008).

Subjects and Data

The subjects used to demonstrate the factor complexity are single-motorcycle accidents. Single-motorcycle accidents are those in which only a single motorcycle is involved. The data include two years (2004–2005) of single-motorcycle accidents, provided by the National Police Agency of Taiwan. The total number of single-motorcycle accidents was 7,634 in 2004 and 9,869 in 2005, with fatal accident rates of 3.52%, and 3.98%, respectively. The extremely low fatal accident rates indicate the adopted dataset is highly imbalanced.

Variables

The dataset contains 29 variables as summarized in Table 1. The dependent variable is the severity of the accidents, coded as a binary variable with value 1 if fatal and 0 otherwise. The

remaining 28 variables include driver characteristics, trip characteristics, driving behavior, weather conditions, and road environment. All the variables are categorical variables except driver's age, speed limit, and hour.

Table 1 Variables to develop single-motorcycle accident models

Category	Variable	Definition	Type
Dependent variable	Severity	Fatal, Injury only	Binary
Driver characteristics	Age		Continuous
	Gender	Male, Female (2 types)	Categorical
	License type	Trucks, Buses, Automobiles, Motorcycles, etc. (16 types)	Categorical
	Occupation	Students, Administration, Education, Engineering, etc. (21 categories)	Categorical
	License condition	With proper license, Drive w/o license, Revoked license, etc. (7 conditions)	Categorical
Trip characteristics	Trip purpose	School, Work, Business, Social activity, Shopping, etc. (9 categories)	Categorical
	Month	January, February, …, December (12 months)	Categorical
	Day of Week	Monday, Tuesday, …, Sunday (7 days)	Categorical
	Hour	0–23	Continuous
	County	Taipei city, Taipei county, etc. (25 counties)	Categorical
Driving behavior	Protection equipment	Wear (helmet), Not wear, Others (3 categories)	Categorical
	Cellphone use	No use, Handheld, Earphone, Hands free, Others (4 types)	Categorical
	Movement prior to accident	Going straight, Left turn, Right turn, etc. (14 types)	Categorical
	Drinking condition	No drinking, BAC < 0.05%, etc. (8 categories)	Categorical
Weather condition	Climate	Sun, Cloud, Rain, Fog, etc. (7 conditions)	Categorical
Road environment	Illumination	Day light, Night with illumination, etc. (4 types)	Categorical
	Road level	Highway, Arterial roads, Streets, etc. (7 levels)	Categorical
	Road type	3-way junctions, straight road, etc. (17 types)	Categorical
	Road location	Within intersections, Fast lane, Mixed lane, etc. (21 types)	Categorical
	Pavement type	Asphalt, Cement, Rubble, Others, None (5 types)	Categorical
	Surface condition	Dry, Wet, Muddy, Slippery, Snow (5 conditions)	Categorical
	Surface deficiency	None, Holes, Bumping, Soft (4 types)	Categorical
	Obstacles	None, Work zone, Fixed objects, Others (5 types)	Categorical
	Sight distance	Good, Curve road, Others, etc. (7 types)	Categorical
	Signal type	Regular traffic light, Flash, etc. (4 types)	Categorical
	Median type	Median, Markers, Marking, etc. (10 types)	Categorical
	Roadside	With marking, Without marking (2 types)	Categorical
	Speed limit	Kilometers per hour	Continuous

Analysis Procedure

Based on the 2004 single-motorcycle accident data, this study develops three types of models: the BRTs models, the logistic regression models, and the classification and regression tree (CART) models. This study develops the BRTs models, mainly following the suggestions by Elith *et al*. (2008). The BRTs models are built using the software *R* (R Development Core Team 2009) with the package *gbm*.

Three parameters are jointly considered to optimize the BRTs models, the number of trees, learning rate, and tree complexity. This study does not particularly control the number of trees as long as it stays at a reasonable size, 1,000–10,000[2]. On the other hand, this study tests the combinations of varying values of learning rates (0.05–0.0001) and tree complexity levels (1–18) to develop the best BRTs model. Meanwhile, to reduce overfitting and improve accuracy, trees are boosted based on random draws from the full training dataset. In this study, 50% of the data are drawn at random without replacement at each iteration.

A model with zero training error is overfit to the training data and will typically generalize poorly. To prevent this problem and determine the best setting of the BRT model for the 2004 single-motorcycle accidents, the cross-validation (CV) technique is applied when the various combinations of learning rates and tree complexity levels are examined. In particular, 10-fold CV is chosen, and predictive deviance is applied to measure the success of the models. Because the dependent variable is binary, the BRTs models are a form of logistic regression that models the probability that a fatal traffic accident occurs, $y = 1$, with explanatory variables $\boldsymbol{X}$, $P(y = 1|\boldsymbol{X}) = f(\boldsymbol{X})$. The Bernoulli loss function[3] is chosen as the deviance for the binary response variable. All 28 explanatory variables listed in Table 1 are used to develop the BRT models. Variables are implicitly selected by down-weighting variable contributions at each iteration (Elith *et al.* 2008), known as a shrinkage method in data mining.

The numerous explanatory variables and categories challenge the model specification of logistic regression models. To comprehensively account for the factor effects, this study applies a general-to-specific approach to develop the logistic regression models; all the explanatory variables are considered in the initial model, and then non-significant variables are dropped based on test statistics including deviance, the Wald statistic, Hosmer-Lemeshow tests, and the Akaike Information Criterion (AIC) measure. For categorical variables, the non-significant categories are collapsed considering their practical definition.

The CART models are developed using the cost-complexity pruning strategy, meaning that the tree growing process is stopped only when some node size is reached that minimizes the cross-validated errors (Hastie *et al.* 2009). The Gini function is chosen as the splitting function.

---

[2] The rule of thumb suggested by Elith et al. (2008) is fitting models with at least 1,000 trees. The analysis results, as presented at the following sections, show that all models converge within 10,000 trees.

[3] The Bernoulli deviance: $-2\frac{1}{\sum w_i}\sum wi(y_i f(\mathbf{x}_i) - log(1 + exp(f(\mathbf{x}_i))))$, where y_i is the reponse, $\mathbf{x}_i$ is the vector of explanatory variables, and w_i are the observation weights (Ridgeway, G., 2007. Generalized boosted models: A guide to the gbm package.) Equal weights are used in this study.

The 2005 single-motorcycle accident data are used to examine the out-of-sample predictive performance for the developed BRT, logistic, and CART models using the 2004 data.

ESTIMATION RESULTS OF BOOSTED REGRESSION TREE MODELS

Optimal Setting

A learning rate of 0.01 is too fast for most BRTs models for which the fitting process stops at a tree size smaller than 1000. The only exception is the model with tree complexity of 1; as shown in the top left panel of Figure 1, the model stops growing trees (i.e., the predictive deviance becomes flat) at a tree size just over 1000 (the vertical gray dotted line). In other words, to obtain robust learning results, the learning rate for the 2004 single-motorcycle accidents should be set to at least 0.005. On the other hand, a learning rate of 0.0005 is only too slow for some low-tree complexity models such as a tree complexity of 1, but appropriate for most tree complexities. Models with a learning rate of 0.0001 are unreported because this extremely slow learning rate makes the growth of most trees unstoppable before the tree size reaches 10,000. Moreover, reducing learning rates does not decrease the predictive deviance when the tree complexity exceeds a certain level. Figure 1 shows that models with a learning rate of 0.001 (green line) generally exhibit lower predictive deviance than those with a learning rate of 0.0005 (blue line), and also lower than those with a learning rate of 0.005 when tree complexity exceeds a certain level, for example 10.

Increasing tree complexity consistently improves the predictive deviance; however, the improvement decreases as the tree complexity increases. The bottom right panel of Figure 1 shows that when the learning rate is fixed at 0.001, the predictive deviance continuously reduces when the tree complexity grows. Yet the predictive deviance lines fitted by the models with a tree complexity of 17 (pink line) and of 18 (yellow line) almost overlap, indicating their close predictive performance. To sum up, the model with a learning rate of 0.001 and tree complexity approximately 18 is the best 2004 single-motorcycle accident BRTs model based on the cross-validation results.

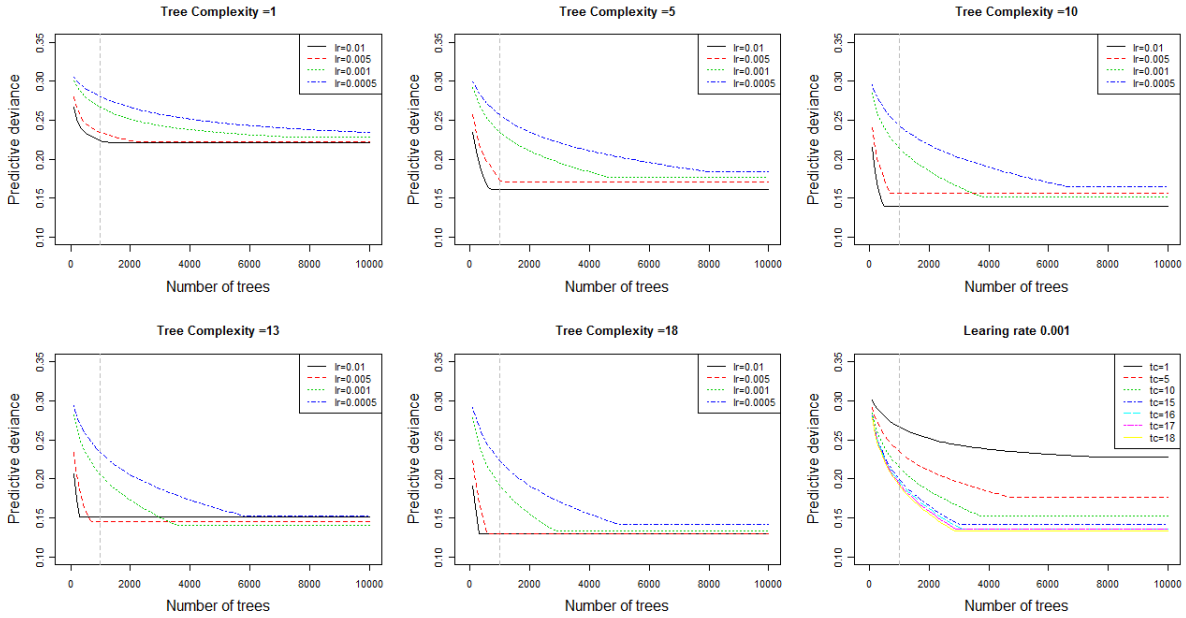

Figure 1 Predictive deviance against number of trees for models fitted with various tree complexities and learning rates

Relative Contributions of Explanatory Variables

With the learning rate fixed at 0.001, the relative contributions of explanatory variables for boosted regression tree models with various tree complexity levels (tc = 3, 8, 11, 17, and 18) are summarized in Table 2. The relative contributions are measured as the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Elith *et al.* 2008).

While the predictive deviance is incrementally reduced with the increase of tree complexity, the relative contributions of explanatory variables are comparatively stable. As shown in Table 2, regardless of tree complexity, the county variable is recognized as the most influential variable, implying that factors that are geographically heterogeneous such as local driving culture, the relatively long distance from hospitals in rural areas, or road design elements (Eiksund 2009) are crucial to distinguish the severity of single-motorcycle accidents. The occupation and month variables are in the second- and third-most influential variables; the occupation variable may suggest the different lifestyles and motorcycle usage (Lin *et al.* 2003b, Bina *et al.* 2006), and the month variable indicates that factors associated with seasonal variation are critical to differentiate the severity of single-motorcycle accidents.

The following two factors, drinking conditions and cellphone use, have been intensively discussed in past studies; driving under the influence of alcohol and using a cellphone while driving are more likely to increase the possibility of traffic accidents and their severity. Road location is recognized in the sixth place of influential variables. Wong and Chung (2007, 2008b) showed that some road locations such as intersections where relatively more fixed objects are set raise the possibility of bump-into-fixed-object traffic accidents and their severity. Usage of protective equipment, i.e., wearing a helmet, and day of the week are the next two influential

variables. Helmet wearing has shown a significant effect on reducing the severity of motorcycle accidents by protecting the brain in both longitudinal and cross-sectional studies (e.g. Hotz *et al.* 2002, Hundley *et al.* 2004). The median-type variable, accounting for 2 to 3 percent of relative contributions, is ranked ninth. While road levels affect the selection and construction of median types, a wide median island creates fixed objects on the road and therefore increases the possibility of traffic accidents and their severity (Wong and Chung 2007, 2008b). The age variable is the only continuous variable in the top 10 most influential variables, and suggests the influence of various physical conditions of motorcyclist age groups on the fatality of traffic accidents.

As shown in Table 2, most of the top 10 most influential variables are variables related to driver characteristics, trip characteristics, and driving behaviors. The only weather-condition variable and most road environment variables contribute only a small portion to the fatality of single-motorcycle accidents.

Table 2 Relative contributions (%) of explanatory variables for boosted regression tree models with various tree complexity levels*

tc = 3		tc = 8		tc = 11		tc = 17		tc = 18	
Variable	Relative contributions	Variable	Relative contributions	Variable	Relative contributions	Variable	Relative contributions	Variable	Relative contributions
County	26.7126	County	22.8054	County	22.6009	County	22.1200	County	21.7453
Drinking condition	15.4440	Occupation	13.8236	Occupation	13.9133	Month	13.9536	Month	14.1118
Occupation	13.5531	Month	11.8642	Month	12.7811	Occupation	13.7416	Occupation	13.7439
Cellphone use	13.5344	Drinking condition	10.4327	Drinking condition	9.6189	Drinking condition	8.5430	Drinking condition	8.5410
Road location	6.3239	Cellphone use	9.0465	Cellphone use	8.4542	Cellphone use	8.1362	Cellphone use	8.0936
Protection equipment	5.6958	Road location	6.6369	Road location	6.7089	Road location	6.8731	Road location	6.8667
Month	5.2931	Protection equipment	4.1089	Day of week	4.1816	Day of week	4.4924	Day of week	4.5932
Day of week	1.8697	Day of week	3.9049	Protection equipment	3.9250	Protection equipment	3.6529	Protection equipment	3.6868
Age	1.4053	Median type	2.2975	Median type	2.7229	Median type	3.0675	Median type	3.0990
License type	1.2994	Age	2.1879	Age	2.2911	Age	2.2750	Age	2.4188
Road type	1.2803	Road level	2.0198	Road level	2.0380	Road level	2.2456	Road level	2.3407
License condition	1.2256	Road type	1.8958	Road type	1.9833	Road type	2.1512	Road type	2.0348
Movement prior to accident	1.1808	Hour	1.6640	Hour	1.7253	Hour	1.8155	Hour	1.8205
Hour	1.0194	Trip purpose	1.4865	Trip purpose	1.4934	Trip purpose	1.5495	Trip purpose	1.5819
Trip purpose	0.9707	License condition	1.3306	License condition	1.4487	License condition	1.4016	License condition	1.3704
Illumination	0.8189	License type	1.2384	License type	1.1236	License type	1.0969	License type	1.0380
Sight distance	0.6142	Illumination	0.9549	Illumination	0.8746	Illumination	0.8923	Illumination	0.9179
Median type	0.6124	Movement prior to accident	0.7745	Movement prior to accident	0.6375	Movement prior to accident	0.5064	Sight distance	0.5180
Road level	0.5785	Sight distance	0.6315	Sight distance	0.5695	Sight distance	0.4739	Movement prior to accident	0.4836
Speed limit	0.4111	Speed limit	0.5032	Speed limit	0.4722	Speed limit	0.4350	Speed limit	0.3999
Climate	0.1321	Climate	0.2227	Climate	0.2546	Climate	0.3021	Climate	0.3073
Gender	0.0127	Roadside	0.0874	Roadside	0.1208	Roadside	0.1598	Roadside	0.1844
Roadside	0.0069	Gender	0.0464	Gender	0.0265	Gender	0.0496	Gender	0.0450
Signal type	0.0031	Surface condition	0.0193	Surface condition	0.0188	Surface condition	0.0291	Surface condition	0.0240
Surface condition	0.0023	Signal type	0.0080	Obstacles	0.0097	Obstacles	0.0205	Obstacles	0.0202
Pavement	0.0000	Obstacles	0.0078	Signal type	0.0056	Signal type	0.0157	Signal type	0.0126
Surface deficiency	0.0000	Surface deficiency	0.0006	Pavement type	0.0000	Pavement type	0.0000	Surface deficiency	0.0006
Obstacles	0.0000	Pavement type	0.0000	Surface deficiency	0.0000	Surface deficiency	0.0000	Pavement type	0.0000

*Learning rates fixed at 0.001

Marginal Effects of Explanatory Variables

To further investigate the effect of the most influential variables, the partial-dependence plots that show the effect of a variable on the fatality of single-motorcycle accidents after controlling for the average effects of all other variables in the model, are illustrated in Figure 2. These results are from the BRT with a learning rate of 0.001 and tree complexity of 17.

Figure 2(a) shows the various effects provided by geographical factors in Taiwan. The most influential regions are those located in the middle of western Taiwan, including HsinChu County, Changhua County, and Chiayi City and County, and one eastern region, Hualien County. These regions are mostly classified in the third levels of administrative bureaucracy in Taiwan, and typically have a tighter budget in public construction including road infrastructure. The poorer road quality implies less protection for motorcyclists when an accident occurs, and thus leads to a higher fatality rate. HsinChu County, Changhua County, and Chiayi City have a population density as high as the counties in the first-level administrative bureaucracy. These regions are expected to have many economic activities and thus create a lot of trips. The many trips in a poorer-quality road network may be a reason for the high fatality rates. The primary industry in Chiayi and Hulien Counties is tourism, and a certain portion of tourists drive motorcycles to visit these regions. The unfamiliar road environment to tourist motorcyclists increases the possibility of traffic accidents and severity. The eastern county, Hualien, has the largest area and the lowest motorcycle density (per kilometer square), which may imply a higher driving speed on average. Finally, the police-to-population ratio of Hualien is one of the lowest in Taiwan, which suggests a lower level of police enforcement and a higher possibility of violations such as speeding.

Figures 2(b) and 2(g) demonstrate the various effects of seasonal factors. Figure 2(b) shows that months January, March, July, and November are associated with higher fatality rates. January, March, and July are the months for lunar New Year, spring vacation, and summer vacation, respectively. Figure 2(g) exhibits a higher fatality rate on typical working days, Tuesday, Wednesday, and Thursday, as well as Sunday.

Figure 2(c) and 2(j) describes two driver characteristics, occupation and age, related to fatality rates. Motorcyclists who are high school students, bus or railroad occupational drivers, or police officers are associated with higher fatality rates in single-motorcycle accidents. High school students who are mostly under age 18 may not legally drive a motorcycle; moreover, the lifestyle of students is typically different from others at a similar age, which might also lead to a higher possibility of accidents and fatality rates (Lin *et al.* 2003a). A certain portion of police officers and occupational drivers require shift work in Taiwan; such workers are more likely to have sleep problems and a higher level of pressure from work, which consequently result in a higher possibility of traffic accidents and severer injury levels.

The age variable demonstrates a nonlinear marginal effect on the probability of fatal single-motorcycle accidents as illustrated in Figure 2(j). As expected, the older motorcyclists are more likely to be involved in a fatal accident than other age groups, especially when the motorcyclists are older than 60. The motorcyclists who are younger than 20 also demonstrate a certain level of marginal effect on the probability of being involved in a fatal single-motorcycle accident. The motorcyclists at an age around 40 show the lowest marginal effect on the probability of being

involved in a fatal single-motorcycle accident. The motorcyclists at this age are expected to have accumulated a certain level of driving experience; they also have relatively mature physical and psychological conditions (compared to young drivers) with a well-functioning body (compared to older drivers) (Yagil 1998). Consequently, this age group is associated with a lower fatality rate.

Figures 2(d), (e), and (h) illustrate the marginal effect of three driving behavioral variables, drinking condition, cellphone use, and protection equipment use, on the probability of fatal single-motorcycle accidents. Figure 2(d) shows that while sober motorcyclists are associated with the lowest level of marginal effects on the fatality rate, a nonlinear relationship is found for those driving under the influence of alcohol. The motorcyclists who are heavily drunk demonstrate the largest effect on the fatality rate, followed by slightly drunk motorcyclists. The motorcyclists with blood alcohol content in the middle range, i.e., between 0.26–0.55 micrograms per liter, have a relatively lower marginal effect. Motorcyclists who are heavily drunk cannot maneuver the motorcycle well, neither can they protect themselves if an accident occurs, and consequently are associated with a high fatality rate. On the other hand, slightly drunk motorcyclists may easily ignore their deteriorating physical conditions, thus leading to a higher fatality rate. Figure 2(e) shows the marginal effect of cellphone use on the fatality rate, and the unknown category exhibits the highest marginal effect. This result indicates the difficulty of reporting cellphone use for traffic accidents. Finally, Figure 2(h) shows that not wearing helmets is connected to an extremely high marginal effect on the fatality rate of single-motorcycle accidents, consistent with past studies (Li *et al.* 2008a).

Figures 2(f), (i), (k), and (l) illustrate the marginal effects provided by the four road-environmental variables. Figure 2(f) indicates the road locations connected to a relatively higher fatality rate of single-motorcycle accidents including exclusive bus lanes, nearby ramps, and motorcycle waiting zones. Exclusive bus lanes are typically designed for areas with a high population density as well as high public transportation demand. A road segment equipped with exclusive bus lanes is typically wider and has a higher speed limit. Its road geometric design is also more complicated than other roads. In other words, the road environment encourages a high driving speed, and requires the motorcyclists to pay attention to the complicated design, consequently leading some motorcyclists into fatal accidents. Similar reasoning can be applied to explain the significant effect of the vicinity of ramps and motorcycle waiting zones. The roads approaching highway ramps are typically wide and have a high speed limit for vehicles to enter highways[4]. The motorcycle waiting zones are designed for motorcyclists to turn left at a wide intersection (i.e., two or more lanes in one direction). Its speed limit is usually high and has a relatively complicated geometry, compared to narrow intersections.

The median types associated with high fatality rates of single-motorcycle accidents are narrow median islands (shorter than 50 centimeter) and markings that prohibit overtaking. While the installation of median islands can prevent conflicts between vehicles from opposite directions, it also creates fixed objects on the road and raises the probability of accidents and severity. The markings that prohibit overtaking are typically drawn on the road segments approaching intersections or without sufficient sight distance. Fatal single-motorcycle accidents at these locations may suggest high speed, which is inappropriate for these locations.

[4] In Taiwan, no motorcycles are allowed to drive on national highways.

Figure 2(k) demonstrates a nonlinear relationship between the road levels and their marginal effect on the fatality rate of single-motorcycle accidents. Significant marginal effects are observed on the high- and low-level roads, while small marginal effects are seen on the middle-level roads. High-level roads such as national and provincial highways have a high speed limit, and the accidents are expected to be severer due to the driving speed. On the other hand, the low-level roads cannot provide sufficient protection for motorcyclists if an accident occurs, and therefore, relatively more fatal accidents are observed on them. Finally, Figure 2(l) shows that a higher fatality rate is associated with road types requiring more sophisticated driving skills including roundabouts, culverts, elevated roads, and graded roads.

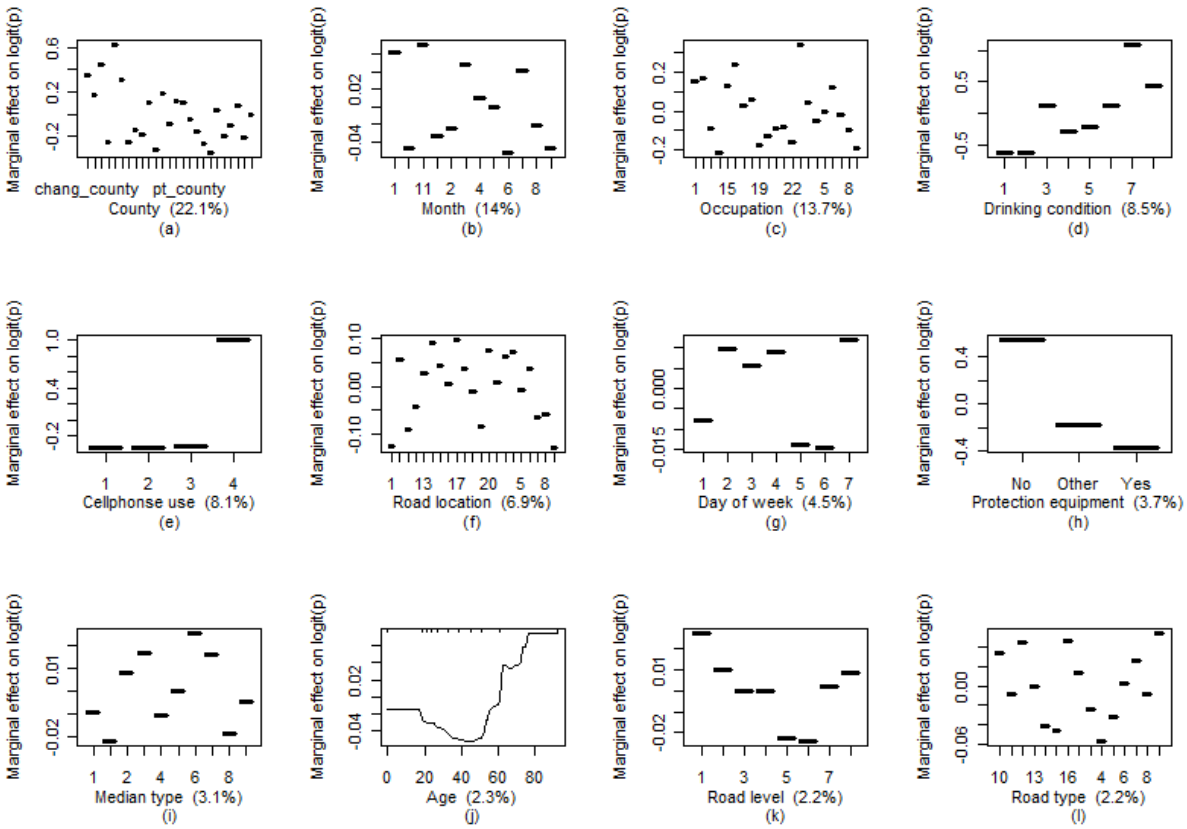

Figure 2 Partial-dependence plots for the 12 most influential variables in the model for fatal single-motorcycle accidents

Important Interactions

The pairwise interactions with effect size greater than 10 for models with various tree complexity levels are illustrated in Figure 3. While the 28 explanatory variables considered can generate 378 combinations of variable pairs, only a few of them play a crucial role in explaining the variance of the fatality for single-motorcycle accidents. No matter what the tree complexity is, the analysis shows that up to seven variable pairs contribute an effect size greater than 10. Moreover, those same variable pairs play the most critical roles across all the BRT models.

Among the seven variable pairs, the two pairs that combine two behavioral variables demonstrate an explicitly upward trend when the tree complexity increases. The first pair is the combination of cellphone use and protection equipment use. As shown in Figure 3, the effect size of the cellphone-protection pair (black line) becomes extremely prominent when tree complexity increases. The other pair is the combination of drinking condition and protection equipment use. Its effect (green line) stably increases as tree complexity rises. The combination of occupation and drinking condition is the last variable pair that shows an upward trend (pink line) although its increase is relatively small. These results indicate that when the occurrence of single-motorcycle accidents is modeled with more complicated interactions (i.e., a higher level of tree complexity), the interaction between behavioral variables plays a more important role. On the other hand, the three interactions that involve the county variable exhibit a bumpy but relatively flat trend, including protection-county (red line), drinking-county (blue line), and cellphone-county (light blue line). These results suggest that no matter how comprehensively the traffic accidents are modeled, the interaction between behavioral variables and geographically heterogeneous factors (represented by the county variable) has a relatively stable effect.

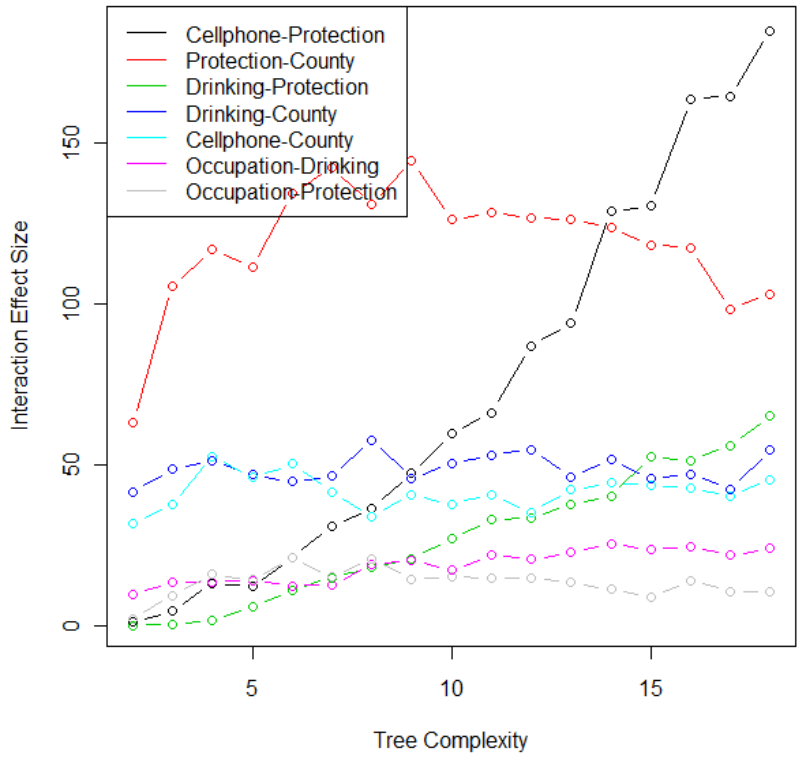

Figure 3 Top seven interaction effects for boosted regression tree models with various tree complexity levels

Out-of-Sample Prediction Using 2005 Data

While the cross-validation results suggest that the BRT model with tree complexity around 18 and learning rate 0.001 has the lowest predictive deviance, the 2005 single-motorcycle accidents are adopted to investigate the out-of-sample predictive performance. A logistic regression and a CART model that were developed using the 2004 single-motorcycle accident data are also tested for their out-of-sample predictive performance with the 2005 data. The best logistic regression model developed with the 2004 data contains 12 variables where variables are selected and categories are merged using deviance and Wald z tests. The Hosmer-Le Cessie omnibus test fails to find evidence of a lack of fit.

The estimation results of logistic models are summarized in Table 3. One driver characteristic, gender, is included in the logistic model, indicating that the odds of female motorcyclists being involved in a fatal single-motorcycle accident are 0.66 times those of male motorcyclists.

The next four variables are trip characteristics. The trip-purpose variable suggests that sightseeing trips have 4.03 times the risk of being involved in a fatal single-motorcycle accident compared to trips with work, school, or business purposes. The month variable indicates a seasonal effect that November is associated with a significantly positive effect on the fatality of single-motorcycle accidents. The county variable significantly contributes to explaining the response variable where almost all the 24 counties (one county is chosen as the reference category) demonstrate a significant effect.

Three behavioral variables are included in the model. The results in Table 3 indicate that motorcyclists wearing helmets have about one-quarter the risk of being involved in a fatal single-motorcycle accident compared to those who do not wear helmets. While motorcyclists who use handheld and hand-free cellphones are insignificantly different from those who do not use cellphones, motorcyclists who have an unknown cellphone use have 6.35 times the risk of being involved in a fatal single-motorcycle accident compared to those who do not use a cellphone. As for the drinking condition, motorcyclists who are slightly drunk or heavily drunk have a significantly high odds of being involved in a fatal single-motorcycle accident; in particular, the slightly drunk motorcyclists have 5.71 times and the heavily drunk motorcyclists have 4.93 times the risk of being involved in a fatal accident compared to those who do not drink. The result reveals a "U"-shaped relationship between alcohol consumption level and accident severity. The relationship may result from two possibilities: reckless driving is more often on intoxicated drivers compared to sober ones, and the adverse physiological effects of alcohol on the body (Bedard *et al.* 2002).

Four road environment variables show significant effects in the logistic model. Some road locations have a significant connection with fatal accidents. A single-motorcycle accident that occurs at the roadside has 3.33 times the risk of being a fatal accident compared to one that occurs within an intersection. A road segment with median markings or without medians is less likely to have a fatal single-motorcycle accident compared to a road segment with median islands or markers; the odds are about 0.78. Roadside and illumination have non-significant estimation results but are selected due to their improvement on the Akaike's Information Criterion.

Table 3 Estimation results of the logistic model using 2004 data

Variable	Category	Estimate	Odds ratio	Variable	Category	Estimate	Odds ratio
Gender	Female	−0.415*	0.660	Protection equipment	Yes	−1.356***	0.258
Trip purpose	Sightseeing	1.394*	4.031		Other	−1.745***	0.175
	Others	0.457*	1.580	Cellphone use	Handheld	−15.440	0.000
Month	November	0.424.	1.528		Handfree	−1.300	0.273
Hour		−0.023*	0.977		Other	1.849***	6.353
County	County 2	−0.961	0.382	Drinking condition	No alcohol response	0.592*	1.808
	County 3	0.071	1.074		BAC < 0.25 mg/L	1.743***	5.714
	County 4	−2.160**	0.115		0.26 < BAC < 0.55 mg/L	0.798	2.220
	County 5	0.883.	2.417		> 0.55 mg/L	1.595***	4.928
	County 6	−0.085	0.919		Cannot detect	3.130***	22.874
	County 7	−1.679**	0.187		Other	2.114***	8.281
	County 8	−1.779**	0.169	Illumination	Nighttime with illumination	−0.225	0.799
	County 9	−15.610	0.000	Road location	Near intersection, median island, fast, slow and mixed lanes	0.505*	1.656
	County 10	−0.162	0.850				
	County 11	−1.717***	0.180		Roadside	1.204***	3.333
	County 12	−16.560	0.000		Other	1.046**	2.846
	County 13	−1.265*	0.282	Median type	Markings or none	−0.248.	0.780
	County 14	−0.758.	0.468	Roadside	With marking	−0.159	0.853
	County 15	−0.802	0.448	Intercept		−3.087***	0.046
	County 16	−1.010**	0.364				
	County 17	−1.135*	0.321				
	County 18	−2.021**	0.133				
	County 19	−2.105***	0.122				
	County 20	−0.882*	0.414				
	County 21	−1.081***	0.339				
	County 22	−0.786.	0.456				
	County 23	−0.583	0.558				
	County 24	−1.997***	0.136				
	County 25	−1.675***	0.187				

***<0.001, **<0.01, *<0.05, .<0.10

The 2004 CART model is developed using the Gini splitting function as illustrated in Figure 4. The model contains the following variables: cellphone use, county, drinking condition, sight distance, road location, month, occupation, and road type. The result shows that most of the variables selected for the CART model are also the influential or significant variables for the BRT and logistics models.

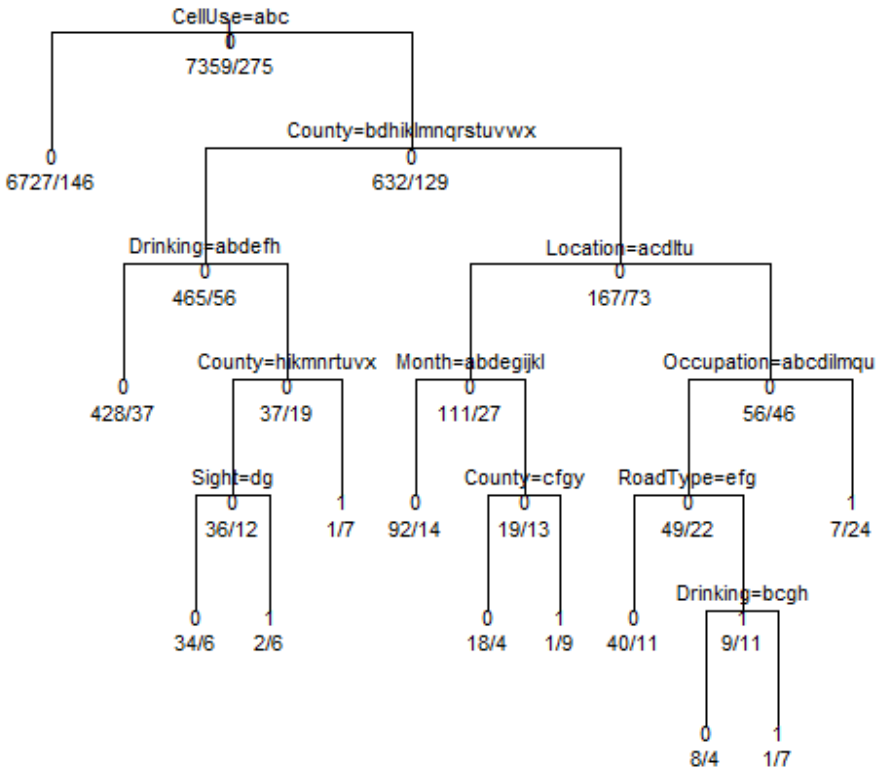

Figure 4 Classification tree and regression tree model using 2004 data

Based on 1,000 simulations where each simulation randomly draws 1,000 samples from the 2005 dataset with a fixed percentage of fatal accidents[5], Figure 5 compares the out-of-sample performance rankings (1 the best, 20 the worst) between the logistic regression model, CART model, and BRT models with various tree complexities. The performance is measured with the indicator of area under the receiver operating characteristic (ROC) curve (AUC). The result clearly indicates that the CART model is the worst model and has the worst performance most of the time. This result is no surprise because conventional CART models tend to focus on the major category when dealing with imbalance datasets (Chang and Wang 2006); ignoring the minor category produces lower AUC values because of the extremely low true-positive rate and high false-negative rate. The BRT model with tree complexity of one is similar to a logistic regression; therefore, their performances are similar.

Figure 5 shows that the out-of-sample predictive performance deteriorates when tree complexity increases. This result is different from the in-sample validation results that the predictive performance improves with a declining trend when tree complexity increases. The variance of predictive performance is large when tree complexity is low and high, but is small when tree complexity is around 7 to 11, as can be seen from the expansion and shrinkage of the boxes. This result may suggest that models with tree complexity below 7 underestimate the complexity of traffic accidents while those with tree complexity above 11 overestimate the complexity of traffic accidents. Generally, the BRT model with tree complexity of eight is preferred because it

[5] The percentage is at a fixed level of 3.98%, the percentage of fatal accidents for the whole 2005 data.

has satisfactory (small bias) and efficient (small variance) predictive performance. In other words, models with eighth-order interactions provide the best transferability.

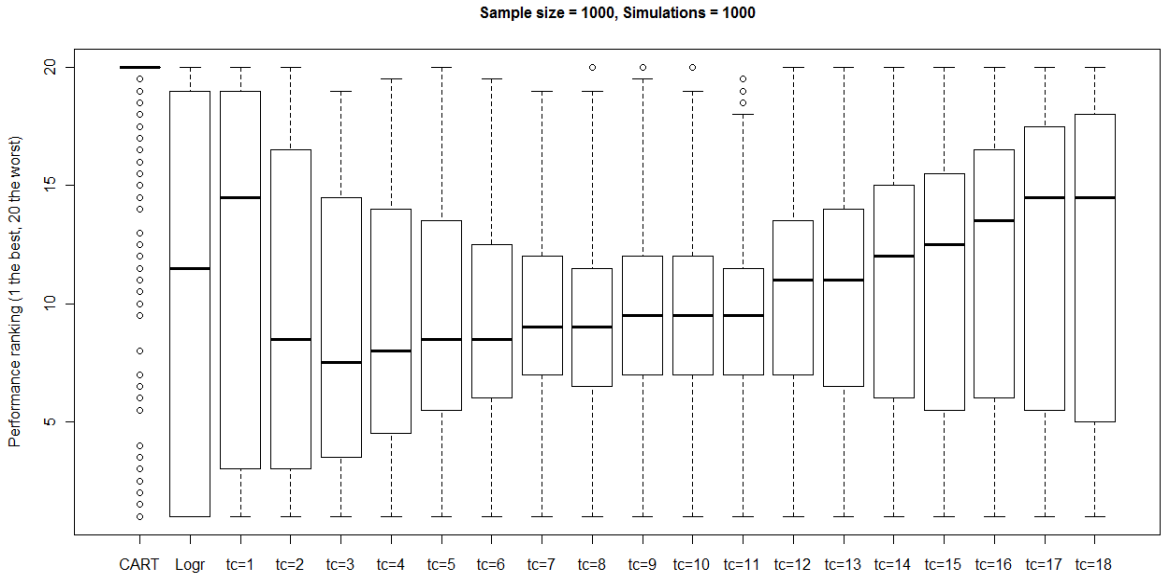

Figure 5 Out-of-sample performance rankings using 2005 data

DISCUSSIONS

This paper investigates the complexity of traffic accidents with a novel method, the boosted regression trees. An empirical dataset, 2004–2005 Taiwan single-motorcycle accidents, is adopted to demonstrate the method's usefulness. The analysis shows the ability of BRTs models to consider a great number of predictors, explore the nonlinear relationship between predictors and the response variable, and have satisfactory, if not better, predictive performance compared to logistic regression and classification and regression tree models.

The BRTs modeling results show that the models considering higher-order interactions exhibit better in-sample and out-of-sample predictive performance than the first-order models (i.e., the traditional logistic regression including only main effects, and the BRTs model with tree complexity of one). This result may suggest the existence of complicated accidents that are difficult to be approximated by models containing merely first- or low-order factors. For these accidents, the effect provided by some factors is conditioned on many other factors. In other words, the factor effects for complicated accidents are highly heterogeneous. On the other hand, accidents that are better approximated by high-order factor interactions account for only a relatively small portion of the total, as can be seen by the cross-validation result that the improvement of predictive performance decreases as the tree complexity increases. In other words, how factors affect the severity of most single-motorcycle accidents is not affected by (conditioned on) other factors. However, for a small portion that accident occurrence is complicated, the factor effects could change dramatically if the driving conditions alter. This result partially explains why good road safety countermeasures are effective to reduce most target accidents, but not all.

Despite the great ability to consider a large number of predictors, the empirical results show that the relative contributions concentrate on a few predictors. In our empirical demonstration, merely three predictors explain approximately half of the variation, indicating that a few predictors can determine the fatality of single-motorcycle accidents. Although the number of influential variables is small, the three most influential predictors, including county, occupation, and month, cannot provide straightforward explanations. The county, occupation, and month predictors respectively represent the geographical, personal, and seasonal factors that might affect the severity of single-motorcycle accidents. However, the exact nature of those factors is unknown. For example, geographical factors can relate to local driving culture or local road quality; both have been regarded as crucial factors to explain traffic accidents (Eiksund 2009, Rakauskas *et al.* 2009). Therefore, more studies are required to study the exact effects of such complex factors. One limitation of the relative contribution measure is the lack of confidence bounds; consequently, it would be difficult to tell the significance of relative contribution differences.

It should be noted that a highly-branching predictor, i.e. one that can be split into distinct classes, does not necessarily have a higher value of relative contribution because the BRTs models use the cross-validation technique to reduce overfitting. For example, as shown in Table 2, the variable cellphone use has only four categories and is ranked at the top-five influential variables; on the other hand, the variable movement prior to accident has 14 categories and is ranked among the least influential variables.

While the logistic regression, CART, and BRTs models discern similar influential (or significant) variables, how those variables are associated with accident severity is recognized differently. The CART model may be the most limited among these models, since only the importance ranking can be determined by observing the entrance order of the variables. Although tree-based models, including CART models, can fit nonlinear relationships, the conventional CART models do not provide quantified results, and thus it is difficult to interpret the relationship between the included predictors and the response variable. On the other hand, the logistic regression model provides the significance of the predictors, but it is difficult to specify the interaction terms in advance. It is recognized that econometric models, including logistic models, are not designed to explore a complicated structure; instead, they should be developed based on economic (or behavioral) theories. A parsimony logistic model should be valued more than a complicated logistic model if both models can capture crucial factor effects. However, modeling traffic accidents usually faces the challenges of numerous confounding factors (and categories) and nonlinear relationships. In particular, if variables such as age or drinking condition, which exhibit a nonlinear relationship with accident severity as shown in the empirical study, are not specified properly, the estimation results of logistic models may be erroneous. One possible way to resolve this problem is to combine the BRTs model and logistic regression model; i.e., using BRTs models to explore the relationship between the considered explanatory variables and the response variable, then transforming the variables, if necessary, to develop a representative model.

Among the predictors considered, the behavioral predictors including drinking condition, cellphone use, and protection equipment use are particularly important to explain unique single-

motorcycle accidents. This result is demonstrated by the increasing effect size of the behavioral interactions when the tree complexity increases. In other words, the harder cases are approximated more closely when the behavioral interactions are valued more in developing the BRTs model. The result may suggest that the relatively unique single-motorcycle accidents result from the combinations of some unexpected or undesired behaviors whose negative effect overrides the protective effect provided by the road design. To reduce such accidents requires safety education to improve drivers' unsafe behavior and attitudes.

Data quality is a typical issue in all the models. For example, the cellphone use variable is one of the most influential variables, and its most prominent category is "unknown". The reason for the prominence of the unknown category is simple: as long as the at-fault motorcyclists die at the crash site, police tend to record the cellphone use as unknown if no witness or further evidence is found. In other words, the significantly positive effect for the unknown category of cellphone use is a mixture of use and no use of cellphone. While this category seems useless to explain the relationship between cellphone use and the severity of single-motorcycle accidents, the BRT model shows that the relative contribution of the unknown category of cellphone use decreases as the tree complexity of BRT models increases. That is, while this unknown category approximates most fatal accidents well, it cannot explain the relatively unique fatal accidents.

CONCLUDING REMARKS

This paper applies the boosted regression tree method to investigate the factor complexity of single-motorcycle accidents. The advantages of BRTs models are demonstrated in the empirical study, including no need to pre-specify function form or to select variables or merge categories, and the abilities to consider numerous predictors and nonlinear relationship, provide satisfactory predictive performance, and offer quantitative results for interpretations. On the other hand, the disadvantages should also be noticed. Like other data mining methods, some parameters need to be tested for their best setting in developing BRTs models; the parameters include tree complexity, learning rate, and bagging fraction. Moreover, as tree complexity increases, the computation time also increases. A balance between computation cost and tree complexity and learning rate needs to be considered while developing boosted regression tree models.

Although providing interpretable statistics, the boosted regression tree is a data-driven approach. Therefore, boosted regression trees may be particularly useful to explore the unknown relationships between accident outcomes and affecting factors, especially complicated and nonlinear relationships. The explored relationships can be considered a basis or reference to develop behavioral theories.

ACKNOWLEDGEMENTS

The author would like to thank for the financial support by the National Science Council of Taiwan (NSC 98-2410-H-424 -018).

REFERENCES

Abdel-Aty, M., Haleem, K., 2011. Analyzing angle crashes at unsignalized intersections using machine learning techniques. Accident Analysis and Prevention 43, 461-470.

340 Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on
341 accident severity. Accident Analysis and Prevention 34 (6), 729-741.
342 Bedard, M., Guyatt, G.H., Stones, M.J., Hirdes, J.P., 2002. The independent contribution of
343 driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis and
344 Prevention 34 (6), 717-727.
345 Bina, M., Graziano, F., Bonino, S., 2006. Risky driving and lifestyles in adolescence. Accident
346 Analysis and Prevention 38 (3), 472-481.
347 Chang, H.L., Yeh, T.H., 2007. Motorcyclist accident involvement by age, gender, and risky
348 behaviors in taipei, taiwan. Transportation Research Part F-Traffic Psychology and
349 Behaviour 10 (2), 109-122.
350 Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: An application of non-
351 parametric classification tree techniques. Accident Analysis and Prevention 38 (5), 1019-
352 1027.
353 De'ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology 88 (1), 243-251.
354 Eiksund, S., 2009. A geographical perspective on driving attitudes and behaviour among young
355 adults in urban and rural norway. Safety Science 47 (4), 529-536.
356 Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. Journal
357 of Animal Ecology 77 (4), 802-813.
358 Freund, Y., Schapire, R.E., Year. Experiments with a new boosting algorithm. In: Proceedings of
359 the Machine Learning: Proceedings of the Thirteenth International Conference.
360 Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Annals of
361 Statistics 29 (5), 1189-1232.
362 Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification
363 trees and random forests. Accident Analysis and Prevention 41, 98-107.
364 Hastie, T., Tibshirani, R.J., Friedman, J.H., 2009. The elements of statistical learning, 2nd ed.
365 Springer-Verlag, New York, New York, USA.
366 Hauer, E., 2006. Cause and effect in observational cross-section studies on road safety. Presented
367 at the 85th annual meeting of the Transportation Research Board. Washington, D.C.,
368 U.S.A.
369 Hotz, G.A., Cohn, S.M., Popkin, C., Ekeh, P., Duncan, R., Johnson, W., Pernas, F., Selem, J.,
370 2002. The impact of a repealed motorcycle helmet law in miami-dade county. Journal of
371 Trauma-Injury Infection and Critical Care 52 (3), 469-473.
372 Hundley, J.C., Kilgo, P.D., Miller, P.R., Chang, M.C., Hensberry, R.A., Meredith, J.W., Hoth,
373 J.J., 2004. Non-helmeted motorcyclists: A burden to society? - a study using the national
374 trauma data bank. Journal of Trauma-Injury Infection and Critical Care 57 (5), 944-949.
375 Kockelman, K.M., Kweon, Y.J., 2002. Driver injury severity: An application of ordered probit
376 models. Accident Analysis and Prevention 34 (3), 313-321.
377 Li, M.D., Doong, J.L., Chang, K.K., Lu, T.H., Jeng, M.C., 2008a. Differences in urban and rural
378 accident characteristics and medical service utilization for traffic fatalities in less-
379 motorized societies. Journal of Safety Research 39 (6), 623-630.
380 Li, X.G., Lord, D., Zhang, Y.L., Me, Y.C., 2008b. Predicting motor vehicle crashes using
381 support vector machine models. Accident Analysis and Prevention 40 (4), 1611-1618.
382 Lin, M.R., Chang, S.H., Huang, W.Z., Hwang, H.F., Pai, L., 2003a. Factors associated with
383 severity of motorcycle injuries among young adult riders. Annals of Emergency Medicine
384 41 (6), 783-791.

385 Lin, M.R., Chang, S.H., Pai, L., Keyl, P.M., 2003b. A longitudinal study of risk factors for
386 motorcycle crashes among junior college students in taiwan. Accident Analysis and
387 Prevention 35 (2), 243-252.
388 Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed
389 logit model: An exploratory empirical analysis. Accident Analysis and Prevention 40 (1),
390 260-266.
391 Mtc, 2007. Monthly statistics of transportation and communications. Taipei, taiwan: Ministry of
392 transportation and communications, executive yuan, taiwan.
393 Nhtsa, 2009. 2007 motorcycles traffic safety fact sheet. National Highway Traffic Safety
394 Administration.
395 Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction
396 for two-lane highway segments. Accident Analysis and Prevention 36 (2), 183-191.
397 R Development Core Team, 2009. R: A language and environment for statistical computing. R
398 Foundation for Statistical Computing, Vienna, Australia.
399 Rakauskas, M.E., Ward, N.J., Gerberich, S.G., 2009. Identification of differences between rural
400 and urban safety cultures. Accident Analysis and Prevention 41 (5), 931-937.
401 Ridgeway, G., 2007. Generalized boosted models: A guide to the gbm package.
402 Rutter, D.R., Quine, L., 1996. Age and experience in motorcycling safety. Accident Analysis and
403 Prevention 28 (1), 15-21.
404 Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S., Barbone, F., 2002. Risk factors
405 for fatal road traffic accidents in udine, italy. Accident Analysis and Prevention 34 (1),
406 71-84.
407 Wong, J.T., Chung, Y.S., 2007. Rough set approach for accident chains exploration. Accident
408 Analysis and Prevention 39 (3), 629-637.
409 Wong, J.T., Chung, Y.S., 2008a. Analyzing heterogeneous accident data from the perspective of
410 accident occurrence. Accident Analysis and Prevention 40 (1), 357-367.
411 Wong, J.T., Chung, Y.S., 2008b. A rule comparison approach for identifying causal factors of
412 accident severity. Transportation Research Record: Journal of the Transportation
413 Research Board 2083, 190-198.
414 Yagil, D., 1998. Instrumental and normative motives for compliance with traffic laws among
415 young and older drivers. Accident Analysis and Prevention 30 (4), 417-424.
416
417