Accident Analysis and Prevention 61 (2013) 107–118

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees

Yi-Shih Chung*

Department of Logistics and Shipping Management, Kainan University, Taiwan

ARTICLE INFO

Article history:
Received 28 November 2011
Received in revised form 24 April 2012
Accepted 16 August 2012

Keywords:
Boosted regression trees
Crash classification
Motorcycle crashes
Machine learning

ABSTRACT

Factor complexity is a characteristic of traffic crashes. This paper proposes a novel method, namely boosted regression trees (BRT), to investigate the complex and nonlinear relationships in high-variance traffic crash data. The Taiwanese 2004–2005 single-vehicle motorcycle crash data are used to demonstrate the utility of BRT. Traditional logistic regression and classification and regression tree (CART) models are also used to compare their estimation results and external validities. Both the in-sample cross-validation and out-of-sample validation results show that an increase in tree complexity provides improved, although declining, classification performance, indicating a limited factor complexity of single-vehicle motorcycle crashes. The effects of crucial variables including geographical, time, and sociodemographic factors explain some fatal crashes. Relatively unique fatal crashes are better approximated by interactive terms, especially combinations of behavioral factors. BRT models generally provide improved transferability than conventional logistic regression and CART models. This study also discusses the implications of the results for devising safety policies.

1. Introduction

Complexity is a typical feature of crash occurrences and challenges the modeling of crash outcomes. Traffic crashes involve various factors. Many studies have addressed the importance of controlling for confounding factors when modeling traffic crashes, especially in cross-sectional studies where causes are unknown a priori (Elvik, 2002; Hauer, 2006). The relationship between the response variable and the explanatory variables may be nonlinear, which increases complexity. For example, the relationship between crash severity and driver age is nonlinear. Young or old drivers are more likely to be involved in a fatal crash than middle-aged drivers, typically because young drivers drive fast and old drivers have more fragile bodies (Rutter and Quine, 1996; Lin et al., 2003a; Chang and Yeh, 2007). Interactions between explanatory variables are also complex. Accident chain theory, which sees a traffic crash as a series of undesired activities or events instead of as a single factor, supports this view of crash occurrences as complex (Reason, 1997; Fleury and Brenac, 2001; Elvik, 2003).

Safety studies have used several descriptive data mining methodologies (Bayam et al., 2005) to unravel the factor complexities of crashes and identify patterns in traffic crashes. Rule-based methods explore a set of rules that describe the interdependence between factors and crash outcomes (e.g., Pande and Abdel-Aty, 2009; Mirabadi and Sharifian, 2010; Saunier et al., 2011; Zhang et al., 2011). Seeing the advantage of handling numerous factors, Montella (2011) used association rule analysis to identify the patterns of factors contributing to roundabout crashes in Italy. Geurts et al. (2005) used the frequent item sets technique to explore and compare crash patterns inside and outside black zones. In a series of studies, Wong and Chung (2007, 2008a,b) employed rough sets theory to explore the circumstances that distinguish crash severity.

Tree-based methods are also a frequent choice. Classification trees, also called decision trees, classify observations by recursively partitioning the predictor space such that the data set is segmented into smaller and more homogeneous groups (Breiman et al., 1984). Chang and Wang (2006) examined injury severity in traffic crashes in Taiwan using classification and regression tree (CART) models. Elmitiny et al. (2010) also applied the CART method to identify the behavioral patterns of driver stop-and-go decisions and red-light running at an intersection. To analyze influential factors in pre-crash maneuvers, Harb et al. (2009) used classification trees to explore the association of potential factors with rear-end, head-on, and angled crashes.

Clustering methods are used when there is no class to be predicted, but rather when observations are divided into natural groups (Witten et al., 2011). Golob and Recker (2004) used *k*-means clustering to identify homogeneous groups of traffic flow regimes under three conditions: dry roads during daylight, dry roads at night, and wet roads. Depaire et al. (2008) also employed

* Tel.: +886 3 341 2500x6083; fax: +886 3 341 2361.
E-mail addresses: yishih.chung@gmail.com, zest@mail.knu.edu.tw

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2012.08.015

a clustering technique – latent class clustering – to classify distinct accident types. Mussone and Kim (2010) applied a self-organizing map method to identify crash clusters by reducing crash data from an N-dimensional space to a two-dimensional plane. All of these descriptive data mining methods can manage large data sets and identify interesting patterns that provide useful information for further statistical analyses.

The studies described demonstrate the utility of various data mining methods, but this study proposes a novel method, namely boosted regression trees (BRT), to investigate crash occurrence complexities for three reasons. First, although conventional data mining methods (especially those inherited from the machine-learning field) can manage large data sets, they normally lack summary measures to indicate correlations between potential factors and crash outcomes. Researchers are required to use other techniques to provide this information. This issue is particularly relevant when the number of rules identified or the resulting classification tree is large. For example, Harb et al. (2009) used the random forest technique, following an analysis of classification trees, to examine the importance of potential factors. Depaire et al. (2008) adopted a multinomial logit model, after latent class clustering analysis, to compare the contribution of predictors between crash types. Kim et al. (2008) employed the logistic regression method to examine the association between hit-and-run collisions and key factors identified by rough set analysis. Second, the classification performance is rarely examined for these descriptive data mining methods. Although their purpose is not prediction, it is preferable that descriptive data mining methods have a level of classification performance to enhance the validity of results, and thus, the effectiveness of the suggested safety countermeasures. Some studies, such as that by Wong and Chung (2008b), used cross-validation (CV) to examine the internal validity of discovered patterns. Montella et al. (2011) applied Bonferroni tests to determine the significance of the explored crash patterns (i.e., trees and rules). Pruning is another common technique to generate good rules or trees in data mining applications (Witten et al., 2011). However, the external validity (i.e., the out-of-sample classification performance) of most studies that use descriptive data mining methods is rarely examined. In other words, the effectiveness of the proposed safety countermeasures (i.e., those used to prevent future crashes or crashes in other areas) might be questionable. Third, most of the data mining methods adopted do not contain mechanisms to handle skewed or imbalanced data sets, that is, data sets where the minority class size is small. This issue is common in crash severity studies where fewer fatal crashes exist than injury-only or property damage-only crashes. This imbalance causes suboptimal classification performance in data mining applications. For example, although they achieved satisfactory overall classification performance, Chang and Wang (2006) showed an extremely low hit rate for fatal crashes using the CART technique. Thus, this study proposes the BRT method.

The BRT method is a tree-based technique; therefore, it retains the advantages of tree models, including that it does not require the pre-specification of function forms and the ability to consider numerous explanatory variables and their possible nonlinear relationships with the response variable (Chang and Wang, 2006). Even more important is that the BRT method incorporates the shrinkage technique, which simultaneously reduces the variance and bias of classification errors. This advantage is crucial for crash modeling because traffic crash classification trees are usually large because of numerous predictors and the uniqueness of traffic crashes (De'ath, 2007). The most relevant advantage of the method is its boosting design. This means that as trees grow sequentially, the method gradually focuses on difficult cases (i.e., relatively unique traffic crashes). This mechanism enables the BRT method to better manage unbalanced crash data sets and capture the features of unique traffic crashes. The boosting design also provides interpretable summary measures. These measures can scrutinize crash complexities. The following section describes the BRT method.

This study uses a single-vehicle motorcycle[1] crash empirical data set to investigate the complexity of crash occurrence where injury levels (fatal vs. non-fatal) are examined. Single-vehicle motorcycle crashes are those that involve only one vehicle (a motorcycle). Although single-vehicle motorcycle crashes account for a relatively small portion of traffic crashes, they are usually serious. Moreover, single-vehicle crashes should be simpler to study than multi-vehicle crashes. This is appropriate for this preliminary study because it investigates the factor complexity of crash occurrence.

The remaining parts of this paper are organized as follows: Section 2 describes the methodology, including the BRT method, the empirical data set and variables, and the analysis procedure. Section 3 provides the analysis results. This is followed by a discussion of this analysis in Section 4. Section 5 presents concluding remarks.

2. Methodology

2.1. Boosted regression trees

2.1.1. Boosted regression trees theory

Two terms characterize BRTs: regression trees and boosting. A BRT model grows a number of trees by bootstrapping the training data, that is, by randomly selecting a certain proportion of observations from the training data without replacement. Each tree grows like a CART develops, a form of binary recursive partitioning. The term "binary" implies that each group of traffic crashes, 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" means that the binary partitioning process can occur repeatedly. The term "partitioning" indicates that the data set is split into sections or partitioned. Splitting functions, which measure the purity (or impurity) of a tree, determine the variable that should be included to split the tree; common functions include Gini, Twoing, and Entropy. To prevent overfitting of data, trees are typically pruned to sever nodes (or branches) resulting in high classification costs (Chang and Wang, 2006). A cost function of misclassification is usually defined to determine which node to prune. CV or out-of-sample validation selects the best tree.

Despite the advantages of CART models, a single tree is occasionally a weak classifier, especially with high-variance data such as traffic crash data. Single-tree models normally use few variables to prevent data overfitting. This makes them an unstable method where a small data change may cause a large change in a tree (Zhang et al., 2005). If several variables are included to control the effects of confounding factors in a single-tree model, the model usually results in high variance and low bias – the bias-variance trade-off (De'ath, 2007). To balance the bias and variance, the bagging technique is introduced. Bagging involves the following steps: (1) take a bootstrap sample from the training data set; (2) fit the tree to this bootstrapped data set; (3) repeat the first 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.

In addition to bagging, the BRT method applies a special mechanism to bootstrap samples, namely boosting. Boosting uses the same principle as bagging, where a weak algorithm is run repeatedly. The computed classifiers are then combined in the final estimation or prediction. In other words, boosting, similarly to bagging, can effectively reduce the variance. However, whereas

[1] Motorcycles in this study include motorcycles (engine capacity greater than 50 cc) and mopeds (engine capacity less than 50 cc).

conventional bagging focuses on randomly selecting observations from the original data with replacement, boosting also considers the difficulty of the training cases. When repeatedly selecting sub-data sets, boosting tends to generate distributions that concentrate on more difficult training cases (Freund and Schapire, 1996). This feature is crucial in safety studies because fatal crashes typically account for a small portion of all crashes.

The algorithm for developing BRT models is as follows: suppose we want to build a function $f(x)$ to approximate 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). This 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 obtain the estimate of γ_m of $\beta b(x; \gamma)$.
   c. Obtain 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 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 BRT method applies a shrinkage strategy. Learning rate ϵ is introduced in Step 2d when the algorithm updates the estimated function:

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

A lower learning rate requires more iterations (i.e., trees) in the boosting sequence. Studies have indicated that a 10-fold reduction in the learning rate requires an approximate 10-fold increase in iterations (De'ath, 2007) and have recommended at least 1000 trees (Elith et al., 2008).

2.1.2. Parameters and summary measures to understand crash complexity

Three parameters require calibration in a BRT model: the number of trees (nt), the learning rate (lr), and the tree complexity (tc). The first two parameters allow analysts to control the boosting speed, such that tree growth stops at a reasonable tree number (usually 1000–10,000).[2] This study focuses on the last parameter – tree complexity. Tree complexity controls whether interactions are fitted; for example, a tc value of 1 fits an additive model (i.e., single decision stumps with two terminal nodes at each tree), and a tc value of 2 allows up to two-way interactions at each tree, and so on (Elith et al., 2008).

In addition to tree complexity, this study presents three summary measures to demonstrate factor complexities of crash occurrence, including relative contribution, marginal effects of predictors, and the magnitude of interaction effects. 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. Relative contribution is scaled, such that the overall sum is 100; a higher relative contribution of an explanatory variable indicates a stronger influence on the crash outcome. Marginal effects demonstrate the effect of an explanatory variable on the crash outcome after accounting for the average effects of all other variables. The magnitude of interaction effects evaluates the amount of residual variance a pair of explanatory variables can explain (Elith et al., 2008).

2.2. Data and variables

The empirical data used to investigate factor complexities are two years (2004–2005) of single-vehicle motorcycle crashes, provided by the National Police Agency of Taiwan. The number of single-vehicle motorcycle crashes was 7634 in 2004 and 9869 in 2005, with fatal crash rates of 3.52% and 3.98%, respectively. The extremely low fatal crash rates reflect the imbalance in the data set.

Table 1 shows that the data set contains 29 variables. The dependent variable is crash severity, coded as a binary variable with a value of 1 if fatal and 0 if not. The remaining 28 variables include driver characteristics, trip characteristics, driving behavior, weather conditions, and road conditions. All variables are categorical variables except driver age, speed limit, and time of the day.

2.3. Analysis procedure

To investigate the complexity of single-vehicle motorcycle crashes, a series of BRT models were developed and examined. A CV technique was used to ensure that the discovered factor complexities (the tree complexity parameter, relative contribution of explanatory variables, marginal effects of explanatory variables, and the magnitude of interaction effects) best described but did not overfit the training data (the 2004 single-vehicle motorcycle crashes). Out-of-sample tests subsequently scrutinized the external validity of the developed BRT models. The 2005 single-vehicle motorcycle crash data were used as the testing data. This study shows that BRT models are more stable and better at out-of-sample classification than logistic regression and CART models. The best CV and out-of-sample-test models were also compared to examine the similarities and differences between the 2004 and 2005 crash patterns.

This study mainly developed the BRT models according to suggestions by Elith et al. (2008). The BRT models were built using the software *R* (R Development Core Team, 2009) with the *gbm* package. Three parameters were jointly calibrated to optimize the BRT models: the number of trees, learning rate, and tree complexity.

[2] The rule of thumb suggested by Elith et al. (2008) is fitting models with at least 1000 trees. The analysis results, as presented in the next sections, show that all models converge within 10,000 trees.

Table 1
Variables to develop single-vehicle motorcycle crash 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 the week	Monday, Tuesday, . . ., Sunday (7 days)	Categorical
	Time of the day	0–23	Continuous
	County	Taipei city, Taipei county, etc. (25 counties)	Categorical
Driving behavior	Protective equipment use	Wear (helmet), Not wear, Others (3 categories)	Categorical
	Cell phone use	No use, Handheld, Earphone, Hands free, Others (4 types)	Categorical
	Movement prior to crash	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

This study did not control the number of trees if it was a reasonable size, between 1000 and 10,000. The combinations of varying values of learning rates (0.05–0.0001) and tree complexity levels (1–18) were also tested to develop the best BRT models. To reduce overfitting and improve accuracy, trees were boosted based on random draws from the full training data set. Fifty percent of the data were drawn at random without replacement at each iteration from the full 2004 data set.

The CV technique was used to prevent overfitting and determine the best BRT model setting for the 2004 single-vehicle motorcycle crashes when various combinations of learning rates and tree complexity levels were examined. Tenfold CV was chosen and predictive deviance was applied to measure the success of the models. Because the dependent variable is binary, the Bernoulli loss function[3] was chosen as the deviance for the binary response variable. Thus, the BRT models become a form of logistic regression that models the probability that a fatal traffic crash will occur, $y = 1$, with explanatory variables $\boldsymbol{X}$, $P(y = 1|\boldsymbol{X}) = f(\boldsymbol{X})$. The AdaBoost exponential loss function can also be used for binary outcomes. This would make the model more akin to the classification tree model. However, studies have shown that the Bernoulli loss function is more likely to yield robust results than the exponential loss function, and was therefore chosen for this study (Ridgeway, 2007; Elith et al., 2008; Hastie et al., 2009). For more information on logistic regression tree analysis, refer to Chan and Loh (2004), Loh (2006), and Hastie et al. (2009). All 28 explanatory variables listed in Table 1 were used to develop the BRT models. Variables were implicitly selected by down-weighting variable contributions at each iteration (Elith et al., 2008), a process called shrinkage in the data-mining field.

The logistic regression and CART models were chosen to conduct the out-of-sample tests because past studies have examined these models extensively. This is particularly true of the logistic regression model (Al-Ghamdi, 2002; Bedard et al., 2002; Valent et al., 2002). Other advanced econometric models such as mixed logit models could have been used for comparison; however, these advanced models usually require delicate model specifications and more function form and parameter assumptions. Comparing the out-of-sample classification performance with logistic regression and CART models provides a basic understanding about the external validity of the BRT models.

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

The CART models were developed using the cost-complexity pruning strategy. This stops the tree-growing process only when a node size that minimizes cross-validated errors is reached (Hastie et al., 2009). The Gini function was chosen as the splitting function.

3. Results

3.1. BRT model for the 2004 single-vehicle motorcycle crash data

Fig. 1 shows the calibration results. Only representative results are presented because of limited space. The gaps between the flat line segments are small, indicating that the variation of learning rates does not exhibit much difference on the converged predictive

[3] The Bernoulli deviance: $-2\left(1/\sum w_i\right)\sum wi(y_i f(X_i) - \log(1 + \exp(f(X_i))))$, where y_i is the response, $\mathbf{X}_i$ is the vector of explanatory variables, and ω_i are the observation weights (Ridgeway, 2007). Equal weights are used in this study.

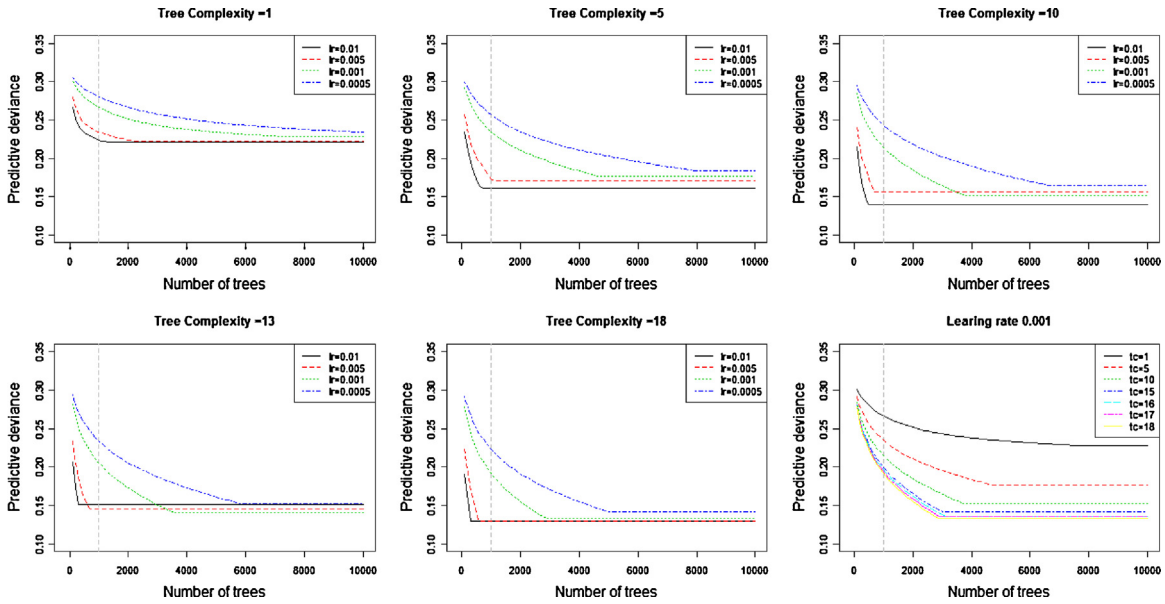

Fig. 1. Predictive deviance against number of trees for models fitted with various tree complexities and learning rates. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

deviation. This result reflects the primary role of learning rates, which is maintaining consistent tree-growing speed instead of improving predictive performance. Fig. 1 shows that a learning rate of 0.001 (the green line) generally produces a BRT model with a reasonable tree size and the lowest predictive deviance.

Increasing tree complexity significantly improves predictive deviation, but at a decreasing rate. The bottom right panel of Fig. 1 shows that with a fixed learning rate of 0.001, the predictive deviance continuously decreases as the tree complexity increases. Predictive deviation continues to decrease until the tree complexity approaches 18. The overlapping pink ($tc = 17$) and yellow lines ($tc = 18$) in Fig. 1 show this. This result indicates that high-order interaction effects are in the 2004 single-vehicle motorcycle crash data, and that *tc*-18 (i.e., trees that allow up to 18-way interactions) is the BRT model with the lowest predictive deviation using 10-fold CV.

3.2. Relative contributions of explanatory variables

To explore the factor effects contributing to the 2004 single-vehicle motorcycle crashes, this study examined the relative contributions of explanatory variables to the BRT models with a fixed learning rate of 0.001 and various tree complexity levels. Fig. 2 plots the relative contributions that evolve with various complexity levels. The relative contribution values for the BRT models with tree complexities of 1 (percentages on the left) and 18 (percentages on the right) are provided for comparison.

Fig. 2 shows that the explanatory variables contribute differently to the *tc*-1 and *tc*-18 BRT models. For example, drinking condition is the most influential factor with a relative contribution of 31.58% to model *tc*-1, but its relative effect declines to 8.54% when tree complexity increases to 18. The variable, month, contributes 0.19% to the *tc*-1 model and it contributes 14.11% to the *tc*-18 model. In summary, 5 of 28 predictors contribute less as tree complexity level increases. These five predictors are mainly driving behavior variables, including drinking condition, cell phone use, protective equipment use (i.e., helmet wearing), and movement prior to a crash. The remaining 23 predictors generally contribute more as the tree complexity level increases. A few predictors account for most of the contribution to the *tc*-1 model. Because most of the predictors contribute more as the level of tree complexity increases, the distribution of relative contribution is more even in the *tc*-18 model. Fig. 2 shows that the change in relative contributions is larger at low tree complexity levels and stabilizes when the tree complexity level exceeds 10.

County, occupation, month, drinking condition, cell phone use, road location, day of the week, protective equipment use, median type, and age are the variables that contribute the most to explaining the dependent variable. Collectively, they account for more than 85% of the total effect of the *tc*-18 model. Irrespective of tree complexity, county variable is the most influential variable, implying that geographically heterogeneous factors such as local driving culture, the relatively long distance from hospitals in rural areas, or road design elements (Eiksund, 2009) are crucial in identifying the severity of single-vehicle motorcycle crashes. The occupation and month variables are 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 between the severity of single-vehicle motorcycle crashes.

Past studies have provided extensive discussions on drinking condition and cell phone use. Driving under the influence of alcohol or using a cell phone while driving increase the possibility of traffic crashes and their severity. Road location is the sixth most influential variable. Wong and Chung (2007, 2008b) showed that some road locations such as intersections, where more fixed objects exist, increase the possibility of bump-into-fixed-object traffic crashes and their severity. Protective equipment use and day of the week are the next two most influential variables. Both longitudinal and cross-sectional studies show that wearing a helmet reduces the severity of motorcycle crashes by protecting the brain (e.g., Hotz et al., 2002; Hundley et al., 2004). The median-type variable is ranked ninth and contributes 2–3%. Road levels affect the selection and construction of median types; a wide median island creates fixed objects on the road and increases the possibility of traffic crashes and their severity (Wong and Chung, 2007, 2008b). The age variable is the only continuous variable of the 10 most influential variables. This suggests that the characteristics of motorcyclist age groups affect the fatality of traffic crashes. For example, young

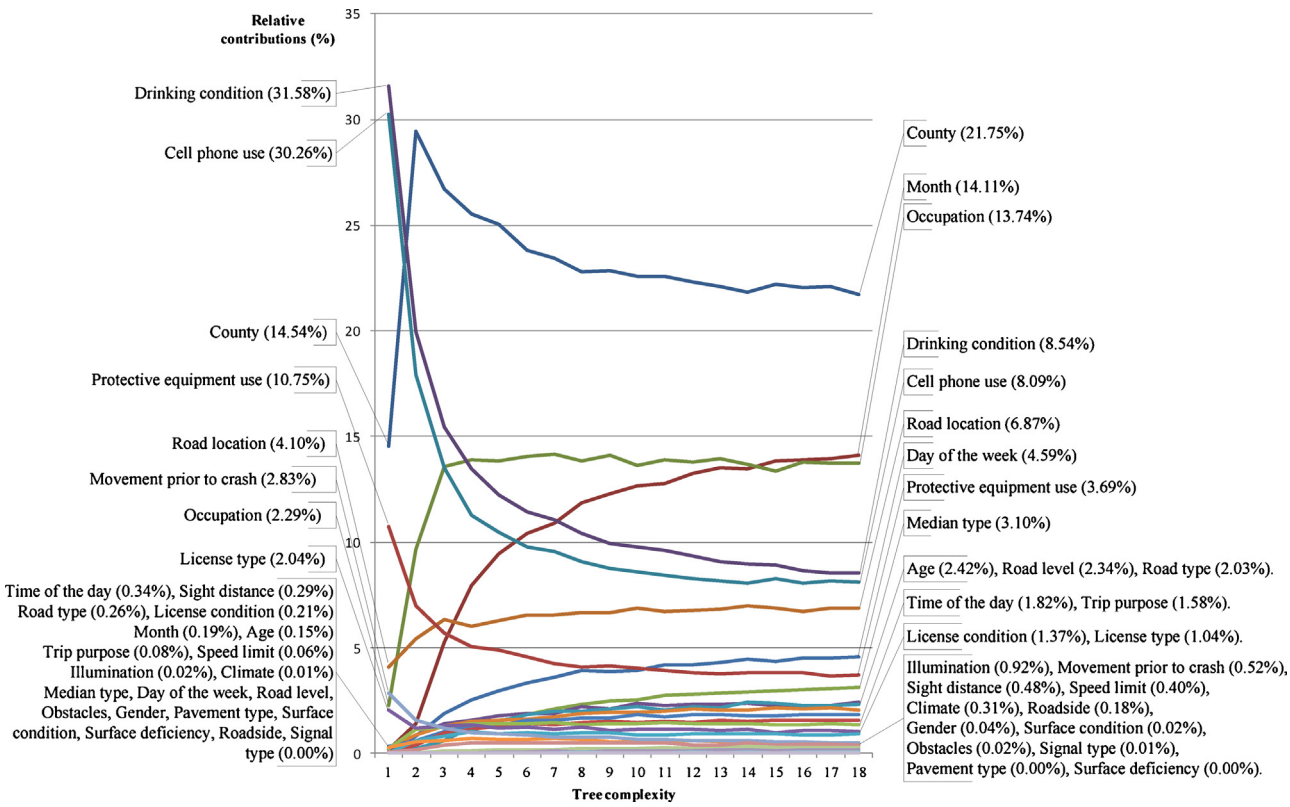

Fig. 2. Relative contributions (%) of explanatory variables for BRT models with a fixed learning rate of 0.001 and various tree complexity levels.

drivers are more antisocial, resulting in reckless driving behaviors and fatal crashes. The declining physical condition of old drivers makes them more fragile in traffic crashes.

Fig. 2 shows that nearly all of the 10 most influential variables in the *tc*-18 model are variables related to driver characteristics, trip characteristics, and driving behaviors. The weather-condition variable and most road environment variables contribute minimally to the fatality of single-vehicle motorcycle crashes.

3.3. Marginal effects of explanatory variables

To further investigate how the explanatory variables affect crash fatality, the partial-dependence that shows the effect of a variable on fatality after controlling for the average effects of all other variables in the model is examined. Results indicate that the partial dependence of explanatory variables stabilizes quickly. If the tree complexity level exceeds 3, the patterns of how the explanatory variables relate to crash fatality rarely change. In other words, the *tc* parameter does not significantly affect the partial dependence of explanatory variables.

The partial dependence shows interesting patterns between the explanatory variables and crash fatality. Fig. 3 shows a summary of the partial dependence of the 12 most influential predictors in the BRT model with a learning rate of 0.001 and a tree complexity of 18.

A nonlinear relationship exists in three variables: age, drinking condition, and road levels. The age variable demonstrates a convex marginal effect on the probability of fatal single-vehicle motorcycle crashes, as shown in Fig. 3(j). Older motorcyclists are more likely to be involved in fatal crashes than other age groups, especially when motorcyclists are older than 60. An age of less than 20 also has a marginal effect on the probability of being involved in a fatal single-vehicle motorcycle crash. An age of approximately 40 years has the lowest marginal effect on the probability of being involved in a fatal single-vehicle motorcycle crash. Motorcyclists at this age are expected to have accumulated a certain level of driving experience; they are also physically and psychologically mature (compared to younger drivers), with healthy bodies (compared to older drivers) (Yagil, 1998). Consequently, this age group has a lower fatality rate.

Fig. 3(d) shows the marginal effect of drinking condition on the probability of fatal single-vehicle motorcycle crashes. There is a nonlinear relationship between driving under the influence of alcohol and the fatality rate. Motorcyclists who are heavily drunk demonstrate the largest effect on the fatality rate, followed by slightly drunk motorcyclists. Blood alcohol content in the middle range (between 0.26 and 0.55 μg/L) has a lower marginal effect. Motorcyclists who are heavily drunk cannot maneuver the motorcycle nor protect themselves if a crash occurs. Consequently, they are associated with a high fatality rate. Slightly drunk motorcyclists may ignore their diminished capacity, thus leading to a higher fatality rate.

Fig. 3(k) demonstrates a nonlinear relationship between road levels and their marginal effect on the fatality rate. High and low-level roads have significant marginal effects, but middle-level roads have a smaller marginal effect. High-level roads, such as national and provincial highways, have high speed limits and crashes are more severe because of driving speed. Low-level roads cannot provide sufficient protection for motorcyclists if a crash occurs, and therefore, more fatal crashes occur on them.

Fig. 3(a) shows how each county relates to crash fatality. The most influential counties 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 counties are generally classified in the third level of administrative bureaucracy in Taiwan, and typically have smaller public

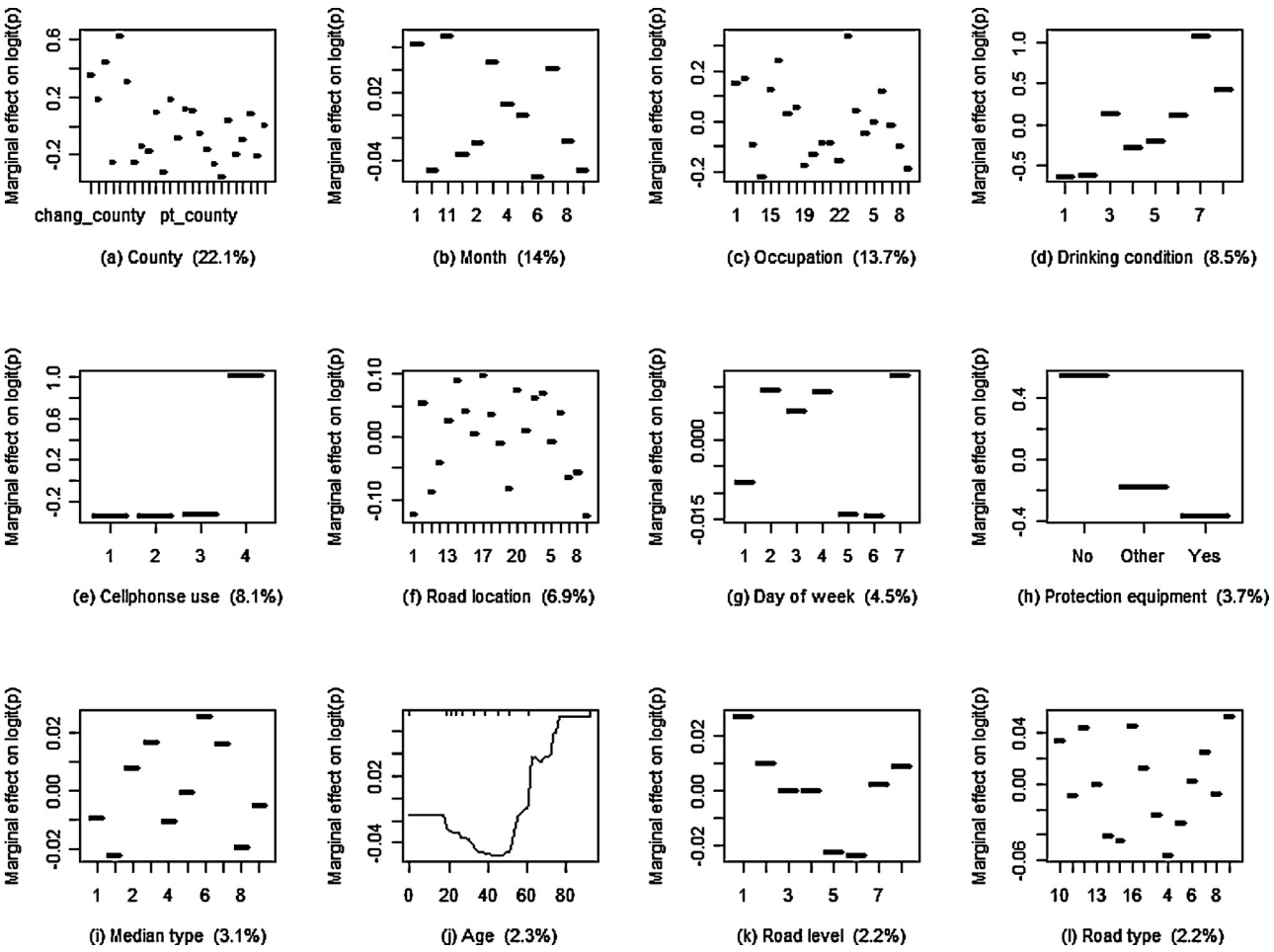

Fig. 3. Partial-dependence plots for the 12 most influential variables.

construction budgets. The poor road quality provides less protection for motorcyclists when a crash occurs, leading to a higher fatality rate.

Fig. 3(b) and (g) shows the effects of seasonal factors. Fig. 3(b) shows that January, March, July, and November are associated with higher fatality rates. Lunar New Year, spring vacation, and summer vacation occur in these months, respectively. Fig. 3(g) shows a higher fatality rate on typical working days, Tuesday, Wednesday, and Thursday, as well as Sunday.

Fig. 3(c) shows how occupation relates to fatality rates. Motorcyclists who are high school students, bus or railroad occupational drivers, or police officers are associated with higher fatality rates. High school students are mostly under 18 years of age and drive motorcycles illegally. Student lifestyles are also typically different from others of a similar age, which might lead to a higher possibility of crash occurrence and fatality rates (Lin et al., 2003a). Some police officers and occupational drivers work night shifts in Taiwan; these workers are more likely to have sleep problems and a higher level of work pressure. This results in an increased possibility of traffic crashes and more severe injury levels.

Fig. 3(e) shows the marginal effect of cell phone use on the fatality rate. It shows that the unknown category exhibits the highest marginal effect. This result reflects the difficulty of reporting cell phone use in traffic crashes. Fig. 3(h) shows that not wearing helmets is correlated with an extremely high effect on the fatality rate, consistent with past studies (Li et al., 2008).

Fig. 3(f), (i), and (l) shows the marginal effects of the three road-environment variables. Fig. 3(f) shows that road locations with higher crash fatality rates include exclusive bus lanes, nearby ramps, and motorcycle waiting zones. Exclusive bus lanes are usually designed for areas with high population densities and demand for public transport. A road segment equipped with exclusive bus lanes is typically wider with a higher speed limit. The road geometric design is also more complex than in other roads. In other words, the road environment encourages a high driving speed and requires motorcyclists to focus on the complex design, resulting in fatal crashes. A similar reasoning explains the significant effect of the vicinity of ramps and motorcycle waiting zones. The roads approaching highway ramps are typically wide with high speed limits for vehicles to enter highways.[4] The motorcycle waiting zones are designed for motorcyclists to turn left at wide intersections (i.e., where two or more lanes are in one direction). The speed limit is usually high and the road geometry is more complex, compared to narrow intersections.

The median types associated with high fatality rates are narrow median islands (smaller than 50 cm) and markings that prohibit overtaking, as shown in Fig. 3(i). Although the installation of median islands can prevent conflicts between vehicles driving in opposite directions, it also creates fixed objects on the road and increases the probability of crashes and crash severity. Markings that prohibit overtaking are typically drawn on road segments approaching intersections or without sufficient sight distance. Fatal single-vehicle motorcycle crashes at these locations suggest high speeds,

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

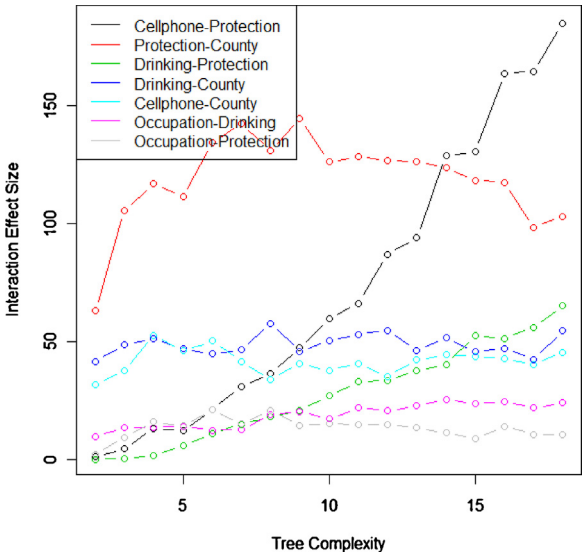

Fig. 4. Top seven interaction effects for BRT models with various tree complexity levels. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

which are inappropriate for these locations. Fig. 3(l) shows that a higher fatality rate is associated with road types that require more sophisticated driving skills, including roundabouts, culverts, elevated roads, and graded roads.

3.4. Important interactions

Fig. 4 shows pairwise interactions with effect sizes greater than 10 for models with various tree complexity levels. Although the 28 explanatory variables generate 378 combinations of variable pairs, few are crucial in explaining the variance in fatality. Regardless of tree complexity, the analysis results show that seven variable pairs contribute an effect size greater than 10. These same variable pairs play the most critical roles across all BRT models.

Among the seven variable pairs, the two pairs that combine behavioral variables demonstrate an explicitly upward trend as tree complexity increases. The first pair is the combination of cell phone use and protective equipment use. As shown in Fig. 4, the effect size of the cell phone-protective equipment pair (black line) is extremely prominent when tree complexity increases. The other pair is the combination of drinking condition and protective equipment use. Its effect (green line) increases consistently as tree complexity increases. These results indicate that when the occurrence of single-vehicle motorcycle crashes is modeled with more complex interactions (i.e., a higher level of tree complexity), the interaction between behavioral variables plays a larger role.

Conversely, the three interactions that include the county variable exhibit a relatively flat trend. These are protective equipment–county (red line); drinking–county (blue line), and cell phone–county (light blue line). These results suggest that, irrespective of how comprehensively traffic crashes are modeled, the interaction between behavioral variables and geographically heterogeneous factors (represented by the county variable) has a relatively stable effect.

The two pairs involving occupation (occupation – drinking and occupation – protective equipment) have a relatively small effect on crash fatality. The combination of occupation and drinking condition shows an upward trend (pink line), but its increase is small. The combination of occupation and protective equipment shows a slightly downward trend as tree complexity increases.

3.5. Out-of-sample classification using 2005 data

The previous sections describe the BRT models developed with the CV technique using 2004 single-vehicle motorcycle crash data. To further investigate the external validity of the BRT models, this section presents the out-of-sample classification performance using 2005 single-vehicle motorcycle crashes as testing data.

For comparison, a logistic regression and a CART model were also developed using the 2004 data. The out-of-sample classification performance of the logistic regression, CART, and BRT models were then compared using the 2005 data with the indicator of area under the receiver-operating characteristic curve (AUC). The AUC indicator is a popular performance measure for classification models. This indicator provides a single figure and has a number of desirable properties, such as independence from a chosen decision-threshold (Bradley, 1997). To enhance the confidence level, this study evaluated the AUC values in 1000 simulation runs, where each simulation randomly drew 1000 samples from the 2005 data set with a fixed percentage of fatal crashes.[5]

The logistic regression model from 2004 contains 12 variables, where variables were selected and categories merged using deviance and Wald z tests. The Hosmer–Le Cessie omnibus test fails to find evidence of a lack of fit. The 2004 CART model was developed using the Gini splitting function. The details of the estimation results of the 2004 logistic regression and CART models are provided in Appendix A (Table A1 and Fig. A1).

Table 2 shows a summary of the significant variables in the logistic regression model, the variables selected by the CART model, and the top 12 variables of the BRT *tc*-1 and *tc*-18 models. The CART model variable list follows the sequence in which the variables developed the tree, reflecting their priority in distinguishing fatal from injury-only crashes. The order of BRT model variables is also based on their relative contribution to crash fatality. The logistic regression model variable list does not reflect the sequence of the variables as used in the model. The results show that four variables – county, cell phone use, drinking condition, and road location – are crucial variables in all four models (i.e., they are significant in the logistic regression model and top the CART and BRT models variable lists). This indicates that these four variables are important, irrespective of how they are modeled. The variables crucial to the CART model are more similar to those in the BRT models than those in the logistic regression model. Variables identified by the CART model are present in either the top 12 BRT *tc*-1 or *tc*-18 models. Gender, trip purpose, illumination, and roadside variables are significant in the logistic model, but not crucial in the other three models.

Fig. 5 shows a summary of the performance rankings of 1000 simulation runs using boxplots. The bold lines in the boxes are median rankings; boxes indicate the interquartile range (IQR) (the difference between the first and third quartiles) of the rankings; whiskers are rankings that are 1.5 times the IQR above or below the median ranking; and dots outside the whiskers are outliers. Ranking first indicates the best performance, whereas ranking 20th is the worst performance. The results suggest that the CART model has the worst ranking most of the time; the first and third quartiles and the median all rank 20th. This is expected because conventional CART models focus on the largest category in unbalanced data sets (Chang and Wang, 2006), ignoring the smallest category produces lower AUC values because of the extremely low true-positive rate and high false-negative rate. The *tc*-1 BRT model is similar to a logistic regression; therefore, their performances are similar.

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

Table 2
Crucial variables for logistic regression, CART, and BRT models.

Logistic regression	CART	BRTs *tc*-1 (top 12 variables)	BRTs *tc*-18 (top 12 variables)
County	**Cell phone use**	**Drinking condition**	**County**
Cell phone use	**County**	**Cell phone use**	Month
Drinking condition	**Drinking condition**	**County**	Occupation
Road location	**Road location**	Protective equipment	**Drinking condition**
Gender	Month	**Road location**	**Cell phone use**
Trip purpose	Occupation	Movement prior to crash	Road location
Month	Sight distance	Occupation	Day of the week
Time of the day	Road type	License type	Protective equipment
Protective equipment		Time of the day	Median type
Illumination		Sight distance	Age
Median type		Road type	Road level
Roadside		License condition	Road type

The bold variables appear as the crucial variables in all the four models.

The out-of-sample classification performance of the BRT models varies with an increase in tree complexity levels. The results generally indicate a deteriorating trend (the upward median ranking) as tree complexity level increases. This is different from the in-sample validation results, which show that classification performance consistently improves with an increase in tree complexity levels. The variance in classification performance, indicated by the width of the IQR, is large when tree complexity is low or high, but small at a mid-level tree complexity of approximately 8. This suggests that models with low tree complexity underestimate the complexity of the 2005 traffic crashes, whereas those with high tree complexity overestimate the complexity of the 2005 traffic crashes. Whereas the *tc*-3 BRT model has the lowest median ranking, the BRT models with *tc*s of approximately 8 have low median rankings and efficient (i.e., low variance) classification performance.

4. Discussion

This study investigated the complexity of traffic crashes with the novel BRT method. An empirical data set, 2004–2005 Taiwanese single-vehicle motorcycle crashes, was used to demonstrate the utility of the BRT method. The analysis results show the ability of BRT models to consider many variables, explore the nonlinear relationship between the explanatory and response variable, and perform satisfactory classification compared to logistic regression and CART models.

The BRT analysis results show that the higher order models exhibit better in-sample and out-of-sample classification performance than the first-order models (the logistic regression model with main effects only and the *tc*-1 BRT model). This suggests that models containing merely first or low-order factors cannot approximate complex traffic crashes well. For complex crashes, the effect of some factors is dependent on many other factors; the factor effects can vary dramatically if driving conditions change. This partially explains why good road safety countermeasures effectively reduce most, but not all, target traffic crashes.

Despite the ability of the model to consider many variables, the results show that a few variables explain most of the variation. Three variables explain approximately half of the variation, indicating that a few variables determine the fatality of single-vehicle motorcycle crashes. Although there are few influential variables, it is unclear why the three most influential variables (county, occupation, and month) have such a large effect. The county, occupation, and month variables, respectively, represent the geographical, personal, and seasonal factors that may affect crash severity. However, the exact nature of these factors is unknown. For example, geographical factors can reflect local driving culture or local road quality; both have been crucial factors in explaining traffic crashes (Eiksund, 2009; Rakauskas et al., 2009). Therefore, further research

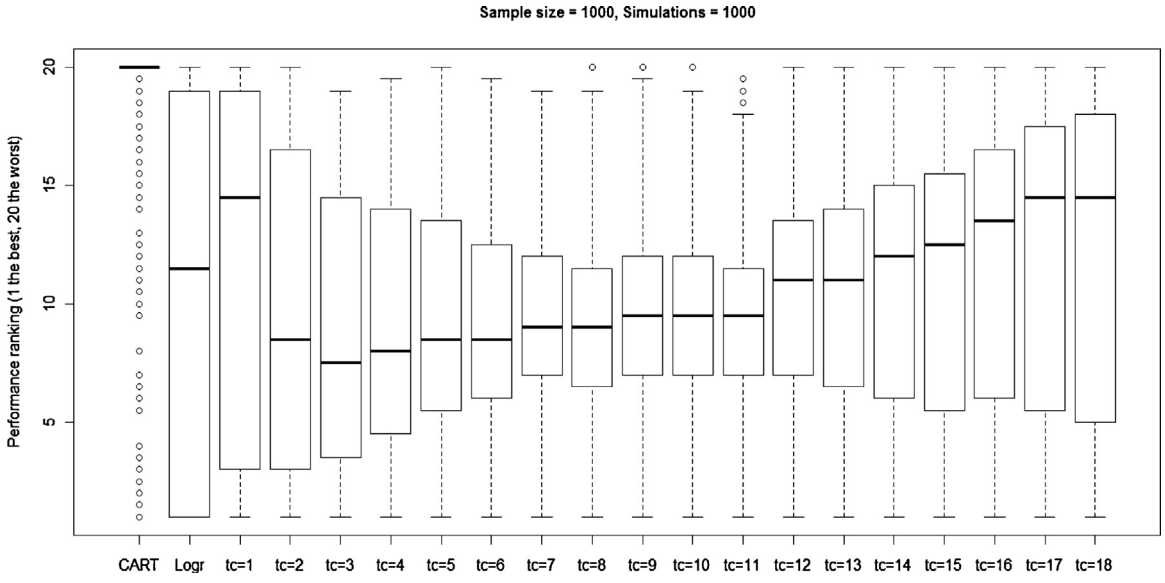

Fig. 5. Out-of-sample performance rankings using 2005 data.

should study the exact effects of such complex variables. One limitation of the relative contribution measure is a lack of confidence bounds; consequently, it is difficult to determine the significance of differences between relative contributions.

A highly branched variable (one that can be split into distinct classes) does not necessarily contribute more because the BRT models use the CV technique to reduce overfitting. For example, Fig. 2 shows that the variable cell phone use has only four categories and is ranked one of the five most influential variables, but the variable movement prior to crash has 14 categories and is ranked among the least influential variables.

Although the logistic regression, CART, and BRT models identify similar influential or significant variables, how these variables are associated with crash severity is different. The CART model is the most limited of these models because variable entrance order is the only method of determining importance ranking. Although tree-based models (including CART models) can fit nonlinear relationships, conventional CART models do not provide quantified results for interpreting the relationship between the explanatory and response variables. The logistic regression model supplies the significance of the explanatory variables, but specifying interaction terms in advance is difficult. Econometric models, including logistic models, are not designed to explore complex structures; instead, they should be used according to economic (or behavioral) theories. A parsimonious logistic model has more value than a complex logistic model if both models capture crucial factor effects. However, the modeling of traffic crashes usually faces numerous confounding factors (and categories) and nonlinear relationships. Particularly, if variables such as age or drinking condition, which exhibit a nonlinear relationship with crash severity, are not specified properly, the estimation results of logistic models may be erroneous. One way to resolve this is to combine the BRT and logistic regression models (i.e., use BRT models to explore the relationship between the considered explanatory variables and the response variable, and then transform the variables to develop a representative logistic regression model).

Of the variables considered, the behavioral variables (drinking condition, cell phone use, and protective equipment use) are particularly important in explaining unique single-vehicle motorcycle crashes. The increasing effect of behavioral interactions as tree complexity increases demonstrates this. In other words, BRT models that value behavioral interactions more approximate difficult cases more closely. This suggests that relatively unique crashes result from unexpected or undesired behaviors, the negative effects of which override the protective effect provided by road design. Safety education to adjust driver behavior and attitudes could reduce these crashes.

Data quality is an issue relevant to all the models. For example, the cell phone use variable is one of the most influential variables and its largest category is "unknown". It is such a large category because when motorcyclists die at crash sites, police record cell phone use as unknown if no witnesses or further evidence is found. Therefore, the significantly positive effect of unknown cell phone use category is a combination of cell phone use and no cell phone use. Although this category seems ineffective at explaining the relationship between cell phone use and crash severity, the BRT models show that the relative contribution of the unknown category decreases as tree complexity increases. Therefore, the unknown category approximates most fatal crashes well, but it cannot effectively explain relatively unique fatal crashes.

5. Concluding remarks

This paper applies the BRT method to investigate the factor complexity of single-vehicle motorcycle crashes. The empirical study demonstrates the advantages of BRT models. These advantages

Table A1
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	Protective equipment	Yes	−1.356***	0.258
Trip purpose	Sightseeing	1.394*	4.031		Other	−1.745***	0.175
	Others	0.457*	1.580	Cell phone use	Handheld	−15.440	0.000
Month	November	0.424~	1.528		Handfree	−1.300	0.273
Time of the day		−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		BAC > 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				

~ <0.10.
* <0.05.
** <0.01.
*** <0.001.

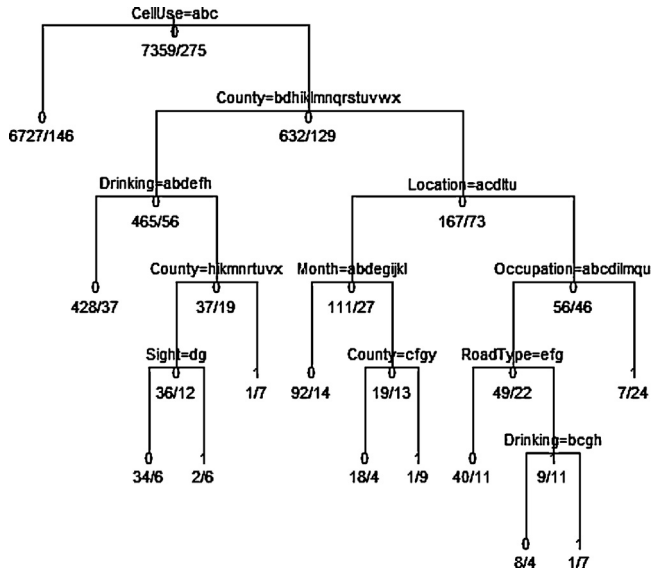

Fig. A1. CART model using 2004 data.

include no requirement to pre-specify function form, select variables, or merge categories; and the abilities to consider several explanatory variables, account for nonlinear relationships, provide satisfactory classification performance, and supply quantitative results. However, the model has disadvantages. As with other data mining methods, some parameters must be tested for optimal settings when developing BRT models. These parameters include tree complexity, learning rate, and bagging fraction. Additionally, as tree complexity increases, computation time also increases. The balance between computation cost, tree complexity and learning rates must be considered when developing BRT models.

Although it provides interpretable statistics, the BRT method is a data-driven approach. Therefore, BRTs may be particularly useful to explore unknown relationships between crash outcomes and influential factors, especially if the relationships are complex or nonlinear. These explored relationships could be the foundation or reference points for developing behavioral theories.

Acknowledgments

The authors would like to thank the reviewers for their helpful comments and the National Science Council of Taiwan for their financial support (NSC 98-2410-H-424-018). Thanks are also due to the participants of the Third International Conference on Road Safety and Simulation, Indianapolis, IN, USA, for providing valuable feedback on the content of this article.

Appendix A.

See Fig. A1 and Table A1.

References

Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis and Prevention 34 (6), 729–741.
Bayam, E., Liebowitz, J., Agresti, W., 2005. Older drivers and accidents: a meta analysis and data mining application on traffic accident data. Expert Systems with Applications 29 (3), 598–629.
Bedard, M., Guyatt, G.H., Stones, M.J., Hirdes, J.P., 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis and Prevention 34 (6), 717–727.
Bina, M., Graziano, F., Bonino, S., 2006. Risky driving and lifestyles in adolescence. Accident Analysis and Prevention 38 (3), 472–481.
Bradley, A.P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition 30 (7), 1145–1159.
Breiman, L., Friedman, J.H.R.A.O., Stonec, J., 1984. Classification and Regression Trees. Wadsworth, New York.
Chan, K.Y., Loh, W.Y., 2004. Lotus: an algorithm for building accurate and comprehensible logistic regression trees. Journal of Computational and Graphical Statistics 13 (4), 826–852.
Chang, H.L., Yeh, T.H., 2007. Motorcyclist accident involvement by age, gender, and risky behaviors in Taipei, Taiwan. Transportation Research. Part F: Traffic Psychology and Behaviour 10 (2), 109–122.
Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38 (5), 1019–1027.
De'ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology 88 (1), 243–251.
Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accident Analysis and Prevention 40 (4), 1257–1266.
Eiksund, S., 2009. A geographical perspective on driving attitudes and behaviour among young adults in urban and rural norway. Safety Science 47 (4), 529–536.
Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77 (4), 802–813.
Elmitiny, N., Yan, X., Radwan, E., Russo, C., Nashar, D., 2010. Classification analysis of driver's stop/go decision and red-light running violation. Accident Analysis and Prevention 42 (1), 101–111.
Elvik, R., 2002. The importance of confounding in observational before-and-after studies of road safety measures. Accident Analysis and Prevention 34, 631–635.
Elvik, R., 2003. Assessing the validity of road safety evaluation studies by analysing causal chains. Accident Analysis and Prevention 35 (5), 741–748.
Fleury, D., Brenac, T., 2001. Accident prototypical scenarios, a tool for road safety research and diagnostic studies. Accident Analysis and Prevention 33 (2), 267–276.
Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29 (5), 1189–1232.
Geurts, K., Thomas, I., Wets, G., 2005. Understanding spatial concentrations of road accidents using frequent item sets. Accident Analysis and Prevention 37 (4), 787–799.
Golob, T.F., Recker, W.W., 2004. A method for relating type of crash to traffic flow characteristics on urban freeways. Transportation Research. Part A: Policy and Practice 38 (1), 53–80.
Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests. Accident Analysis and Prevention 41, 98–107.
Hastie, T., Tibshirani, R.J., Friedman, J.H., 2009. The Elements of Statistical Learning, 2nd ed. Springer-Verlag, New York, NY, USA.
Hauer, E., 2006. Cause and effect in observational cross-section studies on road safety. In: 85th Annual Meeting of the Transportation Research Board, Washington, DC, USA.
Hotz, G.A., Cohn, S.M., Popkin, C., Ekeh, P., Duncan, R., Johnson, W., Pernas, F., Selem, J., 2002. The impact of a repealed motorcycle helmet law in miami-dade county. Journal of Trauma – Injury Infection and Critical Care 52 (3), 469–473.
Hundley, J.C., Kilgo, P.D., Miller, P.R., Chang, M.C., Hensberry, R.A., Meredith, J.W., Hoth, J.J., 2004. Non-helmeted motorcyclists: a burden to society? – a study using the national trauma data bank. Journal of Trauma – Injury Infection and Critical Care 57 (5), 944–949.
Kim, K., Pant, P., Yamashita, E.Y., 2008. Hit-and-run crashes use of rough set analysis with logistic regression to capture critical attributes and determinants. Transportation Research Record 2083, 114–121.
Li, M.D., Doong, J.L., Chang, K.K., Lu, T.H., Jeng, M.C., 2008. Differences in urban and rural accident characteristics and medical service utilization for traffic fatalities in less-motorized societies. Journal of Safety Research 39 (6), 623–630.
Lin, M.R., Chang, S.H., Huang, W.Z., Hwang, H.F., Pai, L., 2003a. Factors associated with severity of motorcycle injuries among young adult riders. Annals of Emergency Medicine 41 (6), 783–791.
Lin, M.R., Chang, S.H., Pai, L., Keyl, P.M., 2003b. A longitudinal study of risk factors for motorcycle crashes among junior college students in taiwan. Accident Analysis and Prevention 35 (2), 243–252.
Loh, W.Y., 2006. Logistic regression tree analysis. In: Pham, H. (Ed.), Handbook of Engineering Statistics. Springer.
Mirabadi, A., Sharifian, S., 2010. Application of association rules in iranian railways (rai) accident data analysis. Safety Science 48 (10), 1427–1435.
Montella, A., 2011. Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. Accident Analysis and Prevention 43 (4), 1451–1463.
Montella, A., Aria, M., D'ambrosio, A., Mauriello, F., 2011. Analysis of powered two wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis and Prevention, http://dx.doi.org/10.1016/j.aap.2011.04.025, in press.
Mussone, L., Kim, K., 2010. The analysis of motor vehicle crash clusters using the vector quantization technique. Journal of Advanced Transportation 44 (3), 162–175.
Pande, A., Abdel-Aty, M., 2009. Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. Safety Science 47 (1), 145–154.
R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Australia.
Rakauskas, M.E., Ward, N.J., Gerberich, S.G., 2009. Identification of differences between rural and urban safety cultures. Accident Analysis and Prevention 41 (5), 931–937.

Reason, J., 1997. Managing the Risks of Organizational Accidents. Ashgate, Aldershot.
Ridgeway, G., 2007. Generalized Boosted Models: A Guide to the gbm Package.
Rutter, D.R., Quine, L., 1996. Age and experience in motorcycling safety. Accident Analysis and Prevention 28 (1), 15–21.
Saunier, N., Mourji, N., Agard, B., 2011. Mining microscopic data of vehicle conflicts and collisions to investigate collision factors. Transportation Research Record 2237, 41–50.
Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S., Barbone, F., 2002. Risk factors for fatal road traffic accidents in Udine, Italy. Accident Analysis and Prevention 34 (1), 71–84.
Witten, I., Frank, E., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann.
Wong, J.T., Chung, Y.S., 2007. Rough set approach for accident chains exploration. Accident Analysis and Prevention 39 (3), 629–637.
Wong, J.T., Chung, Y.S., 2008a. Analyzing heterogeneous accident data from the perspective of accident occurrence. Accident Analysis and Prevention 40 (1), 357–367.
Wong, J.T., Chung, Y.S., 2008b. Comparison of methodology approach to identify causal factors of accident severity. Transportation Research Record 2083, 190–198.
Yagil, D., 1998. Instrumental and normative motives for compliance with traffic laws among young and older drivers. Accident Analysis and Prevention 30 (4), 417–424.
Zhang, M.H., Xu, Q.S., Daeyaert, F.P.J.L., Massart, D.L., 2005. Application of boosting to classification problems in chemometrics. Analytica Chimica Acta 544 (1–2), 167–176.
Zhang, W., Gkritza, K., Keren, N., Nambisan, S., 2011. Age and gender differences in conviction and crash occurrence subsequent to being directed to iowa's driver improvement program. Journal of Safety Research 42 (5), 359–365.