2018

Applying Machine Learning Techniques to Analyze the Pedestrian and Bicycle Crashes at the Macroscopic Level

Md Sharikur Rahman

APPLYING MACHINE LEARNING TECHNIQUES TO ANALYZE THE PEDESTRIAN AND BICYCLE CRASHES AT THE MACROSCOPIC LEVEL

by

MD SHARIKUR RAHMAN
B.Sc. Bangladesh University of Engineering and Technology, 2014

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Civil, Environmental and Construction Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2018

Major Professor: Mohamed Abdel-Aty

ABSTRACT

This thesis presents different data mining/machine learning techniques to analyze the vulnerable road users' (i.e., pedestrian and bicycle) crashes by developing crash prediction models at macro-level. In this study, we developed data mining approach (i.e., decision tree regression (DTR) models) for both pedestrian and bicycle crash counts. To author knowledge, this is the first application of DTR models in the growing traffic safety literature at macro-level. The empirical analysis is based on the Statewide Traffic Analysis Zones (STAZ) level crash count data for both pedestrian and bicycle from the state of Florida for the year of 2010 to 2012. The model results highlight the most significant predictor variables for pedestrian and bicycle crash count in terms of three broad categories: traffic, roadway, and socio demographic characteristics. Furthermore, spatial predictor variables of neighboring STAZ were utilized along with the targeted STAZ variables in order to improve the prediction accuracy of both DTR models. The DTR model considering spatial predictor variables (spatial DTR model) were compared without considering spatial predictor variables (aspatial DTR model) and the models comparison results clearly found that spatial DTR model is superior model compared to aspatial DTR model in terms of prediction accuracy. Finally, this study contributed to the safety literature by applying three ensemble techniques (Bagging, Random Forest, and Boosting) in order to improve the prediction accuracy of weak learner (DTR models) for macro-level crash count. The model's estimation result revealed that all the ensemble technique performed better than the DTR model and the gradient boosting technique outperformed other competing ensemble technique in macro-level crash prediction model.

ACKNOWLEDGMENT

I would like to express the deepest appreciation and gratitude to my honourable supervisor Professor Mohamed Abdel-Aty, for his invaluable guidance, advice, support, and encouragement toward successful completion of the master's degree. I would also like to gratefully acknowledge Signal Four Analytics (S4A) and Florida Department of Transportation (FDOT) for providing access to Florida crash and geospatial data. I wish to acknowledge the support of my respectable committee members, Dr. Samiul Hasan, and Dr. Qing Cai. I would also like to thank my beloved parents who always encourage me a lot. Last but not the least, many thanks to my beloved wife Laizu Akter. She always stood by me and cheered me up through good and bad times. Without her support and encouragements, I cannot achieve anything.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The most active forms of transportation are walking and bicycling which have the lowest impact on the environment and improve physical health of pedestrians and bicyclists. Transportation agencies are increasingly promoting walking and bicycling options for short distance trips to mitigate climate change and obesity problem among adults. However, the most common problem impeding the preference of walking and bicycling is traffic safety concerns. According to the latest traffic safety data from the National Highway Traffic Safety Administration (NHTSA), pedestrian and bicycle deaths have increased by 9.0 % and 1.3 %, respectively in 2016 compared to the calendar year 2015 (NHTSA, 2017a). Thus, the safety challenges associated with pedestrians and bicyclists remain an important concern for transportation policy. The safety risk posed to active transportation users in Florida is exacerbated compared to active transportation users in the US. In 2015, while the national average for pedestrian and bicyclist fatalities per 100,000 population was 1.67 and 2.50, respectively, the corresponding number for the state of Florida was 3.10 (ranked second among all states) and 7.40 (ranked first among all states), which clearly present the challenge faced in Florida (NHTSA, 2017b, 2015). The crash prediction models applied to the pedestrian and bicycle crashes would give some valuable insights for a transportation planner to identify the contributing factors related to pedestrians and bicyclists' crashes which might be helpful for policy implications at a planning level.

1.1 Motivation for The Study

In transportation safety research, crash prediction models are developed for two levels: (1) micro-level (2) macro-level. The former one focuses on crashes at a segment or intersection

to identify the influence of contributing factors with the objective of offering engineering solutions. On the other hand, the macro-level crashes from a spatial aggregation such as traffic analysis zone, census block, census tract, county are considered to quantify the significant factors at a macro-level so that it can provide countermeasures from a planning perspective. Statistical models, such as Poisson and negative binomial regression, have been employed to analyze both micro- and macro-level crashes for many years. However, statistical models have their own model-specific assumptions which lead to inaccurate results of injury likelihood (Chang and Chen, 2005). In this regard, this study contributes to the safety literature by undertaking pedestrian and bicycle crash prediction model using the most widely applied data mining technique: decision tree regression (DTR). To the best of our knowledge, none of the studies have explored data mining techniques in analyzing pedestrian and bicycle crashes at the macro-level. In this regard, three broad categories of predictor variables including traffic, roadway, and socio-demographic characteristics are considered in the DTR model development and validation. In addition, the attributes of the neighboring zones are considered as predictor variables along with the targeted STAZs attributes in DTR models to improve the prediction accuracy of pedestrian and bicycle crashes. Furthermore, the current study has undertaken some ensemble techniques (i.e. bagging, random forest, and gradient boosting) to improve the prediction accuracy of the DTR models considered as weak learner which provides valuable insights on advancing crash prediction modeling techniques for macro-level crash analysis.

1.2 Study Methodology and Objective

The most common approach employed to address the macro level crash risk in safety literature is developing the statistical crash frequency models. In this modelling framework,

the impact of independent/exogenous variables are evaluated for a given dependent variable. However, statistical models have their own model-specific assumptions which lead to inaccurate results of injury likelihood (Chang and Chen, 2005). In our current study, we apply machine learning/data mining approach to develop the pedestrian and bicycle crash prediction model using decision tree regression (DTR). In this regard, three broad categories of predictor variables including traffic, roadway, and socio-demographic characteristics are considered in the DTR model development and validation. In addition, the current study undertakes the attributes of the neighboring zones as predictor variables to improve the prediction accuracy of pedestrian and bicycle crashes. Variable importance of DTR models for both pedestrians and bicyclists crashes were computed in order to perform the policy analysis at macro-level. Furthermore, some ensemble techniques (i.e. bagging, random forest, and gradient boosting) were employed to improve the prediction accuracy of the DTR models considered as weak learner which provides valuable insights on advancing crash prediction modeling techniques for macro-level crash analysis. The models are estimated by using data from Florida at the Statewide Traffic Analysis Zone (STAZ) level for the year of 2010-2012.

1.3 Thesis Structure

The rest of the thesis is organized as follows: Chapter 2 provides a brief review of relevant earlier research. Chapter 3 describes the modelling methodologies such as decision tree regression and the ensemble techniques employed of this study. Chapter 4 discusses a detailed summary of the data source and predictor variables considered for the analysis. Model estimation results are reported in Chapter 5. Finally, a summary of model findings and conclusions are presented in Chapter 6.

CHAPTER 2: LITERATURE REVIEW

The field of crash modeling is vast. Several research efforts have been conducted throughout the years for developing crash prediction models. Generally, there are two types of modelling techniques had been employed throughout the years (1) statistical models (2) data mining techniques. In this chapter, we present a detailed discussion of the various model structures (statistical and data mining) used in existing literature and position our current study in context.

2.1 Earlier Research

Road traffic accidents are highly recognized as a national health problem which affects the society both emotionally and economically (Blincoe et al., 2002; NHTSA, 2005). There is a considerable number of research efforts that have been examined in crash frequency estimation (vehicle, pedestrian, and bicycle) (see (Lord and Mannering, 2010) for a detailed review). These studies have been conducted for different modes of vehicle (automobiles and motorbikes), pedestrian and bicycle, and for different scales - micro (such as intersection and segment) and macro-level (such as census tract, traffic analysis zone (TAZ), county). It is beyond the scope of this paper for exhaustive review of micro-level (see (Abdel-Aty et al., 2016; Eluru et al., 2008; Lord et al., 2005) for detailed micro-level literature review) and macro-level (see (Cai et al., 2017, 2016; Lee et al., 2018) for detailed macro level literature review) crash frequency studies. These studies have heavily focused on econometric statistical modeling approaches (Wang et al., 2018; Yuan and Mohamed Abdel-Aty, 2018) for the prediction of traffic crashes with exploring significant contributing factors related to the crash occurrence. However, statistical models can lead to inaccurate estimations of injury likelihood

if prespecified model assumptions and underlying relationship between dependent and independent variables of these models are invalid (Chang and Chen, 2005).

Moreover, the presence of large number of zeroes in pedestrian and bicycle crashes is one of the major methodological challenges in statistical modeling to analyze the contributing factors related to pedestrian and bicyclist crashes. In crash count models, the presence of excess zeros may result from two underlying processes or states of crash frequency likelihoods: crash-free state (or zero crash state) and crash state (see (Mannering et al., 2016) for more explanation). In the presence of such dual-state, application of single-state model may result in biased and inconsistent parameter estimates. In a statistical framework, the potential relaxation of the single-state count model is zero inflated model for addressing the issue of excess zeros: zero inflated (ZI) model (Shankar et al., 1997). But, several research studies have criticized the application of dual state ZI models for traffic safety analysis (Lord et al., 2007, 2005; Son et al., 2011). A ZI model assumes that two types of zeros exist, i.e., sampling zeros and structural zeros. For traffic safety, the structural zeros correspond to inherently safe conditions implying zero crash by nature and the sampling zeros correspond to potential crash conditions implying zero crash only by chance (Lord et al., 2007, 2005). Hence, the statistical assumptions of having structural zeroes is unrealistic as a traffic crash could occur under any conditions.

Recently, data mining and/or machine learning techniques have become popular in transportation safety research to determine the factors associated with traffic crashes. Unlike statistical models, machine learning techniques are non-parametric methods which do not require any pre-defined underlying relationships between target variable and predictors (Tavakoli Kashani et al., 2014). Among the machine learning techniques, the decision tree model has gained much popularity in transportation safety literature which can identify and

easily explain the complex patterns associated with crash risk (Chang and Chen, 2005; Chang and Chien, 2013; Chang and Wang, 2006; Pande et al., 2010). To overcome the shortcoming of the statistical modelling, decision tree can be a preferred alternative for forecasting traffic crashes with reasonable interpretations. Unlike statistical models, decision trees do not need any predefined model assumption and underlying relationship between dependent and independent variables. It does deal well with multicollinear independent variables and does treat satisfactorily discrete variables with more than two levels (Karlaftis and Golias, 2002; Washington and Wolf, 1997). Moreover, decision tree models can help in deciding how to subdivide heavily skewed target variables (i.e., zero crash counts) into ranges while the statistical modeling has some limitations for dealing with heavily skewed data (Song and Lu, 2015). Therefore, decision tree models might be a preferred option to analyze heavily skewed response variable which is most common in pedestrian and bicycle crashes. A summary of earlier studies employing decision tree models in traffic safety literature is presented in Table 1 (Abdel-Aty et al., 2005; Chang and Chen, 2005; Chang and Chien, 2013; Chang and Wang, 2006; De Oña et al., 2013; Eustace et al., 2018; Iragavarapu et al., 2015; Karlaftis and Golias, 2002; Kashani and Mohaymany, 2011; Montella et al., 2012; Pande et al., 2010; Tavakoli Kashani et al., 2014; Wah et al., 2012; Zheng et al., 2016). The information provided in the table includes the study unit considered, the methodological approach employed, the target variables analyzed in the decision tree framework. The following observations can be inferred from the table. From the table, it is evident that all the existing decision tree-based safety studies are conducted at a micro-level such as roadway segments and intersections. To the best our knowledge, none of the studies have explored decision tree methods in order to build the crash prediction model at the macro-level.

Table 1 Summary of Previous Traffic Safety Studies Using Decision Tree and Ensemble Techniques

Area of Interest	Studies	Study Unit (Scale)	Methodology	Target Variables Analyzed
Decision Tree	Kashani *et al.* (2014)	Roadway segment (Micro)	Classification Tree	Injury severity level - Injury, fatality
	Zheng *et al.* (2016)	Highway-rail grade crossings (Micro)	Classification Tree	Highway-rail grade crossings crash
	Kashani *et al.* (2011)	Two-lane, two-way rural roads segments (Micro)	Classification Tree	Injury severity level- Light injury, Serious injury, Fatality
	Iragavarapu *et al.* (2015)	Road segments-pedestrian crash (Micro)	Classification Tree	Injury severity level- fatal or non-fatal
	Chang *et al.* (2005)	National Freeway (Micro)	Classification Tree	Injury Severity level (0– 4, 4 representing 4 or more crashes)
	Wah *et al.* (2005)	Roadway segments (Micro)	Classification Tree	Category of Frequencies of motorcycle Accidents- Zero frequency (0), Low frequency (1-19), High frequency (20 and above)
	Chang *et al.* (2006)	Roadway segments (Micro)	Classification Tree	Injury severity level- fatality, injury, no-injury
	Pande *et al.* (2010)	Roadway segments (Micro)	Classification Tree	Binary variable-Crash vs Non-crash
	Chang *et al.* (2013)	National freeways (Micro)	Classification Tree	Injury severity level- fatality, injury, no-injury
	Ona *et al.* (2013)	Road Segments-Rural highways (Micro)	Classification Tree	Accident Severity- slightly injured, killed or seriously injured (KSI) (state B)
	Montella *et al.* (2012)	Roadway segments- Powered two-wheeler crashes (Micro)	Classification Tree	Several response variables- severity, crash type, involved vehicles, alignment
	Eustace *et al.* (2018)	Road segments (Micro)	Classification Tree	Injury severity level-fatal/injury, and property damage only
	Abdel-Aty *et al.* (2005)	Road segments (Micro)	Regression Tree	Total crash, Angle crash, Left turn crash, Head on crash, pedestrian crash, rear-end crash, right turn crash, sideswipe crash
	Karlaftis *et al.* (2002)	Road segments (Micro)	Regression Tree	Total number of crash
Ensemble Techniques	Sohn *et al.* (2002)	Road segments (Micro)	Arcing and bagging	Injury severity level-bodily injury and property damage

It is also noticed that most of the model structures employed in developing decision trees are classification trees except for two studies (Abdel-Aty et al., 2005; Karlaftis and Golias, 2002) which conducted hierarchical tree-based regression for developing the micro-level crash prediction model. Within the decision tree structure, those studies did not explore the total number of pedestrian and bicycle crashes while they have predominantly analyzed crash frequency by severity levels or other different attribute levels.

One of the basic assumptions of most of the modelling techniques are that observations are independent from each other. Nevertheless, this assumption is often violated in traffic data because of possible correlation among observations. For instance, some observations that are from the same spatial units may have common unobserved factors. In macro-level analysis, crashes occurring in a spatial unit are aggregated to obtain the crash frequency. However, this aggregation process might introduce errors in identifying the predictor variables for the spatial unit. For example, a crash occurring closer to the boundary of the unit might be strongly related to the neighboring zone than the actual zone where the crash occurred. There is a considerable amount of research that have been undertaken to accommodate for such spatial unit induced bias (Huang et al., 2010; Lee et al., 2015; Siddiqui et al., 2012). The most recent study proposed the consideration of exogenous variables from neighboring zones for accounting for spatial dependency which was called spatial spillover model (Cai et al., 2016). And, the research effort revealed that models with spatial exogenous variables significantly outperformed the model that did not consider the spatial exogenous variables. In our analysis, we introduce spatial predictor (exogenous) variables from neighboring zones for improving the prediction accuracy. Apart from the statistical and data mining methods, simulation techniques can identify the significant contributing factors related to the crash occurrence (Ekram and Rahman, 2018; Rahman et al., 2018; Rahman and Abdel-Aty, 2018).

However, decision trees can be unstable because of the small variations in the data which might result in a completely different tree being generated. This would result in a good prediction for the majority class, but a relatively poor prediction for the minority class.

This problem can be mitigated by using decision trees within an ensemble (36). In machine learning, ensemble methods are used to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Data ensemble combines various results obtained from a single classifier fitted repeatedly based on bootstrap resamples. The advantage of ensemble lies in the possibility that the difference of result caused by the variance of input data may be reduced by combining each classifier's output. To the best of the authors' knowledge, none of the studies have implemented ensemble techniques in the transportation safety field in order to improve the prediction accuracy except for Sohn et al. (2003), which employed arcing and bagging as ensemble techniques (Table 1). The result suggests that ensemble algorithms such as bagging and arcing improved the prediction accuracy of traffic crashes compared to individual classifier decision tree.

In summary, the current study contributes to non-motorized macro-level crash analysis along three directions: (1) evaluate the regression tree models for both pedestrian and bicycle crashes (2) consider spatial predictor variables in crash prediction models (3) introduction of ensemble techniques (i.e., bagging, random forests, and gradient boosting) in order to improve the prediction accuracy of macro-level crash analysis.

2.2 Current Study

The literature review clearly highlights the disadvantages of statistical modeling techniques over data mining frameworks in the burgeoning safety literature. And, it is clearly noted that data mining technique can help in deciding how to subdivide heavily skewed target variables (i.e., zero crash counts) into ranges which is essential for pedestrian and bicycle

crashes. In this context, the current study makes three important contributions for the macro-level crash risk.

First, we apply the data mining techniques for both pedestrian and bicycle crash risk, which is the first application of decision tree regression models in the growing traffic safety literature at macro-level. To facilitate a policy analysis at the macro-level, variable importance of DTR models for both pedestrians and bicyclists crashes were computed.

Second, within the decision tree framework, we also accommodate spatial predictor variables from neighboring STAZs in order to improve the prediction accuracy of DTR models for both pedestrian and bicycle crashes.

Third, we undertake some ensemble techniques such as bagging, random forest, and gradient boosting to improve the prediction accuracy of pedestrian and bicycle crashes. Specifically, we examine performance in model estimation and prediction for bagging, random forest, and gradient boosting techniques compared to decision tree regression model.

Empirically, the study develops crash frequency model for both pedestrian and bicycle crash. The models are estimated using STAZ level crash data for the year 2010-2012 for the state of Florida. The model results offer insights on important variables affecting crash frequency.

2.3 Summary

This chapter presented a detailed summary of modelling techniques employed in earlier studies for predicting crashes. Further, the chapter positioned the current research work in context. The modelling framework employed in this study is described in detail in the subsequent chapter.

CHAPTER 3: METHODOLOGY

There are two types of decision tree-based methods: classification tree and regression tree. The former is designed to partition data based on the discrete nature of categorical target variables, while the latter is to partition (regress) data on the basis of continuous response data. The target variables in this study are pedestrian and bicycle crashes in each STAZ. Hence, this paper focuses on the latter method regression tree and some ensemble techniques applied to improve the forecasting accuracy.

3.1 Regression Tree Framework

A regression tree is referred to a set of rules for dividing a large collection of observations into smaller homogeneous groups based on the predictor (independent) variables with respect to a continuous target (dependent) variable. The methods used to estimate regression trees have been around since the early 1960s and are sometimes referred to as classification and regression tree (CART) (Breiman et al., 1998). Generally, there are two key questions for the development of a regression tree : (1) which variable of all predictor variables offered in the model should be selected to produce the maximum reduction in variability of the response (target) variable, (2) which value of the selected predictor variable (discrete or continuous) results in the maximum reduction in variability of the response variable. Numerical search procedure is undertaken to iterate these two steps until all the observations are portioned into a smaller homogenous group (Washington, 2000).

In this paper, the focus of the regression tree model is to predict the total number of crashes. Let us assume that the response variable, Y_n (total number of crashes), is a column vector of n random variables, and $X_{n,p}$ is a matrix of (p-1) random predictor variables measured

for n cases. The equation system for modeling regression tree, the deviance D or sum of square (SSE) is defined as follows:

$$\text{SSE}=D=\sum_{l=1}^{L}(Y_l-\mu)^2 \quad (1)$$

$$\mu=\frac{1}{L}\sum_{l=1}^{L}Y_l=\text{Arithmetic mean of Y} \quad (2)$$

Where,

D = total deviance of Y, or the sum of squared errors (SSE);

Y_l= lth observation in column vector Y; and

L = sample size over which D is calculated ($L = n$ for total sample)

The observations in Y are partitioned based on a predictor variable X_1 (which variable results in the maximum reduction in variability of the response variable) that results in two subsamples, say samples b and c, each containing M and N of the original L observations (M + N = L). If the overall sample deviance is D_a, then the deviance reduction function is

$$\Delta=D_a-D_b-D_c \quad (3)$$

Where, Δ is the deviance reduction when sample a is partitioned on X_1 to obtain subsamples b and c,

$$D_a=\sum_{l=1}^{L}(Y_{(a)l}-\mu_{(a)})^2=\text{total deviance in sample (node) a} \quad (4)$$

$$D_b=\sum_{l=1}^{M}(Y_{(b)l}-\mu_{(b)})^2=\text{total deviance in sample (node) b} \quad (5)$$

$$D_c=\sum_{l=1}^{N}(Y_{(c)l}-\mu_{(c)})^2=\text{total deviance in sample (node) c} \quad (6)$$

$$\mu_{(b)}=\frac{1}{M}\sum_{m=1}^{M}Y_m=\text{Arithmetic mean of subsample (node) b} \quad (7)$$

$$\mu_{(c)}=\frac{1}{N}\sum_{n=1}^{N}Y_n=\text{Arithmetic mean of subsample (node) c} \quad (8)$$

It is worth mentioning that M is the sample size of subsample (node) b, and N is the sample size of subsample (node) c. In regression tree, predictor variable X_i taken from $X_{n,p}$ is sought to partition the column vector Y such that the deviance reduction function showed in Equation 9 is maximized.

$$\Delta = \sum_{l=1}^{L} (Y_{(a)l} - \mu_{(a)})^2 - \sum_{m=1}^{M} (Y_{(b)l} - \mu_{(b)})^2 - \sum_{n=1}^{N} (Y_{(a)l} - \mu_{(a)})^2 \quad (9)$$

While searching the matrix from $X_{n,p}$, two items must be sought to maximize Equation 9: the variable X_i and the numerical value on which the corresponding partition of Y will produce the maximum reduction of the deviance reduction function. When this maximal partition is found, the original data in node a are partitioned into two subsamples b and c having minimal combined deviance compared with all possible subsamples. Thus, the reduction in node a deviance is greatest when the deviances at nodes b and c are smallest. As mentioned earlier, numerical search procedures are used to maximize Equation 9.

In regression tree, tree growth will continue until there are homogenous observations in each terminal node. At first, the regression tree produces the maximal tree with a complex structure that overfits the training data. However, maximal tree produces good prediction accuracy in training data but worse prediction accuracy in testing sample. To have better understanding, complex tree overfits the training observations which results in overstated confidence in predictions and inclusion of insignificant predictor variables. The most common method used to reduce overfitting problem is called pruning. This method uses criteria about model complexity to trim the full tree model to a smaller and more manageable or practical tree size which reduce overfitting significantly (Washington, 2000; Washington and Wolf, 1997). The last step of building a regression tree is to select an optimal tree from the pruned trees. The principle behind selecting the optimal tree is to find a tree with respect to a measure of misclassification cost on the testing dataset so that the information in the learning dataset will not overfit. The misclassification cost depends significantly on size of the tree and model with the lower misclassification cost for both training and testing sample is the preferred regression tree model. For example, the misclassification cost for the learning (training) data

decreases monotonically with increasing the size of a tree, indicating that the saturated tree always gives the best fit to the learning data. On the contrary, when tree grows larger and larger, the misclassification cost for the testing data decreases first and then increases after reaching a minimum. This indicates that the saturated tree is greatly overfitted when applied to analyze the testing data. The optimal tree can be determined when the misclassification costs reach a minimum for both the training and testing data (see (Breiman et al., 1998) for a detailed review).

3.2 Ensemble Techniques

An ensemble technique is defined by a set of individually trained classifiers whose predictions are combined in order to improve the prediction accuracy of a single classifier (i.e., regression tree). The prediction of an ensemble technique typically requires more computation compared to a single learner so that ensembles techniques compensate poor learning algorithms by performing a lot of extra computation. In this paper, we have undertaken bagging, random forests, and boosting as methods for creating three ensemble techniques of regression tree to construct more powerful prediction models.

The basic idea underlying bagging is to reduce the variance of the decision tree that creates several subsets of data from the training sample with replacement and build the final output averaging all the predictions. To be more specific, if several similar data sets are created by resampling with replacement which is called bootstrapping and a number of regression trees are grown without pruning and averaged, the variance component of the output error is reduced. Mathematically, it is possible to calculate $\hat{f}^1(x)$, $\hat{f}^2(x)$,. . . ., $\hat{f}^B(x)$, using B separate training sets, and averaging them in order to obtain a single low variance statistical learning model, given by Equation 10:

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x) \quad (10)$$

However, this is not practical because the dataset does not have access to multiple training sets. Hence, the sample can bootstrap by taking repeated samples from the training data set (James, G., Witten, D., Hastie, T., & Tibshirani, 2013). This can generate *B* different bootstrapped training data sets and train the model on the b^{th} bootstrapped training set in order to get $\hat{f}^{*1}(x), \widehat{f}^{*2}(x)$ … … $\hat{f}^{*B}(x)$, and finally average all the predictions (See Equation 11)

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x) \quad (11)$$

This empirical formulation is called bagging.

Random forest is similar to bagging in that bootstrap samples are drawn to construct multiple trees. The main difference from bagging is that random forest compute one extra step having the random selection of predictor variables rather than using all variables to grow the trees. The number of predictors used to find the best split at each node is a randomly chosen subset of the total number of predictors. As with boosting tree, the trees are grown to maximum size without pruning, and aggregation is by averaging the trees. Suppose, there are N observations and M predictor variables in the learning dataset. At first, subsets of data from the training sample with replacement are taken from full dataset like bagging. Then, a subset of M predictor variables is selected randomly, and whichever variables give the best split is used to split the node iteratively. The main advantages of random forest over bagging is that random predictor selection diminishes correlations among unpruned trees and constructs a learning model with low bias and variance at the same time.

Boosting is another approach for improving the predictions resulting from a series of decision trees. Like bagging, boosting is an efficient approach that creates several subsets of data which constructs a final output by averaging all the prediction of resulting trees. Unlike

bagging, the training set used for each individual learner is chosen based on the performance of the earlier learner(s). In Boosting, observations that are incorrectly predicted by previous classifiers in the individual learners are chosen more often than observations that were correctly predicted Consequently, boosting attempts to produce new learners for its ensemble that are better able to correctly predict examples for which the current ensemble performance is poor. It is worth mentioning that in bagging, the resampling of the training set is not dependent on the performance of the earlier classifiers. In machine learning, gradient boosting technique has gained much popularity for building powerful predictive models from weak learners. Specifically, gradient boosting techniques uses a base weak learner and try to boost the performance of weak learners by iteratively shifting the focus towards problematic observations that were difficult to predict. This ensemble technique identifies problematic observations by large residuals computed in the previous iterations (Mayr et al., 2014).

3.3 Summary

The main objective of the study is to develop data mining modelling techniques to predict the pedestrian and bicycle crash in a zonal level. This chapter presented a detailed discussion of the modelling techniques employed in this study.

CHAPTER 4: DATA PREPARATION

The previous chapter provided a detailed discussion about the modeling framework employed in the current research effort. This chapter presents characteristics of the data employed for analysis including the source of the data, the compilation of response and predictor variables considered in the analysis.

4.1 Data Source

This study is focused on pedestrian and bicycle crashes at the STAZ level. The data provides crash information for 8,518 STAZs, with an average area of 6.472 square miles. Data for the empirical study were obtained from Florida Department of Transportation (FDOT), Crash Analysis Reporting System (CARS) and Signal Four Analytics (S4A) databases for the years 2010 to 2012. CAR and S4A are long and short forms of crash reports in the State of Florida, respectively. The Long Form crash report is used to obtain detailed information on major crashes such as accident resulting in injuries or crashes involving felonious activities (such as hit-and-run or driving under influence). Short Form crash reports depict the reports based on all other traffic crashes. Thus, when integrated, a complete representation of road crashes in Florida is generated.

4.2 Response Variables

The data provides crash information for 8,518 STAZs. About 16,240 pedestrians and 15,307 bicycles involved crashes that occurred in Florida in these 3 years' period were compiled for the analysis. Among the STAZs, 46.18% of them have zero pedestrian crashes while 49.86% of them didn't have any bicycle crashes. Total number of pedestrian and the bicycle crashes are two response variables that considered in this study.

4.3 Exogenous Variables Summary

The crash records are collected from Florida Department of Transportation, Crash Analysis Reporting (CAR) and Signal Four Analytics (S4A) databases. Roadway characteristics, traffic characteristics, and socio-demographic characteristics - three broad categories of predictors are considered in our study. The response variables are the total number of pedestrian and bicycle crash in each zone. The data employed are obtained from FDOT Transportation Statistics Division and US Census Bureau. The attributes are then aggregated at the STAZ level using geographical information system (GIS). As discussed earlier, the current analysis considered spatial predictor variables which correspond to characteristics of neighboring STAZs along with the target STAZs. Towards this end, for every STAZ, the STAZs that are adjacent are identified. Based on the identified neighbors, a new variable based on the value of each exogenous variable from surrounding STAZs is computed. The descriptive statistics of the response and predictor variables are summarized in Table 2. Specifically, the table provides the predictor values at a STAZ level as well as for the neighboring STAZs.

Roadway characteristics included are road lengths for different functional class, signalized intersection density, length of bike lanes and sidewalks, etc. Intersection density denotes the number of intersections per street mile in a STAZ. Vehicle-miles-traveled and proportion of heavy vehicles in VMT are considered as traffic characteristics. For demographic characteristics, population density, proportion of families without vehicle, proportion of urban area, no of commuters by public transportation, etc. are considered.

4.4 Summary

In this chapter, data compilation procedures are discussed. Further, descriptive statistics for both dependent and independent variables are provided. The empirical analysis results are summarized in the next chapter.

Table 2 Sample Characteristics of the Road Accidents Attributes

Variables name	Definition	Targeted TAZs			Neighboring TAZs		
		Mean	S.D.	Max[a]	Mean	S.D.	Max[a]
Crash Variables							
Pedestrian crash	Total number of pedestrian crashes per STAZ	1.907	3.315	39.000	-	-	-
Bicycle crash	Total number of pedestrian crashes per STAZ	1.797	3.309	88.000	-	-	-
Traffic & Roadway Variables							
VMT	Total vehicle miles travel in the STAZ	31381.0	41852.3	684742.8	195519.7	169120.3	2103376.3
Proportion of heavy vehicle in VMT	Total heavy vehicle VMT in STAZ /Total vehicles VMT in STAZ	0.067	0.052	0.519	0.070	0.045	0.350
Proportion of length of arterial roads	Total length of arterial road/ Total road length in the STAZ	0.221	0.275	1.000	0.144	0.125	1.000
Proportion of length of collectors	Total length of collector road/ Total road length in the STAZ	0.191	0.246	1.000	0.156	0.136	1.000
Proportion of length local roads	Total length of local road/ Total road length in the STAZ	0.572	0.329	1.000	0.680	0.200	1.000
Signalized intersection density	Number of intersections per mile in each STAZ	0.227	0.578	8.756	0.378	5.552	495.032
Length of bike lanes	Total length of bike lanes in each STAZ	0.303	1.096	28.637	1.909	3.847	38.901
Length of sidewalks	Total length of sidewalk in each STAZ	0.993	1.750	25.683	6.304	6.745	77.720
Socio-Demographic Variables							
Population density	Population density per square mile	2520.3	4043.3	63069.0	2330.2	3489.7	57181.9
Proportion of families without vehicle	Total number of families with no vehicle in STAZ/Total number of families in STAZ	0.095	0.123	1.000	0.095	0.108	1.000
School enrollments density	Total school enrollment per square miles in STAZ	775.02	5983.05	255147.24	684.22	2900.54	102285.73
Proportion of urban area	Total urban area in STAZ/Total area in STAZ	0.722	0.430	1.000	0.650	0.434	1.000
Distance to the nearest urban area	Distance of the STAZ to the nearest urban area	2.140	5.441	44.101	-	-	-
Hotels, motels, and timeshare rooms density	Hotels, motels, and timeshare rooms density per square mile	172.49	941.71	32609.84	121.678	528.078	11397.148

Variables name	Definition	Targeted TAZs			Neighboring TAZs		
		Mean	S.D.	Max[a]	Mean	S.D.	Max[a]
No of total employment	Total employment in STAZ	1140.10	1722.45	31932.15	6917.245	6725.135	76533.000
Proportion of industry employment	Proportion of industry employment	0.176	0.232	1.000	0.183	0.177	1.000
Proportion of commercial employment	Proportion of commercial employment	0.299	0.235	1.000	0.305	0.177	1.000
Proportion of service employment	Proportion of service employment	0.525	0.257	1.000	0.495	0.186	1.000
No of commuters by public transportation	No of commuters using public transportation	18.813	54.273	934.000	119.582	246.299	3559.985
No of commuters by cycling	No of commuters using bicycle	5.894	19.804	775.000	90.869	128.399	1902.135
No of commuters by walking	No of commuters by walking	14.354	34.680	1288.000	37.566	74.484	1634.530

[a] The minimum values for all variables are zero.

CHAPTER 5: MODEL ANALYSIS AND RESULTS

The results for the models described in Chapter 3 are presented in this chapter. Basically, the model estimation process involved estimating four models as follows (1) DTR aspatial model for pedestrian crashes (2) DTR spatial model for pedestrian crashes (3) DTR aspatial model for bicycle crashes (4) DTR spatial model for bicycle crashes. This chapter presents the modelling results with the explanation of significant predictor variables associate with the pedestrian and bicycle crash risk.

5.1 Model Specification and Overall Measure of Fit

In this study, from the 8518 STAZs, 70% of the STAZs were randomly selected as training set for model development while 30% were employed as testing set for model validation. In the first step, the model estimation process involved estimating four models as follows (1) DTR aspatial model for pedestrian crashes (2) DTR spatial model for pedestrian crashes (3) DTR aspatial model for bicycle crashes (4) DTR spatial model for bicycle crashes. Prior to discussing the model results, we compare the estimated models in Table 3. The table presents the Average Squared Error (ASE) and Standard Deviation of Errors (SDE) for the four DTR models with training and testing samples. It is worth mentioning that, a series of trees have been produced in order to achieve the best DTR models for each of the four models mentioned above. The model with the lower ASE and SDE is the preferred DTR model. Across pedestrian and bicycle crash prediction models, the models with spatial predictor variables (spatial model) offer substantially better prediction models in terms of ASE and SDE in both training and testing date sets. Thus, this result highlighted that inclusion of predictor variables of adjacent STAZs improve crash prediction models using data mining techniques (DTR

models) which confirmed the same results obtained using statistical modeling techniques on Cai et al. (Cai et al., 2016).

Table 3 Comparison of Predictability Between Different Models

Pedestrian Crashes		
Training (N=5963)	**Without Spatial Predictor Variables**	**With Spatial Predictor Variables**
No of predictor variable used	10	12
ASE	5.597	5.142
SDE	2.366	2.268
Testing (N=2555)	**Without Spatial Predictor Variables**	**With Spatial Predictor Variables**
No of predictor variable used	10	12
ASE	6.328	6.178
SDE	2.516	2.485
Bicycle Crashes		
Training (N=5963)	**Without Spatial Predictor Variables**	**With Spatial Predictor Variables**
No of predictor variable used	9	12
ASE	5.413	5.092
SDE	2.327	2.257
Testing (N=2555)	**Without Spatial Predictor Variables**	**With Spatial Predictor Variables**
No of predictor variable used	9	12
ASE	6.724	5.926
SDE	2.594	2.435

5.2 DTR Model Estimation and Interpretation

As previously mentioned, DTR partitions the data into relatively homogeneous terminal nodes, and it takes the mean value observed in each node as its predicted value. The empirical analysis involved a series of DTR model estimations in order to achieve the lowest possible ASE and SDE. In presenting the DTR framework, we will restrict ourselves to the discussion of the decision tree regression graphically. The main objective of this paper is to explore the DTR models in order to obtain the important contributing factors (either using spatial predictor variables or not) for pedestrian and bicycle crashes and then substantially improving the

prediction model by applying ensemble techniques to the DTR models. Toward this end, lists of variables are entered into each model and their relative importance were also produced. Variable importance is calculated based on deviance (D) or sum of squared errors (SSE) of each variable which indicates a measure of the dispersion. The first partition of the observations in the DTR models is undertaken based on the most important predictor variable resulting in the maximum reduction in variability of the response variable. Then, further partitions are made based on the hierarchy of most important variables. The importance value of the most important variable is 1. Then all other variables are assigned with a relative importance. The variable importance result of four models (2 model types with and without spatial predictor variables of neighboring STAZs) of pedestrian and bicycle crashes each are displayed in Table 4 and Table 5, separately. Across the four models for either pedestrian or bicycle crashes, the significant importance variable are quite comparable. While the variables with relative importance results for all DTR models across pedestrians and bicycle crashes are presented, the discussion focuses on the DTR model with spatial predictor variables that offers the best model.

5.2.1 DTR models for pedestrian crash

For DTR spatial model, seven predictor variables of targeted STAZs and five predictor variables of neighboring STAZ are found to be most important variables for forecasting pedestrian crash. Five significant predictor variables of neighboring STAZ confirmed the importance of including spatial variables in order to predict the pedestrian crashes at the macro-level. The results of the variable importance for both models (aspatial and spatial) for pedestrian crashes are presented in Table 4. To emphasize the predictor variables, we also ranked each variable based on their variable importance – with 1 as the highest important

variable and 12 as the lowest important variable in spatial model.

Table 4 Variable Importance for Pedestrian Crash of STAZs

Predictor variables	Aspatial	Ranking	Spatial	Ranking
STAZ predictor variables				
Number of commuters by public transportation	1.0000	1	1.0000	1
Number of total employments	0.5236	2	0.5372	2
Signalized intersection density	0.3999	3	0.4191	3
Number of commuters by walking	0.3744	4	0.3673	4
Vehicle miles travelled (VMT)	0.2968	5	0.3405	6
Length of sidewalks	0.2883	6	0.3479	5
Length of bike lanes	0.1359	7	0.1394	9
Distance to nearest urban area	0.0511	8	-	-
Hotels, motels, and timeshare rooms density	0.0459	9	-	-
Proportion of urban area	0.0215	10	-	-
Spatial predictor variable				
Number of commuters by public transportation in neighboring STAZs	-	-	0.3200	7
Number of commuters by walking in neighboring STAZs	-	-	0.1703	8
Population density in neighboring STAZs	-	-	0.1372	10
Proportion of families without vehicle in neighboring STAZs	-	-	0.1304	11
School enrollment density in neighboring density	-	-	0.0530	12

The following observations can be made based on the results presented in Table 4. The most important variable for determining the number of pedestrian crashes in macro-level is number of commuters using public transport with relative importance 1.0. The statistical modelling results intuitively support that commuters by public transportation reflect zones with higher pedestrian activity resulting in increased crash risk (Abdel-Aty et al., 2013). The next most important variable to predict the pedestrian crashes is total employment which is surrogate measures of pedestrian exposure (Siddiqui et al., 2012). Hence, it is expected that total employment has a higher impact on crash frequency. The variables including signalized intersection density, number of walk commuters, length of sidewalks, and length of bike lanes

represent the likelihood of pedestrian access. Therefore, these variables are found to be significant variables in the DTR model. The VMT variable is a measure of vehicle exposure and as expected a significant predictor for pedestrian crashes. It is interesting to note that the variables distance to nearest urban area, hotel, motel, and timeshare room density, and proportion of urban area are significant predictor variables (rank-8,9,10) in DTR aspatial model, while those variables are not found significant variables in DTR spatial models which offers the better fit. Among the significant important spatial predictor variables, the number of commuters by public transportation offers the most important variable to predict pedestrian crashes. Cai et. al., (2014) (Cai et al., 2016) proved that the commuters by public transportation in neighboring STAZ has a positive impact on pedestrian crashes. Moreover, the number of commuters by walking, population density, proportion of families without vehicle, and school enrollment density in neighboring STAZs are significant spatial variables of pedestrian crashes at the macro-level.

5.2.2 DTR models for bicycle crash

In the DTR model with spatial variables presented in Table 5, eight variables of the targeted STAZs and four variables of the neighboring STAZs are responsible for predicting bicycle crash frequency. The impact of some predictor variables in the pedestrian and bicycle crash prediction models are quite similar. A possible reason is that STAZs with high pedestrian activity are also likely to experience high bicyclists activity. Among the parent (targeted) STAZ variables, number of total employments is the most important predictor variable of bicycle crashes. The other important variables for the bicycle crash propensity are vehicle miles travelled (VMT), number of commuters using bicycle, number of commuters by walk, length of sidewalks, number of commuters using public transport, signalized intersection density, and

proportion of urban area, respectively. There are three main differences in the STAZ variable impacts between pedestrian and bicyclists crash frequency in terms of variable importance.

Table 5 Variable Importance for Bicycle Crash of STAZs

Predictor variables	Aspatial	Ranking	Spatial	Ranking
STAZ predictor variables				
Number of total employments	1.0000	1	1.0000	1
Number of commuters using bicycle	0.6684	2	0.5688	4
Number of commuters using public transport	0.6523	3	0.1543	9
Vehicle miles travelled (VMT)	0.4875	4	0.5909	3
Length of bike lanes	0.4403	5	0.2922	7
Proportion of urban area	0.3015	6	0.0369	12
Distance to nearest urban area	0.2040	7	-	-
Signalized intersection density	0.1955	8	0.1387	10
School enrollment density	0.0749	9	-	-
Number of commuters by walk	-	-	0.3254	6
Spatial predictor variable				
Number of commuters using bicycle in neighboring STAZs	-	-	0.6464	2
Population density in neighboring STAZs	-	-	0.5270	5
Length of bike lanes in neighboring STAZs	-	-	0.2037	8
School enrollment density in neighboring STAZs	-	-	0.1327	11

First, the density of hotel, motel and time share rooms is not a significant variable for predicting bicycle crash. This result is intuitive because tourists are less likely to use bicycles. Second, the school enrolment density does have significant impact on bicycle crashes as it is possible that students are more likely use bicycles for traveling to schools. Third, the length of sidewalks in the STAZ does not have significant importance to predict bicycle crashes, whereas, sidewalk length is found to be significant variable for predicting pedestrian crashes. In terms of spatial variables effect, the important variables have mixed effects between pedestrian and bicyclists. Population density and the school enrolment density in neighboring STAZs offers important spatial variables for both pedestrian and bicycle crash prediction

models. Number of commuters using bicycle and the length of bike lanes in neighboring STAZs are found significantly associated with bicycle crashes.

5.2.3 Ensemble Techniques Results

To improve the prediction accuracy of the DTR models, we have used ensemble techniques using three structures: 1) Bagging, 2) Random Forest, 3) Gradient Boosting. Figure 1 illustrates the basic framework of the three ensemble techniques proposed in the pedestrian and bicycle crash prediction models. Some observations can be made from this framework. All the three ensemble techniques combine several decision trees to produce better predictive performance than utilizing a single decision tree. Bagging create several subsets of data by bootstrap resampling while the random forest utilizes the same process in addition to taking the random subset predictors.

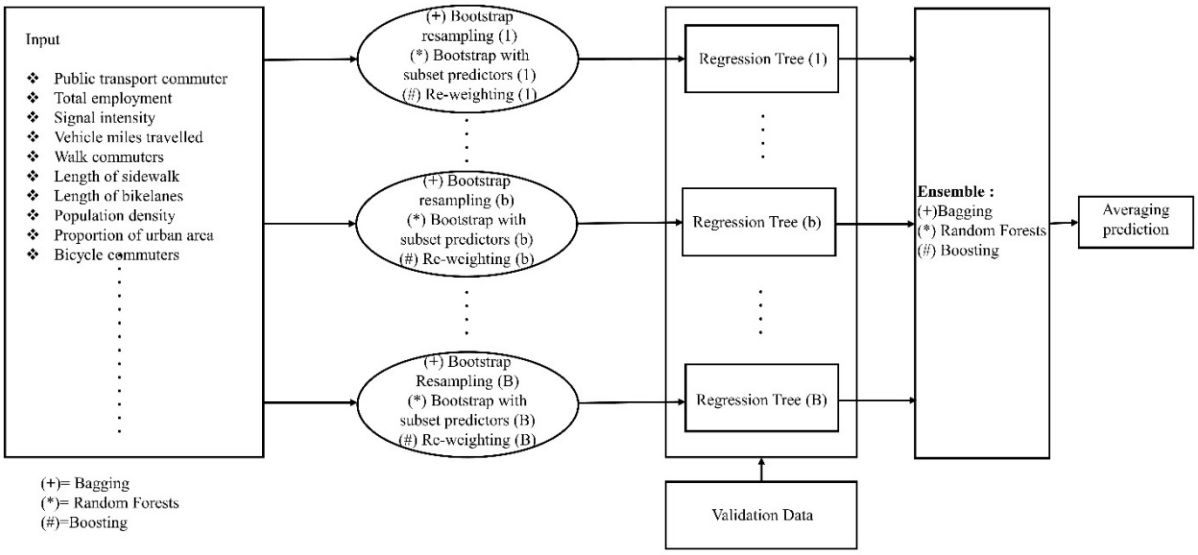

Figure 1 Ensemble technique framework: Bagging, Random Forests, and Boosting.

Unlike bagging and random forest, boosting generate multiple training samples by re-weighting which can improves the accuracy of single learner. Finally, bagging, random forest, and boosting estimate the final prediction by averaging multiple estimates of individual trees.

The aforementioned three ensemble techniques were implemented based on the methodology showed in Figure 1 and the goodness of fit measure such as ASE and SDE are calculated for the spatial models of pedestrian and bicycle crashes. The comparison results of the ensemble techniques along with the DTR models (weak learners) for both pedestrian and bicycle crashes are presented in Table 6. The table presents the ASE and SDE for the ensemble techniques and DTR model for training and testing samples. Three significant conclusions can be made from the results highlighted in Table 6. First, all models with ensemble techniques perform better than the original DTR model. Second, gradient boosting provides the best performance in all ensemble techniques compared to the other counterparts. Third, Random forests is better than bagging in terms of goodness of fit measures.

Table 6 Comparison of Predictability Across Ensemble Techniques

Measure of effectiveness	**Decision Tree**	**Bagging**	**Random Forests**	**Gradient Boosting**
Pedestrian crashes with spatial predictor variables				
Training (N=5963)				
ASE	5.142	5.016	4.975	4.856
SDE	2.268	2.239	2.230	2.203
Testing (N=2555)				
ASE	6.178	6.089	6.015	5.915
SDE	2.485	2.468	2.453	2.432
Bicycle crashes with spatial predictor variables				
Training (N=5963)				
ASE	5.092	4.965	4.912	4.821
SDE	2.257	2.228	2.216	2.196
Testing (N=2555)				
ASE	5.926	5.868	5.821	5.712
SDE	2.435	2.422	2.413	2.390

5.3 Summary

This chapter presented the modelling results of decision tree regression considering with or without spatial predictor variables. Variable importance of the predictor variables provides an indication of policy analysis in the macro-level crash risk. Some ensemble techniques results are also presented to improve the prediction accuracy of the DTR models. To summarize, based on the empirical results, it is clear that the gradient boosting algorithms outperformed competing two ensemble techniques which found the best technique for predicting the pedestrian and bicycle crash in macro-level.

CHAPTER 6: CONCLUSIONS

This study applied data mining techniques for pedestrian and bicycle crash analyses that captures the effects of important predictor variables at the macro-level. The study conducted decision tree regression (DTR) modeling analysis to highlight the importance of various traffic, roadway, and socio-demographic characteristics of the STAZ on the pedestrian and bicycle crash occurrence. To the best of the authors' knowledge, this is the first attempt to employ such DTR models at the macro-level. The study also considered spatial predictor variables from neighboring STAZs in order to improve the prediction accuracy of DTR models for both pedestrian and bicycle crashes. It was found that the introduction of spatial predictor variables on DTR models clearly outperformed the DTR models that did not consider the spatial variables in terms of goodness-of-fit measures. To facilitate a policy analysis at the macro-level, variable importance of DTR models for both pedestrians and bicyclists crashes were computed. The variable importance results clearly highlighted the significant predictor variables of the targeted and neighboring STAZs including traffic (such as VMT), roadway (such as signalized intersection density, length of sidewalks and bike lanes, etc.) and sociodemographic characteristics (such as population density, commuters by public transportation, walking and bicycling) for both pedestrian and bicycle crashes. In terms of the planning perspective, it is important to identify zones with high public transit commuter, employment area, pedestrian and bicyclist commuters and undertake infrastructure upgrades to improve safety. Finally, the study undertook some ensemble techniques such as bagging, random forest, and gradient boosting to improve the prediction accuracy of pedestrian and bicycle crashes. The results revealed that, all the ensemble techniques offer substantially better fit compared to original DTR models. Moreover, Random forests is better than bagging in

terms of goodness of fit measures. Finally, gradient boosting algorithms outperformed competing two ensemble techniques which found the best technique for predicting the pedestrian and bicycle crash in macro-level.

The paper is not without limitations. While the decision tree regression is considered, we do not consider other data mining techniques to check the prediction accuracy. It will be an interesting exercise to model the other data mining techniques such as neural network, support vector machine and their ensembles. Moreover, it might be beneficial to explore the similar models for multiple spatial units and several years.

REFERENCES

Abdel-Aty, M.A., Lee, J., Eluru, N., Tammam, N., Yasmin, S., 2016. Joint Modeling of Pedestrian and Bicycle Crashes: A Copula Based Approach. Transp. Res. Rec. 2601, 119–127.

Abdel-Aty, M., Keller, J., Brady, P., 2005. Analysis of Types of Crashes at Signalized Intersections by Using Complete Crash Data and Tree-Based Regression. Transp. Res. Rec. J. Transp. Res. Board 1908, 37–45.

Abdel-Aty, M., Lee, J., Siddiqui, C., Choi, K., 2013. Geographical unit based analysis in the context of transportation safety planning. Transp. Res. Part A Policy Pract. 49, 62–75.

Blincoe, L., Seay, A., Zaloshnja, E., T.Miller, Romano, E., S.Luchter, R.Spicer, 2002. The economic impact of motor vehicle crashes, 2000. DOT HS, 809, 446.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone., C.J., 1998. Classification and Regression Trees. Chapman & Hall/CRC.

Cai, Q., Abdel-Aty, M., Lee, J., Eluru, N., 2017. Comparative analysis of zonal systems for macro-level crash modeling. J. Safety Res. 61, 157–166.

Cai, Q., Lee, J., Eluru, N., Abdel-Aty, M., 2016. Macro-level pedestrian and bicycle crash analysis: Incorporating spatial spillover effects in dual state count models. Accid. Anal. Prev. 93, 14–22. https://doi.org/10.1016/j.aap.2016.04.018

Chang, L.Y., Chen, W.C., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Safety Res. 36, 365–375.

Chang, L.Y., Chien, J.T., 2013. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Saf. Sci. 51, 17–22.

Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accid. Anal. Prev. 38, 1019–1027.

De Oña, J., López, G., Abellán, J., 2013. Extracting decision rules from police accident reports through decision trees. Accid. Anal. Prev. 50, 1151–1160.

Ekram, A.-A., Rahman, M.S., 2018. Effects of Connected and Autonomous Vehicles on Contraflow Operations for Emergency Evacuation: A Microsimulation Study. Proceeding 97th Annu. Meet. Transp. Res. Board.

Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accid. Anal. Prev. 40, 1033–1054.

Eustace, D., Alqahtani, T., Hovey, P.W., 2018. Classification Tree Modelling of Factors Impacting Severity of Truck-Related Crashes in Ohio. Transp. Res. Board 97th Annu. Meet.

Huang, H., Abdel-Aty, M., Darwiche, A., 2010. County-Level Crash Risk Analysis in Florida Bayesian Spatial Modeling. Transp. Res. Rec. J. Transp. Res. Board 2148, 27–37.

Iragavarapu, V., Lord, D., Fitzpatrick, K., 2015. Analysis of injury severity in pedestrian crashes using classification regression trees. Transp. Res. Board 94th Annu. Meet.

James, G., Witten, D., Hastie, T., & Tibshirani, R., 2013. An introduction to statistical learning. New York springer. 112.

Karlaftis, M.G., Golias, I., 2002. Effects of road geometry and surface on speed and safety. Accid. Anal. Prev. 34.3 34, 357–365.

Kashani, A.T., Mohaymany, A.S., 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Saf. Sci. 49, 1314–1320.

Lee, J., Abdel-Aty, M., Choi, K., Huang, H., 2015. Multi-level hot zone identification for pedestrian safety. Accid. Anal. Prev. 76, 64–73.

Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M., Cai, Q., 2018. Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed fractional split multinomial logit

modeling approach with spatial effects. Accid. Anal. Prev. 111, 12–22.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transp. Res. Part A Policy Pract. 44, 291–305.

Lord, D., Washington, S., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accid. Anal. Prev. 39, 53–57.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. Accid. Anal. Prev. 37, 35–46.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.

Mayr, A., Binder, H., Gefeller, O., Schmid, M., 2014. The evolution of boosting algorithms: From machine learning to statistical modelling. Methods Inf. Med. 53, 419–427.

Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F., 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accid. Anal. Prev. 49, 58–72.

NHTSA, 2017a. 2016 motor vehicle crashes: overview. Traffic Saf. facts Res. note 1–9.

NHTSA, 2017b. Traffic Safety Facts: Pedestrian.

NHTSA, 2015. Traffic Safety Facts: Bicyclists and Other Cyclists.

NHTSA, 2005. Motor Vehicle Traffic Crashes as a Leading Cause of Death in the United States 2002 1, 3.

Pande, A., Abdel-Aty, M., Das, A., 2010. A classification tree based modeling approach for segment related crashes on multilane highways. J. Safety Res. 41, 391–397.

Rahman, M.S., Abdel-Aty, M., 2018. Longitudinal safety evaluation of connected vehicles' platooning on expressways. Accid. Anal. Prev. 117, 381–391. https://doi.org/10.1016/j.aap.2017.12.012

Rahman, M.S., Abdel-Aty, M., Wang, L., Lee, J., 2018. Understanding the Highway Safety

Benefits of Different Approaches of Connected Vehicles in Reduced-Visibility Conditions. Transp. Res. Rec. J. Transp. Res. Board.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. Accid. Anal. Prev. 29, 829–837.

Siddiqui, C., Abdel-Aty, M., Choi, K., 2012. Macroscopic spatial analysis of pedestrian and bicycle crashes. Accid. Anal. Prev. 45, 382–391.

Son, H.D., Kweon, Y.J., Park, B.B., 2011. Development of crash prediction models with individual vehicular data. Transp. Res. Part C Emerg. Technol. 19, 1353–1363.

Song, Y.-Y., Lu, Y., 2015. Decision tree methods: applications for classification and prediction. Shanghai Arch. psychiatry 27, 130–5.

Tavakoli Kashani, A., Rabieyan, R., Besharati, M.M., 2014. A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. J. Safety Res. 51, 93–98.

Wah, Y.B., Nasaruddin, N., Voon, W.S., Lazim, M.A., 2012. Decision Tree Model for Count Data. World Congr. Eng. I, 4–9.

Wang, X., Yuan, J., Schultz, G.G., Fang, S., 2018. Investigating the safety impact of roadway network features of suburban arterials in Shanghai. Accid. Anal. Prev. 113, 137–148. https://doi.org/10.1016/j.aap.2018.01.029

Washington, S., 2000. Iteratively specified tree-based regression: theory and trip generation example. J. Transp. Eng. 126, 482–491.

Washington, S., Wolf, J., 1997. Hierarchical Tree-Based Versus Ordinary Least Squares Linear Regression Models: Theory and Example Applied to Trip Generation. Transp. Res. Rec. 1581, 82–88. https://doi.org/10.3141/1581-11

Yuan, J., Mohamed Abdel-Aty, 2018. Approach-Level Real-Time Crash Risk Analysis for Signalized Intersections. Accid. Anal. Prev.

Zheng, Z., Lu, P., Denver, T., 2016. Accident Prediction for Highway-Rail Grade Crossings using Decision Tree Approach: An Empirical Analysis. Transp. Res. Rec. J. Transp. Res. Board 2545, 115–122.