Accident Analysis and Prevention 45 (2012) 317–325

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Aggregate nonparametric safety analysis of traffic zones

Chowdhury Siddiqui[a,*], Mohamed Abdel-Aty[a], Helai Huang[b]

[a] Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816-2450, United States
[b] School of Traffic and Transportation Engineering, Central South University (Railway Campus), 68 Shao-Shan-Nan Road, Changsha, Hunan 410075, PR China

ARTICLE INFO

Article history:
Received 15 April 2011
Received in revised form 15 June 2011
Accepted 26 July 2011

Keywords:
Macro-level safety analysis
Total crash
Severe crash
Data mining
Random forest
Safety planning

ABSTRACT

Exploring the significant variables related to specific types of crashes is vitally important in the planning stage of a transportation network. This paper aims to identify and examine important variables associated with total crashes and severe crashes per traffic analysis zone (TAZ) in four counties of the state of Florida by applying nonparametric statistical techniques such as data mining and random forest. The intention of investigating these factors in such aggregate level analysis is to incorporate proactive safety measures in transportation planning. Total and severe crashes per TAZ were modeled to provide predictive decision trees. The variables which carried higher weight of importance for total crashes per TAZ were – total number of intersections per TAZ, airport trip productions, light truck productions, and total roadway segment length with 35 mph posted speed limit. The other significant variables identified for total crashes were total roadway length with 15 mph posted speed limit, total roadway length with 65 mph posted speed limit, and non-home based work productions. For severe crashes, total number of intersections per TAZ, light truck productions, total roadway length with 35 mph posted speed limit, and total roadway length with 65 mph posted speed limit were among the significant variables. These variables were further verified and supported by the random forest results.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The recent emphasis on transportation safety planning (TSP) issues has shed light on the importance in dealing with macro-level crash data. Incorporating safety in the transportation planning stage has been challenging from different perspectives. For example, aggregated crash data is criticized of having problems with heterogeneity. Also, planning for an unused land area or even for an existing area requires considerations from socio-economic, demographic and different traffic-related factors. The interplay between these factors becomes a significant problem for prediction during the planning stage.

For the last decade researchers have been trying to capture significant factors associated with crashes at different aggregation levels. In this study we consider traffic analysis zones (TAZs) as the base spatial unit of analysis. That means, crashes were aggregated for a TAZ. TAZs are thought to be homogenous in demographic and land use patterns (You et al., 1997). Also they are widely used building blocks of transportation planning networks and frequently used by metropolitan planning organizations (MPOs) in their travel demand models (TDMs) and various traffic related analysis.

This paper aims to identify and examine important variables associated with total and severe crashes per TAZ. In doing so, different trip related variables and road-traffic related factors were investigated. Very few studies have addressed the effect of trip related variables onto crashes. Trips usually depend on demographic and land use patterns of an area. Thus, trip variables were thought to be advantageous in accounting for different correlated effects between demographic and land use factors. Investigating these factors in such aggregate level analysis will be helpful in incorporating proactive safety measures for long range transportation planning. Data mining techniques such as decision trees and random forests were applied to find important factors and develop predictive crash models. These techniques are nonparametric in nature and do not depend on any functional form. Also, they account for missing values and outliers and can handle a dataset with large number of predictors without being affected by the multicollinearity problem between them. The superiority and robustness of tree based algorithms are extensively discussed in Karlaftis and Golias (2002), Chang and Wang (2006), Harb et al. (2009), and Yan et al. (2010).

2. Review of literature

Macroscopic safety modeling has been addressed so far at different spatial aggregation starting from census block to state level. Noland (2003) analyzed how various road infrastructure

* Corresponding author at: Department of Civil, Environmental and Construction Engineering, University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816-2450, United States. Tel.: +1 407 823 5657; fax: +1 407 823 3315.
E-mail address: kawsar_arefin@knights.ucf.edu (C. Siddiqui).

0001-4575/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2011.07.019

improvements affect traffic-related fatalities and injuries while controlling for other factors like several demographic variables, per capita alcohol consumption, seat-belt legislation, etc. which are known to affect overall safety. County level aggregated data has been analyzed using socio-economic characteristics (Amoros and Laumon, 2003; Noland and Oh, 2004; Aguero-Valverde and Jovanis, 2006), weather conditions and amount of travel (Aguero-Valverde and Jovanis, 2006), and transportation infrastructure (Noland and Oh, 2004; Aguero-Valverde and Jovanis, 2006). Wier et al. (2009) looked at the vehicle–pedestrian injury collisions at 176 San Francisco, California census tracts. The predictor variables examined in their study included street, land use, and population characteristics. Noland and Quddus (2004) analyzed ward level crash data for England using land use types, road characteristics and demographic data. Levine et al. (1995) examined the number of motor vehicle crashes per census block group to population, employment and road characteristics.

This study analyzed total and severe crashes at 1349 TAZs of four counties of the state of Florida with variations in urban and rural landscape. The study considered several road network characteristics and various trip related predictors in the crash analysis.

A wide spectrum of modeling approaches has been incorporated in aggregate level safety research. For example, negative binomial (NB) models have been widely used in TAZ (Hadayeghi et al., 2003, 2006; Guevara et al., 2004), ward (Noland and Quddus, 2004), and county (Karlaftis and Tarko, 1998; Amoros and Laumon, 2003; Noland and Oh, 2004; Aguero-Valverde and Jovanis, 2006) level aggregate crash analyses. Wier et al. (2009) used ordinary least square regression to predict vehicle–pedestrian injury collision for the census tracts. Other methodologies found in the literature include log linear models (Washington et al., 2006) and geographically weighted regression (GWR) models (Hadayeghi et al., 2003) for TAZs, spatial lag models for block groups (Levine et al., 1995), and more recently Bayesian hierarchical models for the wards (Quddus, 2008) and Bayesian spatial models for the counties (Huang et al., 2010). Quddus (2008) found that in many cases the results from the non-spatial NB models were similar to the Bayesian hierarchical models. Also, Aguero-Valverde and Jovanis (2006) found no evidence of spatial correlation in fatal crashes for the counties of Pennsylvania and their results concerning the effects of covariates on fatal and injury crash risk were mostly consistent in direction and magnitude for both NB and Full Bayes models.

The application of classification and regression tree (CART) in road safety analysis has been advocated due to its nonparametric nature and for providing strong, yet simple, variable exploration and predictive ability. CART has the advantage of handling a relatively large number of independent predictors. Also, being a nonparametric tool no functional form is required in modeling with CART and has both theoretical and applied ascendancy over multiple linear and NB models in analyzing crash rates (Karlaftis and Golias, 2002).

The safety literature reflects several examples of CART application in analyzing crashes. For example, Stewart (1996) applied CART models to analyze motor vehicle crash data to estimate measures of driver injury severity when the crash consisted of a vehicle striking a fixed object on the roadside. Stewart (1996) concluded that the tree structure has the advantage of determining complex interaction among the variables easily; however, if the target variable can be approximated accurately by a known functional form of the independent predictors, standard regression procedure would be expected to perform better than CART. It is well known that crash pattern often violates underlying assumptions made by the traditional/classical regression functions. To compensate for multicollinearity between variables, missing observations, and the fact that the true model form was unknown for different factors affecting crashes, by type of collision, at signalized intersections Abdel-Aty et al. (2005) used tree based regressions. Chang and Chen (2005) analyzed two years of crash data from national freeway in Taiwan using CART and traditional negative binomial models using different geometric, traffic characteristics, and environmental factors. Chang and Chen (2005) study demonstrated CART to be a good alternative in analyzing freeway crash frequencies compared with negative binomial regression models based on the prediction performances. More recently Yan et al. (2010) used hierarchical tree-based regression (HTBR) to explore train–vehicle crashes occurring at highway–rail grade crossings. The authors developed several predictive crash models and assessed the effectiveness of the stop-sign treatment utilizing HTBR. Application of nonparametric regression trees in the other safety literature include-predictive modeling of rural crash frequency by Karlaftis and Golias (2002), analyzing rear end crashes by Yan and Radwan (2006), classification of intersection crashes by Qin and Han (2008), and exploring factors associated with crash avoidance maneuvers by Harb et al. (2009).

This paper investigated important variable rankings along with the predictive trees. Random Forest technique was incorporated in this regard to examine and rank important variables. Washington (2000) commented on tree based regression as a valuable data exploration technique for understanding the data structure or relationships among variables. Harb et al. (2009) used similar technique to classify drivers with evasive action and drivers with no evasive actions in their study exploring several driver, vehicle, and environmental characteristics associated with crash avoidance maneuvers. Abdel-Aty et al. (2005) extensively used tree based variable importance ranking to identify important factors related with different collision types occurring at signalized intersections. Also, Abdel-Aty et al. (2008) argued that random forest is more robust variable selection tool as it exhausts a collection of multiple tree classifiers as compared to one single decision tree.

The objectives of this paper are two-fold. The paper investigates important variables associated with total and severe crashes per traffic analysis zone (TAZ). It also illustrates the application of data mining techniques such as decision trees and random forests to find these important factors. Application of nonparametric methods (such as decision trees) is relatively new in safety modeling. Therefore, we have emphasized its robustness as compared to classical/traditional regression models.

As discussed above, several previous studies looked into the aggregated crash data and several regression techniques were applied to predict crashes and to capture significant variables associated with aggregate crashes. However, application of data mining techniques to investigate important variables and the association between the predictors and target at aggregated levels has not been performed. In this study, this has been incorporated through application of CART and random forest.

The rest of the paper is structured as follows. The following section presents the traffic analysis zones and the data preparation steps undergone for this study. The next sections provide a brief overview of classification and regression trees and random forest techniques. The sections thereafter present modeling and results of the decision tree analyses, and explain prediction ability of the tree models. In the following section variable importance rankings from random forest are presented. Finally the paper ends with a discussion summarizing the results, limitation of the study and the future scope.

3. Data description

Traffic analysis zones (TAZs) are special areas delineated by state and/or local transportation officials for tabulating traffic-related

data and are defined as part of the Census Transportation Planning Package (US Census Bureau, web link). A TAZ is a spatial aggregation of census blocks and its size is in part a function of population (Peters and MacDonald, 2004). As cited by You et al. (1997) the most important criteria used to define TAZs include spatial contiguity, homogeneity, compactness, etc. Also TAZs have commonly been considered as a basis for the aggregate modeling process (Miller and Shaw, 2001). The study in this paper was based on 1349 TAZs of four counties in the state of Florida. The counties considered were Hillsborough, Citrus, Pasco, and Hernando. Crashes which occurred during the years 2005 and 2006 in these four counties were used for the analysis in the study. The Geographic Information System (GIS) maps with crash locations were collected from Florida Department of Transportation (FDOT). Each crash in the GIS map was geocoded and provided with several attributes. The roadway characteristics were identified from separate GIS shape files provided by FDOT. These GIS maps represented actual locations of crashes and lengths of roadways.

The TAZ cartographic boundary maps of the study counties and the number of trip attraction and production per day per TAZ for thirteen different categories were collected from FDOT District Seven Intermodal Systems Development Unit. The latest trip data based on 'taz2004' data structure (generated from Tampa Bay Regional Planning Model version 6.0) was used in the study.

According to the United States Department of Transportation (USDOT), planners use home as a base to predict travel patterns since typically household characteristics related to travel are easier to identify and forecast. The conventional meaning of a trip origin and destination is not valid for trip productions and attractions. The following are a few rules suggested by the USDOT (National Transportation Laboratory):

- *Rule 1*: If a trip begins or ends at the traveler's home, the trip is produced at the home end.
- *Rule 2*: If a trip begins or ends at the traveler's home, the trip is attracted to the non-home end.
- *Rule 3*: If a trip begins at a non-home location and ends at a non-home location, the trip is produced at the origin and attracted to the destination.

Classifying trips by trip purposes is typical in trip-based travel demand models adopted by different agencies (such as, MPOs). This enables the simulation of trip behaviors more accurately. For example, a social-recreational trip is more likely to have higher number of occupants in a vehicle as compared to a work trip. Home based work, home based non-work and non-home based work attractions/productions are usually personal trips. Trips in the other categories portend vehicle trips. A home based work trip is made between traveler's home and his/her workplace. A home based non-work trip (may be for various purposes, e.g., shop, school, social recreational, etc.) is made between home and any other destination except workplace. A non-home based trip has neither ends at home, regardless of purpose. A trip with one end inside the study area and the other end outside is defined as an external–internal trip. Also a trip may be categorized by the vehicular usage (taxi, light truck, heavy truck, etc.) and specific purposes (such as college, airport, etc.)

All variables were aggregated at the TAZ level using ArcMap™ version 9.2. The complete list and descriptive statistics of different responses and predictors are provided in Table 1.

4. Classification and regression tree (CART)

Classification and regression trees (CART) are mainly used for variable selection, variable importance, interaction detection, stratified modeling, missing value imputation, model interpretation, predictive modeling, etc. This study mainly focuses on the variable selection, variable importance and predictive modeling for total and severe crashes occurring at TAZs. As mentioned previously CART is a nonparametric method and invariant to monotonic transformation of its independent predictors.

The CART analysis consisted of growing the large initial tree, calculating tree pruning criteria such as AIC, BIC, and deviance values for different tree sizes, and finally pruning the tree to determine the best tree structure. Initially the tree was grown by binary recursive partitioning by dividing the variables into $Y < a$ and $Y \geq a$, where Y is any variable that belongs to roadway or trip characteristics, and a is the splitting value for that variable. The objective of such kind of splitting was to reduce impurity in the child nodes. An optimal splitting rule was being used by the CART procedure where the reduction in impurity in the child nodes was maximized. Gini criterion and misclassification rates were used in this study as the potential node impurity measures. As the initial large tree grew, node impurity of offspring from every node compared to the parent node was minimized. The measure of the node impurity for any node t using Gini criterion follows:

$$i(t) = \sum_{j \neq i} p(i|t)p(j|t) \tag{1}$$

is defined where $i(t)$ is the node impurity of node t, and $p(i|t)$ and $p(j|t)$ are the fraction of cases in node t for which the splitting variable falls into category i or j. For a tree with N terminal nodes misclassification rate is given by,

$$\text{Misclassification rate} = \sum_{t=1}^{N} p(t)[1 - p(j|t)] \tag{2}$$

where, $p(t)$ is the proportion of the data belonging to node t.

The impurities were calculated using *learning* dataset. The *learning* dataset constituted 67% of the total sample size. As the size of the initial tree increases, the nodal impurity decreases. The optimal splitting rules developed for the initial tree was then used in the *test* dataset to calculate model performance. Finally, the initial large tree was snipped based on the selected pruning criterion. CART process exhausts the most important variables to construct the final tree complying with the pruning criterion. Readers who are interested in deeper understanding of the CART procedure would be referred to Chang and Chen (2005) and Breiman et al. (1984) as they provide an extensive description of the tree algorithm.

The CART analyses were also performed using 10-fold cross validation (CV) technique. In general, a k-fold CV is a relatively more efficient remedy than the traditional holdout data (i.e., partitioning dataset into *learning* and *testing*). Performance measures of the trees are averaged over k models. Each model is then fit with $(k-1)/k$ of the data and assessed on the remaining $1/k$ of the data. Then the average over the k holdout datasets is used to estimate the performance of the model fitted to the full dataset.

5. Random forest

Random forest is a relatively new tool in exploring the importance ranking for the variables. Like other data mining tools random forest can handle large number of variables for a large dataset. Random forest technique exploits bootstrap sampling to grow a collection of tree classifiers. The results from these trees are then used to assess the importance of the variables used in the tree modeling. Along with bootstrap sampling random forest also applies Gini impurity criterion so that the Gini impurity for the child nodes is less than the parent nodes. Gini decrease for each individual variable are added up for all trees in the forest and used to gen-

Table 1
Variable description.

Variable name	Definition	*N*	Mean	Standard deviation	Min	Max
TAZ2004	Traffic analysis zone (as per year 2004)	1349	–	–	–	–
Response variables (per TAZ in years 2005–2006)						
log_totcr	Natural logarithm of total number of crashes	1349	3.41	1.269	0	6.18
log_sevcr	Natural logarithm of total number of fatal and severe injury crashes	1349	1.47	0.925	0	3.87
Independent variables related to roadway characteristics (total roadway segment length within a TAZ with)						
seglen15	15 mph posted speed limit	1349	0.22	0.434	0	4.063
seglen25	25 mph posted speed limit	1349	8.27	11.729	0	244.595
seglen35	35 mph posted speed limit	1349	1.40	1.749	0	24.934
seglen45	45 mph posted speed limit	1349	0.11	0.422	0	6.709
seglen55	55 mph posted speed limit	1349	0.13	0.621	0	11.248
seglen65	65 mph posted speed limit	1349	0.22	0.711	0	10.221
SUM_SEG_LEN	All roads	1349	10.76	14.050	0	265
Intersection	Total number of intersections per TAZ	1349	12.32	12.055	1	119
Independent variables related to various trip productions and attractions						
HBWP	Home base work productions	1349	864.44	940.174	0	8056
HBWA	Home base work attractions	1349	852.97	1262.400	0	17788
HBSHP	Home base shop productions	1349	889.98	929.371	0	7363
HBSHA	Home base shop attractions	1349	851.82	1402.270	0	15842
HBSRP	Home base social recreational productions	1349	422.59	436.225	0	3173
HBSRA	Home base social recreational attractions	1349	400.12	649.275	0	8127
HBSCP	Home base school productions	1349	247.14	286.247	0	2965
HBSCA	Home base school attractions	1349	246.75	684.592	0	6832
HBOP	Home base other productions	1349	587.35	614.641	0	4533
HBOA	Home base other attractions	1349	556.74	795.465	0	7992
NHBWP	Non-home base work productions	1349	215.03	299.583	0	3606
NHBWA	Non-home base work attractions	1349	215.03	299.583	0	3606
NHBOP	Non-home base other productions	1349	575.41	860.352	0	10144
NHBOA	Non-home base other attractions	1349	575.41	860.352	0	10144
LTRKP	Light truck productions	1349	268.62	231.951	0	2264
LTRKA	Light truck attractions	1349	268.62	231.951	0	2264
HTRKP	Heavy truck productions	1349	68.76	102.811	0	1591
HTRKA	Heavy truck attractions	1349	68.76	102.811	0	1591
TAXIP	Taxi productions	1349	20.26	21.849	0	323
TAXIA	Taxi attractions	1349	20.26	21.849	0	323
EIP	External–internal productions	1349	0	0	0	0
EIA	External–internal attractions	1349	40.21	56.448	0	647
AIRPP	Airport productions	1349	12.58	34.145	0	540
AIRPA	Airport attractions	1349	0	0	0	0
COLP	College productions	1349	79.64	161.303	0	3234
COLA	College attractions	1349	38.40	341.085	0	5069

erate variable importance measure (Breiman and Cutler, web link). Abdel-Aty et al. (2008) and Harb et al. (2009) provide more details on how random forest important variable ranking algorithm works.

Statistical computing software R (The R Project) and SAS® were used in this study to perform various data mining analyses.

6. Models and results

Regression trees were developed for predicting total crashes and severe crashes occurring at the study TAZs. Natural logarithmic transformations of the response variables were considered to stabilize variance and to minimize average square error. Two regression trees were modeled and their results were compared. The first tree was developed based on the *learning* and *test* samples. The second tree was based on the 10-fold cross validation (CV) technique which used the complete dataset. Table 2 provides the summary statistics of the regression trees. It can be noted that regression trees developed by 10-fold CV provided better goodness-of-fit (i.e., lower residual mean deviance and average square error). Therefore these trees (in Figs. 2 and 3) were used in predictive interpretation of crashes and are discussed in the following sub-sections and next sections.

Please note that in order to compare regression trees with the parametric models, negative binomial (NB) models were developed for the same predictor sets. It was found that the mean/average square error for the NB model developed for 'log of total crash' was 1.116 which is clearly higher than that of both the trees developed for the similar response variable. Average square error of the NB model for 'log of severe crash' was found to be 0.649 which is again higher than that of the regression trees developed considering 'log

Table 2
Goodness-of-fit statistics for different pruned trees.

Tree types	Residual mean deviance	Average square error	No. of variables used	No. of terminal nodes
Log of total crash as response				
Regression tree-1 (test sample)	0.9165	0.911	4	5
Regression tree-2 (CV)	0.7419	0.734	7	11
Log of severe crash as response				
Regression tree-1 (test sample)	0.5855	0.581	5	7
Regression tree-2 (CV)	0.551	0.547	6	9

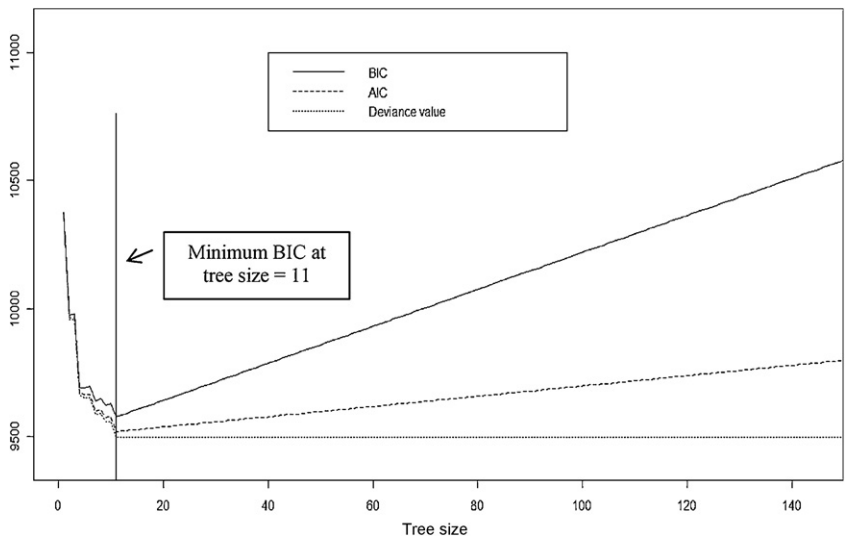

Fig. 1. Tree size selection for regression tree-2 (log of total crash as response) developed using 10-fold CV technique.

of severe crash' as the response. Since a smaller value is desirable for this error, the regression trees developed using cross-validation (CV) technique provided the best fit.

6.1. Modeling total crashes

The initial large tree provided one hundred and forty-three terminal nodes based on 24 variables selected from the batch of all input variables. Fig. 1 shows the Bayesian Information Criterion (BIC), Alkaike Information Criterion (AIC), and deviance values for different tree sizes developed for modeling log of total crashes. The large tree was pruned based on the least BIC value. It can be observed that the trends of all three Goodness-of-fit statistics follow similar pattern. Although BIC is chosen for tree size selection, minimum AIC and deviance values are also achieved by the same number of terminal leaves. This was also found for severe crashes (discussed in the following section). The pruned tree (Fig. 2) produced eleven terminal nodes using seven significant variables which were – total number of intersections per TAZ, light truck productions, total roadway length with 15 mph posted speed limit (PSL), with 35 mph PSL, with 65 mph PSL, airport productions, and non-home based work productions.

6.2. Modeling severe crashes

Similar to the total crash model, a regression tree was developed using 10-fold CV technique and pruned using minimum BIC criterion. The pruned tree (Fig. 3) had nine terminal nodes with six variables. These variables were – total number of intersections per TAZ, light truck productions, total roadway length with 15 mph

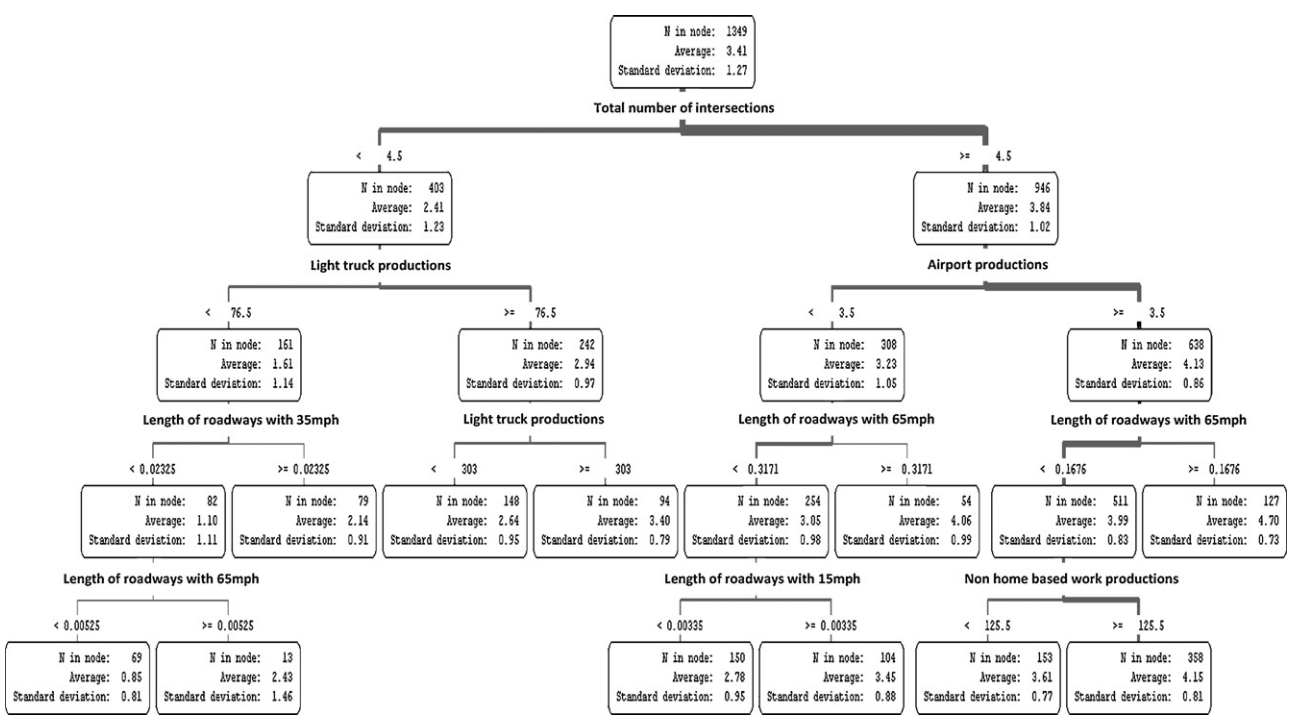

Fig. 2. Regression tree for total crashes using 10-fold CV technique.

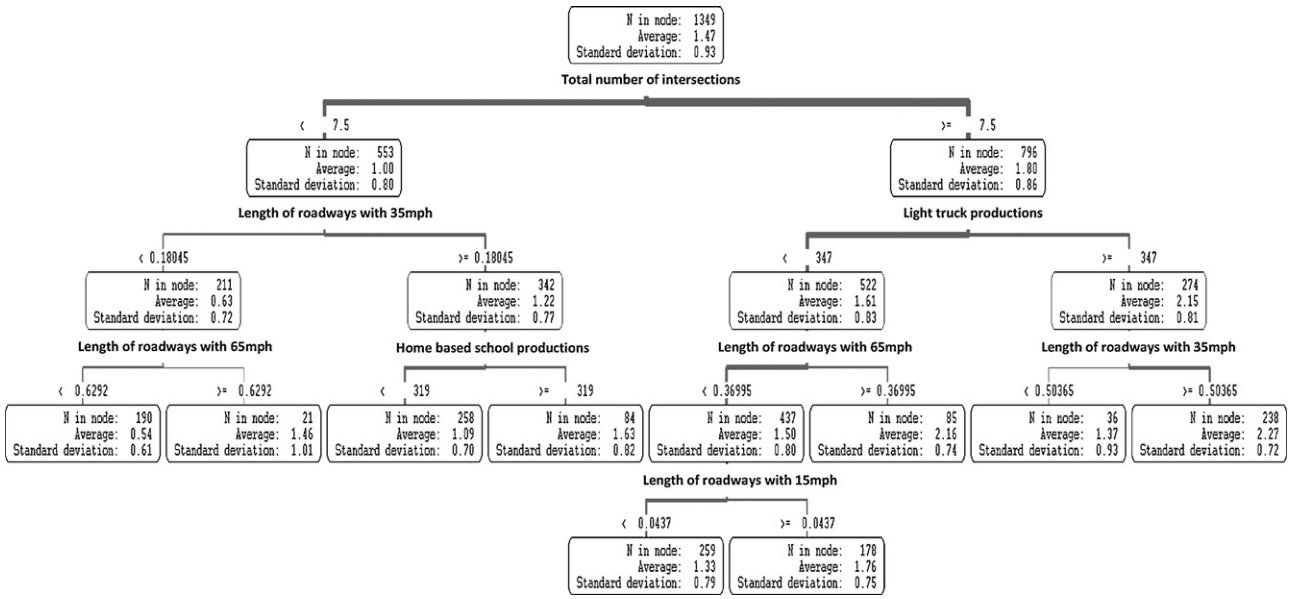

Fig. 3. Regression tree for severe crashes using 10-fold CV technique.

PSL, with 35 mph PSL, with 65 mph PSL, and home-based school productions. It was interesting to note that all variables used in regression tree-1 were also used for regression tree-2. The residual mean deviance for regression tree-2 was found to be 0.551 which was less than that of tree-1.

7. Crash prediction models using trees

7.1. Regression tree model for total crashes

The final tree structure for total crashes used seven significant variables and had eleven terminal nodes. As shown in Fig. 2, total number of intersections per TAZ was used to create the first split of the data. This classified total crashes into two groups: if the number of intersections was less than 4.5 per TAZ, the tree predicted 11 crashes/TAZ ($e^{2.41}$) and for TAZs with total number of intersections ≥4.5, the tree predicted 46.5 crashes/TAZ ($e^{3.84}$). Being used as the first splitting variable, indeed, total number of intersections per TAZ is considered to be the most important variable by the decision tree. The complexity of intersections itself has demanded separate research in the safety studies. Therefore, it is expected that the number of total crashes is dependent on the increase in the number of intersections within a TAZ.

Light truck productions (LTRKP) and airport productions (AIRPP) were used in the second level split. For the TAZs with less number of intersections (<4.5) and LTRKP ≥ 76.5 trips, the tree predicted 19 crashes ($e^{2.94}$). On contrary, TAZs with more number of intersections (≥4.5) and higher number of AIRPP (≥3.5 trips/TAZ), the tree predicted 62.2 total crashes/TAZ ($e^{4.13}$).

The third level of the tree splitting occurred with three variables which were total roadway length with 35 mph PSL (seglen35), light truck productions (LTRKP) and total roadway length with 65 mph PSL (seglen65). Following the branch which had less number of intersections (<4.5/TAZ) but higher light truck production trips (≥76.5 trips/TAZ at second level and then ≥303 trips/TAZ at third level), the tree predicted an average of 30 total crashes/TAZ. On the other side, if a TAZ had total intersections ≥4.5, less airport production trips (<3.5 trips/TAZ), and total roadway length with 65 mph PSL ≥ 0.3171 miles, the tree predicted 58 total crashes for that TAZ.

The branch with more intersections (≥4.5/TAZ) and more airport production trips (≥3.5 trips/TAZ) had the highest propensity of total crashes. As the split indicates- if total length of roadway with 65 mph PSL is less than 0.1676 miles/TAZ, the tree predicts 54 total crashes/TAZ ($e^{3.99}$); and 110 total crashes/TAZ ($e^{4.70}$), otherwise. Therefore, from the tree structure, it can be noted that the variables 'light truck productions' and 'total roadway length with 35 mph PSL' so far were associated with the occurrence of relatively less number of total crashes. On the other side, the variables 'airport productions' and 'total roadway length with 65 mph PSL' were associated with TAZs which were predicted to experience a higher number of total crashes. Hasty attitude to reach the destination may influence airport productions to be at risk. Also airport trips could be related to non-familiar travelers (rental cars). Among all the vehicles that are involved in crashes between 2005 and 2006, automobiles and light trucks constituted about 81% with light trucks having its own share of approximately 21%. Higher involvement in crashes justifies the significance of 'light truck productions' as an important variable.

In the fourth and final level of split, non-home based work production (NHBWP) type trips were associated with highest number of total crashes. Following the splitting rule: if for any particular TAZ, the total number of intersections is ≥4.5, airport productions ≥3.5 trips, and total roadway length with 65 mph PSL is <0.1676 miles, the tree predicted 37 total crashes for having NHBWP <125.5 trips; and 63.4 total crashes, otherwise. Naderan and Shahi (2010) found non-home based work productions as a significant predictor in every crash generation models they developed for total, severe, injury and PDO crashes. In agreement with these authors (Naderan and Shahi, 2010) it is thought that since these trips mostly occur during rush hours and on a periodic basis, they bear more exposure to risk.

7.2. Regression tree model for severe crashes

The best regression tree for severe crashes per TAZ (Fig. 3) exhausted six variables to produce nine terminal nodes. Similar to the total crashes number of intersections per TAZ was identified as the first splitting variable, thereby considering as the most important predictor associated with the severe crashes. It may be

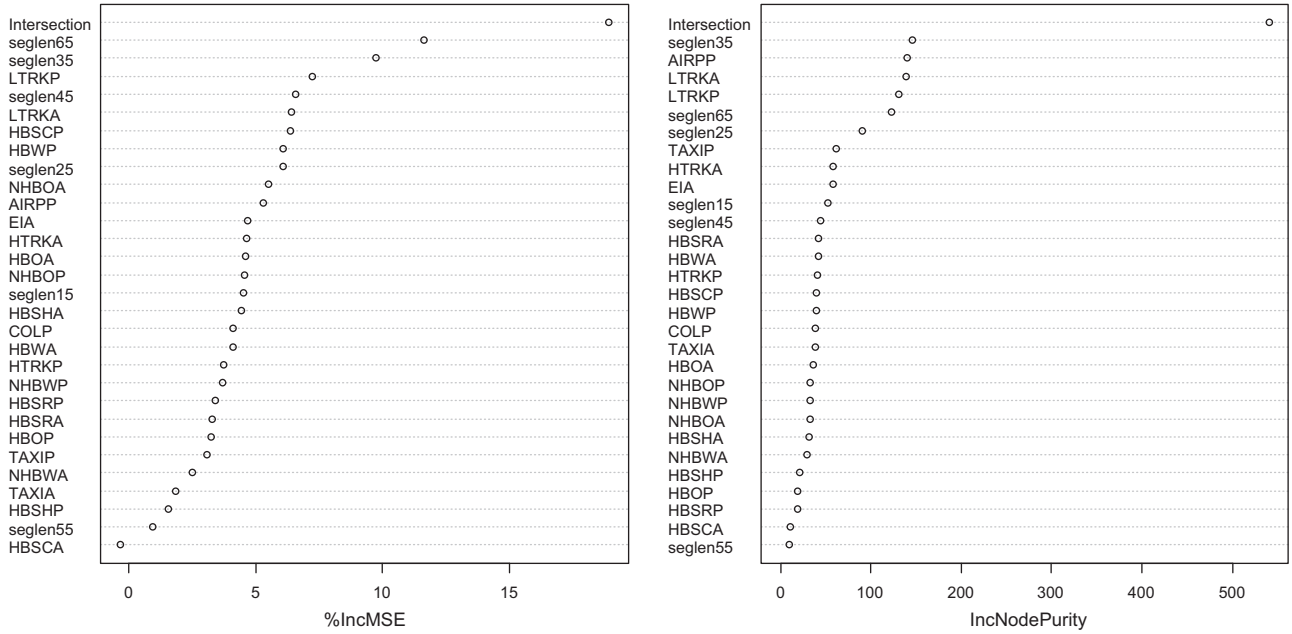

Fig. 4. Variable importance ranking for regression tree developed for total crashes.

observed that, compared to the first level of split for total crashes, the splitting rule associated with the total number of intersections and severe crashes/TAZ shifted to the right (4.5 versus 7.5). This indicates to the fact that a high density of intersection in a TAZ will propagate the risk for severe crashes.

At the second level of the tree, for TAZs with less number of intersections (<7.5) but higher (≥0.18045 miles) roadway length with 35 mph PSL (seglen35), the tree predicted 3.4 severe crashes ($e^{1.22}$). On contrary, for TAZs with more intersections (≥7.5) the split was performed using light truck productions (LTRKP) in such a way that if LTRKP was ≥347 trips, the tree predicted 8.6 severe crashes ($e^{2.15}$); and 5 severe crashes ($e^{1.61}$), otherwise.

The four subgroups created in second level split were further split using three different variables. Variable 'total roadway length with 65 mph PSL' was used twice in the third level splitting rules. For TAZs with less number of intersections (<7.5), higher roadway lengths with 35 mph PSL (≥0.38045 miles), and higher (≥319) number of home-based school production (HBSCP) type trips, the tree predicted 5.1 severe crashes ($e^{1.63}$). For TAZs with higher intersections (≥7.5), lower number (<347) of light truck productions (LTRKP), and higher length (≥0.36995 miles) of roadways with 65 mph PSL, the tree predicted 8.7 severe crashes ($e^{2.16}$). However, following LTRKP ≥ 347 trips/TAZ, for the higher length (≥0.50365 miles) of roadways with 35 mph PSL, the tree predicted 9.7 severe crashes/TAZ ($e^{2.27}$) which is the highest number of predicted severe crashes among all the terminal nodes.

Roadways with 65 mph PSL represent roads with higher speed limit and higher speeds tend to worsen crash severity (Kloeden et al., 2002; Elvik et al., 2004). The length of roadways with 35 mph speed limit was the second highest (13.83%) after roadways with 25 mph posted speed limit. Tignor and Warren (1990) in the study on driver speed behavior found that driver compliance with speed limits is poor and on average, 7 out of 10 motorists exceeded the posted speed in urban areas. Tignor and Warren (1990) also reported that the compliance tends to be worse on low-speed roads. Possibly, a road that is not designed for higher speed (e.g., roads with 35 mph PSL) is riskier for the travelers who tend to drive above the posted speed limit. Consequently, this driving behavior increases the probability of a crash to be more severe. Educational trip production was found to be a significant variable in severe crash and injury crash generation models developed by Naderan and Shahi (2010) for the 380 TAZs of Iran. Finding Home-based school productions as an important splitting variable comply with their findings.

8. Variable importance ranking using random forest

The variables which were used in the pruned tree indicate their relatively higher level of importance among the variables used in this study. To further verify the relative importance of these variables random forest technique was adopted. The R package 'randomForest' was used in analyzing and displaying the variable importance plots. The variable importance ranking (in Figs. 4 and 5, right panel) was calculated using the total decrease in node impurities from variable splits and then average over all the trees. Also, the node impurity was measured using residual sum of squares (Figs. 4 and 5, left panel).

As shown in Fig. 4, total number of intersections per TAZ was top ranked for both cases. Total roadway segment length with 65 mph and 35 mph posted speed limit were ranked second and third, respectively, based on the percentage mean square error. These two variables along with airport productions and light truck productions were within the top one-third important variables according to the node purity. Light truck production was ranked fourth by the mean square error whereas airport productions were ranked eleventh by the same. All four variables selected in regression tree-1 were used in regression tree-2; therefore indicating higher importance for those variables. These variables were – number of intersections per TAZ, airport productions, light truck productions, and roadway length with 65 mph posted speed limit.

Variable importance rankings for the regression tree built for severe crashes are shown in Fig. 5. According to the increased node purity total number of intersections per TAZ and total roadway segment length with 35 mph posted speed limit were ranked first and second most important variables, respectively. Among the other variables which were used in the trees, total roadway segment length with 65 mph posted speed limit, light truck productions, and home-based school productions were within the top one-third of

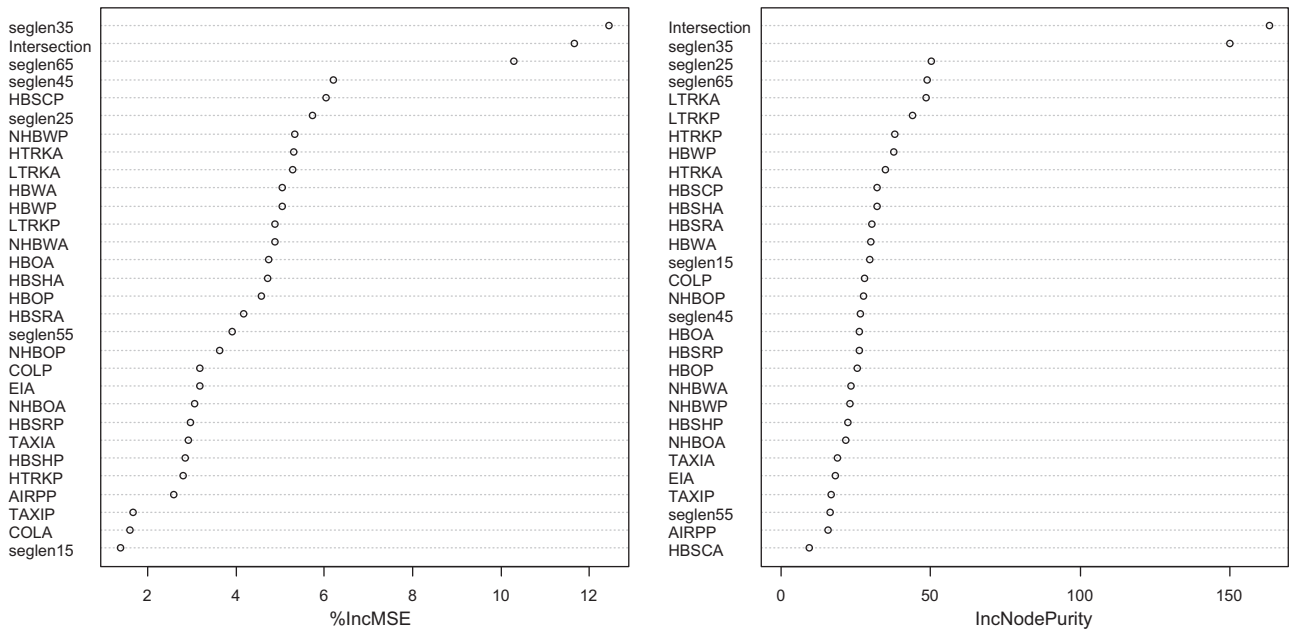

Fig. 5. Variable importance ranking for regression tree developed for severe crashes.

the most important variables ranked in terms of both residual sum of squares and increased node purity.

9. Conclusions

The regression trees developed in this study provided simple 'if-then' relationship of the significant variables affecting total and severe crashes per TAZ. Macro-level crash analysis had been undergone in several studies during the past few years to capture important variables associated with different crash environment and types. However, the challenge for this kind of work sustains in the sense that there are different sets of predictors that may interplay with crashes or among themselves in numerous ways. Several researchers used demographic and socio-economic characteristics for different spatial aggregation. Using trip variables as the predictors for crashes is thought to be privileged as to the fact that by nature of their generation characteristics, trips are function of different demographic and socio-economic factors and are also affected by the land use pattern of an area. Therefore, trips themselves take account into multicollinearity between several factors. Trips ultimately are directly related to crashes and exposure. On the contrary, the cumulative effect of these sets of predictors may to some extent make the interpretability of the trip variables difficult while predicting aggregated crashes (frequency or rates). However, this outcome is not surprising since aggregating crashes for a spatial unit is basically accumulating multiple types of crashes for that geographic boundary. And multiple types of crashes are the phenomena resulting from the complex interactions of those crash specific attributes.

As seen in the regression trees, and also from Figs. 4 and 5, it is evident that few variables were associated with both total and severe crashes. Five variables in particular were present in the top six of the variable rankings in terms of increasing node 'purity' for both total and severe crashes. These variables were the total number of intersections, total roadway length with 35 mph PSL, total roadway length with 65 mph PSL, light truck productions and attractions. Except light truck attractions, the other four variables were used as splitting variables in the regression trees (in Figs. 2 and 3). Therefore, both trees and random forest outputs indicate that the effect of these variables should be considered while developing a strategy for improving the safety of a zone. For example, a TAZ with higher number of intersections can be prioritized for allocating funds for safety treatment, if necessary. TAZs with higher lengths of roadways with 35 mph posted speed limit may be scrutinized carefully by transportation officials to reduce severe crashes as well as total crashes. Alternatives such as installing speed calming devices or lowering the speed limit may be additionally taken into account to improve safety. Similarly, increasing number of signage and driving directions may be considered for unfamiliar drivers in a TAZ with greater amount of airport production. It appears that some of the variables are contributing to both total and severe crashes. However, the trees also indicated the unique predictors related to these two crash response variables. For example, Airport productions were uniquely associated with total crashes per TAZ and were also identified as one of the top important variables according to the random forest rankings. Home based school productions on the other hand appeared to be only associated with severe crashes per TAZ.

Aggregate level analyses have been criticized due to the problem of heterogeneity and multicollinearity. TAZs are relatively homogeneous geographic entities as compared to other spatial units (such as, census blocks, and tracts), they are thought to compensate for the heterogeneity problem. Also, the CART procedure is not affected by the collinearity among the predictors. However, one of the limitations of the study is disregarding the spatial effect which might play a role in the neighboring TAZs. Considering spatial effect in nonparametric crash modeling e.g., data mining, artificial neural network, and hierarchical Bayesian analysis would definitely be an important extension to this analysis.

The authors believe that examining significant variables related to specific types of crashes are vitally important in the planning stage of a transportation network. The need to predict future safety at the planning level has been reflected in NCHRP 8-44(2) project (Washington et al., 2010) which develops a planning-level decision support tool by addressing socio-demographic changes. It is wise to act proactively before crashes would occur in the roadway network.

The long range implication of this work and other along the same line leads to a new direction where safety considerations could be successfully integrated into the planning level. The variables found significant in this study provide a guideline toward the factors that

could be assessed critically in evaluating macro level safety. These variables may be considered as critical factors in a newly built area and/or transportation network where transportation planners can address and insert appropriate safety measures depending on the intensity of these factors.

Acknowledgements

The authors would like to thank the Federal Highway Administration's (FHWA) program for Transformational Changes and Revolutionary Advances for Transportation Planning, and Metro-Plan Orlando for funding this research.

References

Abdel-Aty, M., Keller, J., Brady, P.A., 2005. Analysis of types of crashes at signalized intersections by using complete crash data and tree-based regression. Transportation Research Record 1908, 37–45.
Abdel-Aty, M., Pande, A., Das, A., Knibbe, W.J., 2008. Assessing safety on Dutch freeway with data from infrastructure-based intelligent transportation systems. Transportation Research Record 2083, 153–161.
Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis and Prevention 38, 618–625.
Amoros, E., Laumon, L.M., 2003. Comparison of road crashes incidence and severity between some French counties. Accident Analysis and Prevention 35, 537–547.
ArcMap, Version 9.2, Environmental Systems Research Institute (ESRI) Inc.
Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA.
Breiman, L., Cutler, A., 2010. Random forests. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (accessed 31.03.10).
Chang, L.-Y., Chen, W.-C., 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research 36, 365–375.
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, 1019–1027.
Elvik, R., Christensen, P., Amundsen, A., 2004. Speed and Road Accidents. An Evaluation of the Power Model. TØI report 740/2004. Institute of Transport Economics TOI, Oslo.
Guevara, F.L.D., Washington, S.P., Oh, J., 2004. Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record 1897, 191–199.
Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2003. Macrolevel accident prediction models for evaluating safety of urban transportation systems. Transportation Research Record 1840, 87–95.
Hadayeghi, A., Shalaby, A.S., Persaud, B.N., Cheung, C., 2006. Temporal transferability and updating of zonal level accident prediction models. Accident Analysis and Prevention 38, 579–589.
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.
Huang, H., Abdel-Aty, M., Darwiche, A., 2010. County-level crash risk analysis in Florida: Bayesian spatial modeling. Transportation Research Record 2148, 27–37.
Karlaftis, M.G., Tarko, A.P., 1998. Heterogeneity considerations in accident modeling. Accident Analysis and Prevention 30, 425–433.
Karlaftis, M.G., Golias, I., 2002. Effects of road geometry and traffic volume on rural roadway accident rates. Accident Analysis and Prevention 34, 357–365.
Kloeden, C.N., McLean, A.J., Glonek, G., 2002. Reanalysis of Travelling Speed and the Rate of Crash Involvement in Adelaide South Australia. Report No. CR 207. Australian Transport Safety Bureau ATSB, Civic Square, ACT.
Levine, N., Kim, K.E., Nitz, L.H., 1995. Spatial analysis of Honolulu motor vehicle crashes: II zonal generators. Accident Analysis and Prevention 27, 675–685.
Miller, H.J., Shaw, S.-L., 2001. Geographic Information System for Transportation, Principles and Applications. Oxford University Press.
Naderan, A., Shahi, J., 2010. Aggregate crash prediction models: introducing crash generation concept. Accident Analysis and Prevention 42, 339–346.
National Transportation Laboratory, U.S. Department of Transportation. An Introduction to Urban Travel Demand Forecasting – A Self Instructional Text. http://ntl.bts.gov/DOCS/UT.html (accessed 8.06.11).
Noland, R.B., 2003. Traffic fatalities and injuries: the effect of changes in infrastructure and other trends. Accident Analysis and Prevention 35, 599–611.
Noland, R.B., Oh, L., 2004. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data. Accident Analysis and Prevention 36, 525–532.
Noland, R.B., Quddus, M.A.A., 2004. Spatially disaggregate analysis of road casualties in England. Accident Analysis and Prevention 36, 973–984.
Peters, A., MacDonald, H., 2004. Unlocking the Census with GIS. ESRI Publication.
Qin, X., Han, J., 2008. Variable selection issues in tree-based regression models. Transportation Research Record 2061, 30–38.
Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accident Analysis and Prevention 40, 1486–1497.
SAS Software, Version 9.1.3 Service Pack 3. SAS Institute Inc., Cary, NC, USA.
Stewart, J.R., 1996. Application of classification and regression tree methods in roadway safety studies. Transportation Research Record 1542, 1–5.
Tignor, S.C., Warren, D., 1990. Driver Speed Behavior on U.S. Streets and Highways. 1990 Compendium of Technical Papers. Institute of Transportation Engineers, Washington, DC. http://www.ibiblio.org/rdu/sl-irre0.html (accessed 2.06.10).
The R Project for Statistical Computing. http://www.r-project.org/ (accessed 13.02.10).
U.S. Census Bureau. http://www.census.gov/geo/www/cob/tz_metadata.html (accessed 13.02.09).
Washington, S., 2000. Iteratively specified tree-based regression: theory and trip generation example. Journal of Transportation Engineering 126, 482–491.
Washington, S., Schalkwyk, I. V., Mitra, S., Meyer, M., Dumbaugh, E., Zoll, M., 2006. Incorporating Safety into Long-range Transportation Planning. NCHRP Report 546. Transportation Research Board of the National Academics, Washington, DC.
Washington, S., Schalkwyk, I.V., You, D., Shin, K., Samuelson, J.P., 2010. PLANSAFE: Forecasting the Safety Impacts of Socio-demographic Changes and Safety Countermeasures. Report No. NCHRP 8-44(2). Transportation Research Board, Washington, DC.
Wier, M., Weintraub, J., Humphreys, E.H., Seto, E., Bhatia, R., 2009. An area-level model of vehicle–pedestrian injury collisions with implications for land use and transportation planning. Accident Analysis and Prevention 41, 137–145.
Yan, X., Radwan, E., 2006. Analyses of rear-end crashes based on classification tree models. Traffic Injury Prevention 7, 276–282.
Yan, X., Richards, S., Su, X., 2010. Using hierarchical tree-based regression model to predict train–vehicle crashes at passive highway–rail grade crossings. Accident Analysis and Prevention 42, 64–74.
You, J., Nedović-Budić, Z., Kim, T.J., 1997. A GIS-based traffic analysis zone design: technique. Transportation Planning and Technology 21, 45–68.