A classification tree based modeling approach for segment related crashes on multilane highways

Anurag Pande , Mohamed Abdel-Aty , Abhishek Das

A B S T R A C T

Introduction: This study presents a classification tree based alternative to crash frequency analysis for analyzing crashes on mid block segments of multilane arterials. *Method:* The traditional approach of modeling counts of crashes that occur over a period of time works well for intersection crashes where each intersection itself provides a well defined unit over which to aggregate the crash data. However, in the case of mid block segments the crash frequency based approach requires segmentation of the arterial corridor into segments of arbitrary lengths. In this study we have used random samples of time, day of week, and location (i.e., milepost) combinations and compared them with the sample of crashes from the same arterial corridor. For crash and non crash cases, geometric design/roadside and traffic characteristics were derived based on their milepost locations. The variables used in the analysis are non event specific and therefore more relevant for roadway safety feature improvement programs. First classification tree model is a model comparing all crashes with the non crash data and then four groups of crashes (rear end, lane change related, pedestrian, and single vehicle/off road crashes) are separately compared to the non crash cases. The classification tree models provide a list of significant variables as well as a measure to classify crash from non crash cases. ADT along with time of day/day of week are significantly related to all crash types with different groups of crashes being more likely to occur at different times. *Conclusions:* From the classification performance of different models it was apparent that using non event specific information may not be suitable for single vehicle/off road crashes. *Impact on Industry:* The study provides the safety analysis community an additional tool to assess safety without having to aggregate the corridor crash data over arbitrary segment lengths.

1. Introduction

Crash counts or rates remain a popular approach for the assessment of safety on multilane arterials (or any roadway for that matter). Multilane arterials are defined as roadways with two or more lanes in each direction that have signalized and unsignalized intersections joined by mid block segments. Crash counts are traditionally estimated using negative binomial regression models (e.g., Abdel Aty & Radwan, 2000; Knuiman, Council, & Reinfurt, 1993). In crash frequency analysis the dependent variable (i.e., frequency of crashes) is calculated by aggregating the crash data over specific time periods (months or years) and locations (Abdel Aty & Pande, 2007; Golob, Recker, & Alvarez, 2004). In terms of locations, intersections are defined entities within the multilane arterials. Therefore, individual intersections act as logical units for aggregating the crash data in the form of crash frequencies (e.g., Wang & Abdel Aty, 2006). Crash frequency analysis for roadway segments, on the other hand, requires aggregation of crash data over segment(s) of certain length(s). For example, Caliendo, Guida, and Parisi (2007) divided each direction of a four lane arterial into segments with constant horizontal curvature and longitudinal slope. Donnell and Mason (2006) analyzed the crash frequencies for ½ mile segments. The selection of the length(s) of segments used to aggregate the crash data is arguably arbitrary. The results obtained from crash frequency analysis are likely to be sensitive to the lengths over which data are analyzed.

The objective of this study is to outline some of the problems associated with crash frequency analysis and propose a classification tree based alternative for identifying traffic and highway design parameters significantly associated with crashes on mid block segments of multilane arterials. The study is based on the crash data from U.S. Route 19 (also known as SR 55) in Pasco County Florida. The highway has at least two lanes in each direction and is *not* a limited access facility (i.e., expressway/freeway).

The problem here is setup as a classification problem between crash and non crash cases and classification trees are used as the analysis tool. Crash data are compared with non crash cases that are essentially random combinations of time of day and milepost locations on the same highway. The comparisons of non crash data with crash data proposed in this study allow for using crashes themselves as the unit of analysis for assessing safety on arterials as a function of geometric design, time of

day, and so forth. Mid block segment crashes used for this analysis are reported to have at least a non incapacitating injury so that the analysis proposed here is not affected by under representation of the least severe crashes in the documented crash data (Abdel Aty & Keller, 2005). Furthermore, in the proposed study we have not used variables that are event specific (such as injury severity or alcohol involvement). Milton, Shankar, and Mannering (2008) have demonstrated that the insights provided by models *event specific* explanatory variables have limited application in safety improvement programs since these *event specific* explanatory variables are required to produce useable output or inferences. The approach proposed herein has the advantages of the methodology used by Milton et al. (2008) as it uses *non event (i.e., crash) specific* factors affecting crashes on roadway sections.

The analysis presented herein is based on 545 crashes (reported from year 2004 through 2008) on 19.659 mile corridor of U.S. Route 19 in Pasco County that at least involved a non incapacitating injury. The aforementioned corridor consists of signalized intersections as well as access points without signal control (i.e., unsignalized intersections). These 545 crashes are mid block segment crashes that are not affected by the intersecting traffic streams and may be attributed only to the segments of corresponding roadways. These crashes are identified based on an extensive review of crash reports and by using the following information available in the crash database: type of crash, traffic control device, site location, and contributing cause. The comparison group for these crashes (to identify significant factors associated with their occurrence) is a sample of non crash cases that is generated by randomly selecting milepost locations, time of day, and day of week combination on this arterial. These randomly selected time and locations on the arterial (when no crash was observed) are then used as the comparison dataset for the crashes.

In the following section, details of the crash and non crash data used in this study are provided. It is then followed by information on problems associated with crash frequency analysis. The section after that discusses the classification tree models. The process of estimating generic classification tree model for comparing crash and non crash cases is then followed by classification models for specific crash type (i.e., rear end, pedestrian). The models are followed up with a discussion of the results and concluding remarks.

2. Data extraction and exploration

As mentioned earlier, the crashes attributable to mid block segments of U.S. Route 19 are the focus of this investigation. These segment crashes are defined as the crashes that are not related with the traffic on the intersecting streets. To identify these crashes, first, crashes with first harmful event characterized as "Collision with Motor Vehicle in Transport (Left turn)" and "Collision with Motor Vehicle in Transport (Right turn)" were eliminated from the database. The next task was to identify which of the *remaining* crashes may be attributable to arterial segments and *not* to (signalized or unsignalized) intersections. A detailed review of crash reports revealed that the parameter "Site location" by itself was a weak indicator. It was observed that it is possible for a crash to be not attributable to a signalized intersection even if it may have occurred very close to one. In fact, "traffic control" in combination with the "site location" did a superior job in attributing crashes to one of the three roadway elements (i.e., segments, signalized intersections, and unsignalized intersections) associated with the event of crash (Das, Abdel Aty, & Pande, 2009). Also, crashes with "Collision with Motor Vehicle in Transport (Angle)" as the identified first harmful event were excluded from the sample *if* the contributing cause for the crash was noted as "Improper turn" or "Failed to yield Right of Way." These crashes are caused by vehicles making right/left turns and/or by vehicles that fail to yield right of way to through vehicles. Crashes now remaining in the database are not attributable to signalized/unsignalized intersections and may be attributed to the segments of the multilane highways. Five hundred forty five of these crashes involved at least a non incapacitating injury and only those crashes were retained in the database.

2.1. Segment crash frequency and data aggregation level

It was mentioned previously that the crash frequency analysis may be affected by the length of the segments over which crash data are aggregated. A simple demonstration of this effect is provided in this section. Based on the process described above it was found that for the corridor under consideration there were 545 crashes resulting in at least a non incapacitating injury. The segment crash frequency was then plotted as two histograms with the corridor divided into ¼ mile (Fig. 1(a)) and ½ mile segments (Fig. 1(b)). Along with the histograms the figures also show the top five segments with the highest frequency of crashes. It may be observed that there is some difference in the locations with the highest frequency depending on if we divide the corridor in ¼ mile or ½ mile segments. It indicates that an alternative to crash frequency analysis needs to be explored for segments of the arterials. Also, according to Golob et al. (2004), aggregate studies can be susceptible to the problem of ecological fallacy. The ecological fallacy is a widely recognized error in the interpretation of statistical data, whereby inferences about the nature of individuals are based solely upon aggregate statistics collected for the group to which those individuals belong (Robinson, 1950). The studies analyzing data at individual crash level (i.e., the approach being proposed in this study) are in theory free from this fallacy.

2.2. Crash types

The crash data for analysis are divided into four collision types: (a) Rear end crashes, (b) Pedestrian related crashes, (c) Lane change related crashes, and (d) Single vehicle/off road crashes. This categorization is obtained by logically combining categories of "first harmful event" in the crash database. For example, crashes with first harmful events "Motor vehicle ran into Ditch/Culvert" and "Ran off road into water" were part of the crash type "Single vehicle off road." Lane change related crashes consist of crashes with first harmful event as "Collision with Motor Vehicle in Transport (Sideswipe)" and "Collision with Motor Vehicle in Transport (Angle)" where the contributing cause is neither "Improper turn" nor "Failed to yield Right of Way." Hence, we are considering only the angle crashes attributable to the arterial segments, which by definition are not affected by traffic streams (either from or turning on to) on intersecting roadways. The authors postulated that these crashes would never be right angle crashes. Therefore, the crashes for which the first harmful event has been noted as "Collision with Motor Vehicle in Transport (Angle)" (by the law enforcement personnel on crash site) are essentially lane change related crashes. This postulation was verified by manually reviewing 70 randomly selected crash reports for such crashes. Table 1 shows the proportion of crashes of each type in the database. The head on crashes are only 2.94% of the total sample and therefore, even with four years of crash data there were less than 20 head on crashes.

2.3. Extraction of non crash cases

A sample of non crash cases has been used in the analysis that acts as comparison data for the binary classification tree models. These non crash cases were drawn randomly from the corridor. To draw these cases, any one year period may be divided into 35,040 15 minute periods (4 (15 minute periods per hour) *24 hours*365 days = 35,040 15 minute periods), which would be the number of options available to choose the "time of non crash." Similarly, pool of possible milepost locations for the corridor consisted of mileposts starting at beginning milepost (0.0 in this case) and culminating at the ending milepost (19.635 in this case) with an increment 0.001 miles. For example, this corridor with beginning milepost 0.0 and ending milepost 19.659, there would be 688,851,360 (35040*(19.659/0.0001) = 688,851,360) options

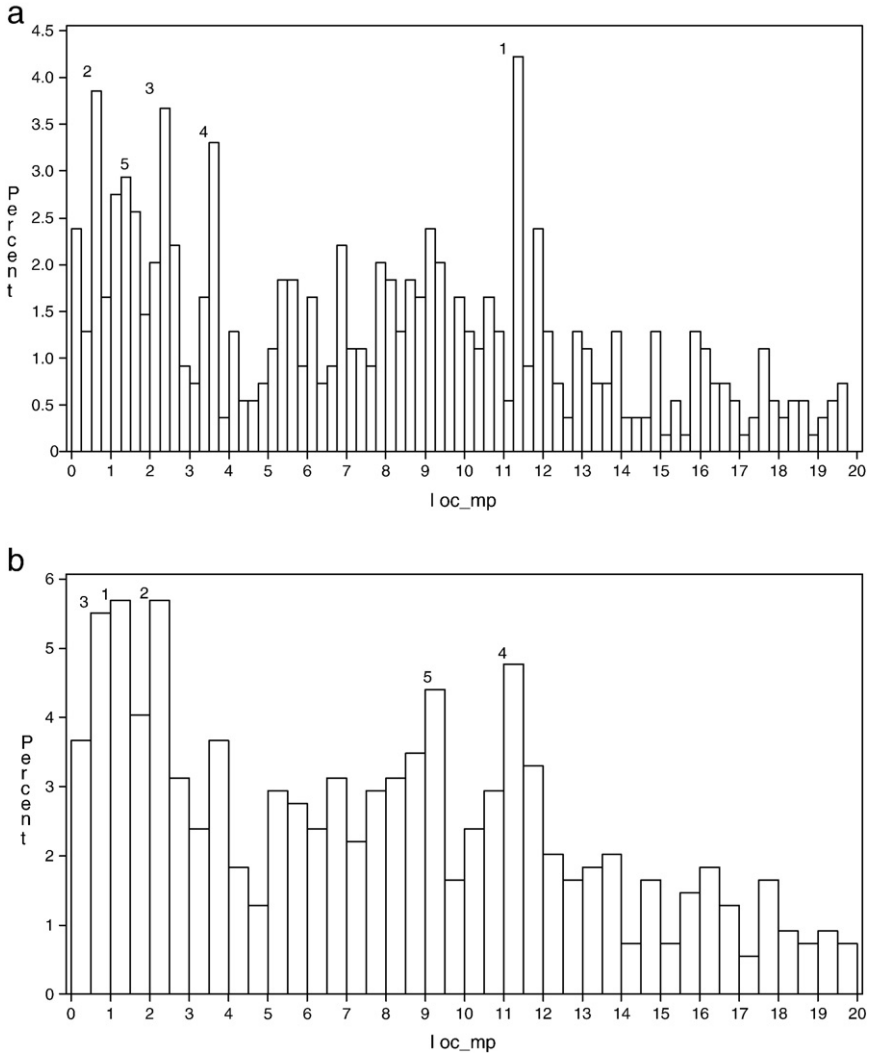

Fig. 1. (a). Histogram of mid-block segment crash frequency with US 19 corridor divided into ¼-mile segments. (b). Histogram of mid-block segment crash frequency with US 19 corridor divided into ½-mile segments.

to select (day, time, and location of) non crash cases. Required non crash cases were drawn randomly from these available options for U.S. Route 19 in Florida. The overall dataset was populated with 4,905 non crash cases such that the overall database had 10% (545) crash cases and 90% non crash cases.

2.4. Traffic/geometric information for crash and non crash cases

The next step was to extract geometric design features such as the curvature, median width, sidewalk, and so forth, for crash and non crash cases. These relevant variables were extracted from the Roadway Characteristics Inventory (RCI) database (Florida Department of Transportation, 2001). The extraction of traffic/geometric information was based on the milepost locations and the roadway ID for the arterial corridor. The roadway ID for U.S. Route 19 in Florida was "14030000." For crashes, it was the actual mile post location of the crash from the FDOT (Florida Department of Transportation) crash database and for non crash cases it was assigned milepost using the procedure described in the previous section.

RCI database provides information on Florida's state maintained road network indexed by data segments. RCI features are listed in the handbook (*RCI Features and Characteristics Handbook*, 2001) published by FDOT (Florida Department of Transportation) and Table 2 details the relevant variables extracted from this database. Note that most of the variables tabulated are not in the same form as the original database. The original categories of the variables in the RCI database were combined to create variables with fewer categories. Table 2 also provides the percentages of crash and non crash cases for all categories of the variables listed. It may be compared to the overall percentage of crash and non crash cases in the database (found in the header row of Table 2) to get a descriptive estimate of the variables or categories associated with crash occurrence on multilane highways. For an easy comparison, the categories with more than 10% (overall proportion of crash cases in the database) crash cases are showed in a different (lighter) shade compared to categories with less than 10% crashes (darker shade).

Table 1
Proportion of various crash types in the dataset.

Crash type	Percentage among the crashes with known crash type (%)
Rear-end	43.44
Head-on	2.94
Lane-change related	17.22
Pedestrian	19.96
Single-vehicle/Off-road	16.44
Total	100

Table 2
Variables used in the analysis.

Variable Description	Categories	Percentage of non-crash cases (90%)*	Percentage of crash cases (10%)*
Posted speed limit	Speed limit = 45 MPH,	87.79	12.21
	Speed limit >45 MPH (50 or 55 MPH)	94.88	5.12
ADT (Annual daily traffic)	ADT <= 54,125	94.06	5.94
	and ADT >54,125	87.24	12.76
Average Truck Factor	T-factor <= 4.6423	87.99	12.01
	and T-factor >4.6423	95.08	4.92
Combination of day of week and time of day	Afternoon Peak Weekday	81.52	18.48
	Friday or Saturday Night	87.39	12.61
	Morning Peak Weekday	92.84	7.16
	Other Off-peak Periods	90.88	9.12
Median width	Median width = 18 ft.	88.64	11.36
	Median width = 24 ft.	88.15	11.85
	Median width = 28 ft.	90.3	9.7
Presence of Sidewalk (Binary)	No	90.13	9.87
	Yes	89.54	10.46
Presence of on-street parking (Binary)	No	89.96	10.04
	Yes	91.67	8.33
Presence of horizontal curvature (Binary)	No	89.96	10.04
	Yes	90.17	9.83

*Represents overall percentage of crash and non-crash cases in the database.

It may be observed from Table 2 that originally continuous variables, ADT, and Percentage of trucks (T factor) were transformed into binary categories. To create these binary variables the original continuous variables were recursively split into groups until the association of the resultant grouping with the binary target *y (y = 1 for crash cases and y = 0 for non crash cases)* is maximized. This transformation of these two variables was deemed necessary based on an observation by Strobl, Boulesteix, Zeileis, and Hothorn (2007). It was noted by Strobl et al. that if there is a wide variation in the number of categories of various variables used in the analysis, the classification trees are biased toward concluding variables with a large number of categories as more important.

Time of crash (and non crash cases), along with day of week, were combined into one variable representing day of week *and* time of day. The four categories of this variable include weekday morning peak hour, weekday afternoon peak hour, Friday/Saturday night, and other off peak periods. Note that the weekend night time was separated from the other off peak periods because of the increased likelihood of alcohol impaired driving. Three binary variables representing the presence of horizontal curvature, sidewalk, and roadside parking were also used in the analysis. Note that the variable median width with three different levels is not used as ordinal variable but as a nominal variable. The nominal scale ensures that one is able to capture the non monotonous nature of the relationship between median width and crash occurrence.

Some of the other variables that were considered include median type, pavement surface conditions, and K factor (design hour volume as a percentage of ADT). These variables could not be included in the analysis for the lack of sufficient variation in their values along the 19.659 mile corridor. The variables shown in Table 2 are *not* event specific characteristics (such as driver characteristics and seat belt use) which, as Milton et al. (2008) argued, allows for a more general, non event specific interpretation of factors.

3. Modeling methodology

The proposed approach is based on classification tree that is one of the more popular data mining algorithms. Data mining is the analysis of large "observational" datasets to find unsuspected relationships that might be useful to the data owner (Hand, Mannila, & Smyth, 2001). It typically involves analysis where objectives of the data analysis have no bearing on the data collection strategy. RCI database (maintained by FDOT) is a good example of such "observational" database. The output of the classification tree models is a set of simple rules that can be interpreted easily. It gives classification tree a big advantage over other data mining tools such as neural networks where the results are hard to interpret. The basic idea in the classification tree construction is to split the dataset such that the resulting dataset is 'purer' than the parent. Classification tree strives toward nodes that are pure in the sense that they contain observations belonging to a single class. To achieve this, a set of candidate split rules is created, which consist of all possible splits for all variables included in the analysis. A measure indicating how far a node is from this ideal situation is called an impurity measure (SAS/STAT® 9.1 User's Guide, 2004).

In this study these splits are evaluated and ranked based on Gini reduction criterion to choose amongst the available splits at every non terminal node. While developing a classification tree, this criterion is applied recursively to the descendents to achieve child nodes having maximum worth. Child nodes in turn become the parents to successive splits, and so on. The splitting process is continued until there is no (or less than a pre specified minimum) reduction in impurity and/or the limit for the minimum number of observation in a leaf is reached (SAS/STAT® 9.1 User's Guide, 2004). Gini reduction criterion measure the "worth" of each split in terms of its contribution toward maximizing the homogeneity through the resulting split. If a split results in the splitting of one parent node into B branches, the "worth" of that split may be measured as follows (SAS/STAT® 9.1 User's Guide, 2004):

$$Worth = Impurity(Parent\ node) \quad \sum_{b=1}^{B} P(b)^{*}Impurity(b) \tag{1}$$

Where *Impurity (Parent node)* denotes the Gini measure for the impurity (i.e., non homogeneity) of the parent node and *P (b)* denotes the proportion of observations in the node assigned to branch b. The impurity measure, *Impurity (node)*, may be defined as follows:

$$Impurity(node) = 1 - \sum_{i}^{classes} \left(\frac{number\ of\ class\ i\ cases}{all\ cases\ in\ the\ node}\right)^2 \tag{2}$$

$$= 1 - \left[(p_{crash})^2 + (p_{non\ \ crash})^2\right]$$

If a node is 'pure' (i.e., consists of only crash or only non crash cases) than the Gini measure will have minimum value, and its value will be higher for less homogeneous nodes. Classification trees developed in this study serve two purposes: (a) the models provide tools for classification between crash and non crash cases, and (b) they provide a variable importance measure for each of the variables used in the analysis.

Breiman, Friedman, Olshen, and Stone (1984) devised variable importance measure (VIM) based on classification trees. In a classification tree with T total nodes, let $S(x_j, k)$ be the split at the k^{th} internal node using the variable x_j. The variable importance measure for variable x_j is the weighted average of the reduction in the Gini impurity measure (defined in Eq. (2)) achieved by all splits using the variable x_j across all internal nodes of the tree and the weight is the node size. If N is the total number of observations in the training sample, then the formula for the importance for variable x_j may be given by the following:

$$VIM\left(x_j\right) = \sum_{t=1}^{T} \frac{n_t}{N} \Delta Gini\left(S\left(x_j, t\right)\right) \tag{3}$$

Where $\Delta Gini\ (S(x_j, t))$ is the reduction in Gini measure of impurity (defined in Eq. (2)) achieved by splitting the variable x_j at node t, and $\frac{n_t}{N}$ represents the proportion of the observations in the dataset that belong to node t (also see Pande & Abdel Aty, 2006).

Eq. (3) represents the variable importance measure as proposed by Breiman et al. (1984). In this study, however, the VIM used has been scaled by maximum importance for the tree so that the measure lies between 0 and 1. In the following section the classification analysis of

Table 3
List of significant variables based on the generic model.

Name	Importance
Speed Limit	1.0000
Time of Day/Day of Week	0.9167
Curvature	0.3680
ADT	0.3237
Sidewalk	0.2651
Roadside Parking	0.0000
T-factor	0.0000
Median Width	0.0000

crash and non crash data is presented along with significant variables and evaluation of classification performance.

4. Crash versus non-crash classification: Generic model

The first step in the analysis was to estimate generic classification tree models, where the binary target variable y represents *crash (y = 1)* versus *non crash* cases (y = 0). The dataset used here includes all 545 crashes that were compared to the 4,905 non crash cases. The dataset was partitioned into 70% training and 30% validation set. Table 3 shows the variables found significantly associated with all crashes based on the estimated VIM. The variables with higher VIM values are the most significant. In other words, they are the most critical for distinguishing between crash and non crash cases.

According to the generic tree model the speed limit posted on the highway is the most important factor followed by time of day/day of week and presence of curvature. Presence of roadside parking, T factor, and median width were not found to be significant (i.e., VIM = 0.0000). However, since the sample used to calibrate the tree model providing this list comprises all types of crashes, one cannot make effective conclusions about how these parameters lead to increased likelihood of crashes.

The classification performance of this generic model over the validation dataset is measured based on the lift plot instead of classification accuracy based on a pre determined threshold. If the classification tree model is applied to the validation dataset the output of the model (for each observation) is the posterior probability of the event of interest (i.e., a crash). Posterior probability is a number between 0 and 1. The closer it is to 1 the more likely, according to the model, it is for that observation to be a crash. To assess the classification performance of the model the observations in validation dataset were sorted by the output posterior probability. In the sorted group, top *10%* observations would be the *10%* observations that are the most likely to be a crash, according to the model. The performance of a model may be measured by determining the proportion of crashes in the validation dataset captured within various deciles[2] of posterior probability. It is worth noting that the overall classification accuracy over validation dataset would not be a good measure for model performance evaluation. With only *5%* crashes in the sample, classification accuracy as high as 95% could be achieved by a model that merely classifies every observation as non crash. Such a model would of course be useless for the objectives of this study.

Fig. 2 shows the lift plot for the generic classification tree model. The curve shows the percentage of crashes in the validation dataset captured within various deciles of posterior probability by the model on y axis. On the x axis the percentiles are shown at equal intervals of 10. Fig. 2 also demonstrates '*performance*' of a random baseline model that represents the expected percentage of crashes identified in the validation dataset if one randomly assigns validation dataset observations as crash and non crash. A model can be assessed for its performance by examining the separation of the corresponding lift curve from the random baseline curve. Larger separation from the random baseline model indicates better classification performance.

Note that the model presented here is generic in nature (i.e., single generic model has been used to identify all crashes regardless of their type rear end, sideswipe, or angle). Highway design parameters associated with crashes are likely to differ by type of crash and therefore the classification tree models should also be type (of crash) specific in nature. The disaggregate models would also be insightful while devising remedial measures to improve the safety on the highways since the countermeasures would also differ for each type of crash. Note that using the models by specific crash type would also result in improved classification performance of the models. To demonstrate the improved performance based on the specific crash type models, the generic model based preliminary analysis has been retained in this study (Fig. 2).

5. Crash versus non-crash classification: Specific crash types

Extended classification tree based analysis where specific crash types are compared separately to non crash cases is presented in this section. Four such classification tree models were considered in all with the set of non crash cases compared to: (a) Rear end crashes, (b) Pedestrian related crashes, (c) Lane change related crashes, and (d) Single vehicle/off road crashes. Note that the sample of head on crashes was too small to estimate the corresponding classification tree model. These four models yielded the most significant factors associated with occurrence of each crash type. The classification performance was also evaluated individually for each model and the results are presented herein.

Table 4 provides the factor significant to all four types of crashes along with the corresponding VIMs. The last column of the table refers to the information from Table 3 (i.e., significant factors for the generic model). Each cell of Table 4 also includes a ranking in parenthesis corresponding to that variable's relative significance for each crash type. In the discussion that follows, the relationship between these parameters and crash occurrence has been explored. The discussion is based on the set of simple rules provided by the four separate classification tree models.

It is worth noting that the speed limit that was one of the most significant factor when analyzing all crashes combined (Table 3) either has relatively low VIM or no significance at all (in case of Single vehicle/Off road) crashes. Time of day/day of the week is the most significant parameter associated with all four groups of crashes. Examining each of the classification tree models (i.e., the set of rules from each model) closely, it was found that different times of day are susceptible to different types of crashes.

Rear end crashes are more likely to occur on sections with higher ADT and during afternoon peak period on weekdays. It was also found that during off peak period sections with higher ADT (>54,125) and 24 ft. wide median were also more likely to have rear end crashes. The classification tree model for rear end also showed that terminal node corresponding to Friday/Saturday night, sections with higher speed limit (50 or 55 MPH) and lower ADT (<=54,125) was a pure node with no (rear end) crashes. Afternoon peak hours on weekdays are also more likely to have lane change related crashes. Traffic congestion during afternoon peak period on weekdays will prompt drivers to change lane more frequently. Interestingly, during morning peak hours sections with low Truck factor (<=4.6423%) and no sidewalk are *least* likely to have lane change related crashes. On the other hand, during Friday/Saturday nights sections with high truck factor (>4.6423%) are more likely to have lane change related crashes. Truck factor has highest relative significance for the lane change crashes (compared to other crash types), which is consistent with findings from one of our previous studies (Pande & Abdel Aty, 2009). Note that curvature is a significant factor for rear end crashes (with a low VIM), however, it is a more significant factor in case of lane change related crashes.

[2] Decile is defined as any of nine points that divide a distribution of ranked scores into equal intervals with each interval containing one-tenth of the scores.

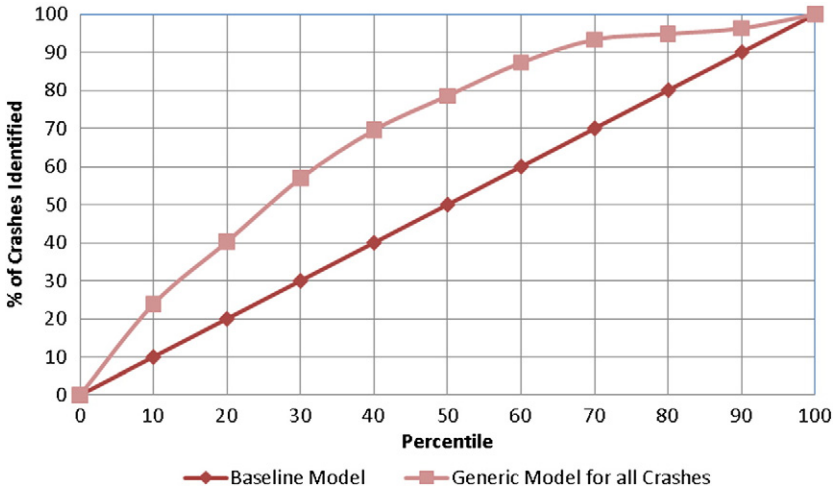

Fig. 2. Classification model performance of the Generic model and the baseline model.

Pedestrian and single vehicle/off road crashes are more likely to occur on Friday and Saturday night, when the road users (drivers and/or pedestrians) are more likely to drive/walk under influence. The set of conditions that made the pedestrian crashes most likely was wide medians (24 or 28 ft.) on Friday/Saturday night on sections with higher ADT (>54,125). Presence of roadside parking is also found to be significant in pedestrian crashes, while interestingly the presence of sidewalk has no significant association with them. ADT is the second most significant factor in all crashes but the single vehicle/off road type. The single vehicle/ off road crashes might be influenced more by driver and vehicle related *event specific* factors compared to traffic or geometric factors being considered here. Corresponding classification tree model showed that Friday/Saturday night time was the single most important factor in determining likelihood of single vehicle/off road crashes. During other times of day/day of week, presence of on street parking increased the likelihood of this crash type. However, as we shall observe next, the classification performance of the tree model leading to these interpretations was poor and therefore the results for this particular tree model may not be as reliable.

It is interesting to note that the performance of the classification tree model for single vehicle crashes is much worse compared to all the other models (Fig. 3). In fact for the first two deciles it is at or below the random "baseline" model. It may be explained by the fact that single vehicle crashes are likely more influenced by driver behavior and not by the highway design parameters that are used here. Hence, these crashes are harder to 'predict' using these variables and it shows in the corresponding lift plots depicted in Fig. 3. The classification model with the best performance is the one for pedestrian crashes since the lift plot corresponding to it is consistently above the other four curves. Note that these lift plots are created by applying the classification tree model on the validation dataset, which were not used for training the models.

6. Concluding remarks

This study provides a classification tree based alternative to crash counts based analysis for identifying significant factors related with crash risk on mid block segments of multilane arterials. The fundamental difference between this approach and crash frequency analysis is that crash counts do not need to be aggregated over roadway segments of arbitrarily selected length value that may influence the results. Potential problems related with such aggrega tion were also demonstrated in the study. In this study crashes are differentiated by type and variables potentially affecting crash occurrence are included explicitly in the classification tree models.

Crash versus non crash binary classification can also be accom plished by logistic regression model. However, one has to ensure that the assumptions of the model structure are not violated. For example, explicitly using two or more correlated independent variables in a logistic regression model may violate underlying model assumptions. One can use a subset of variables to stratify the data and then estimate separate logistic regression models for the stratified samples to ensure

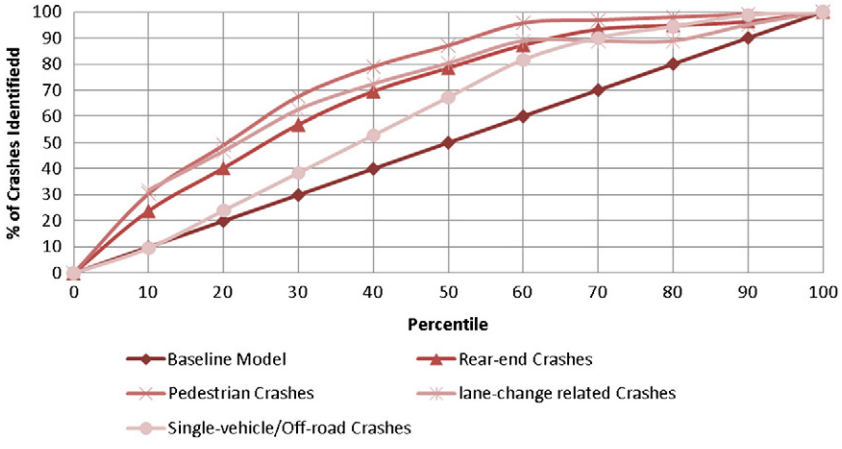

Fig. 3. Classification performance of the models for each crash type.

Table 4
List of Significant Variables and corresponding VIMs.

Variables	Crash Types				
	Rear-end	Pedestrian	Lane-change related	Single vehicle/ Off road	All Crashes (Table 3)
Speed Limit	0.2240 (5)	0.2407 (5)	0.4948 (4)	0.0000 (NA)	1.0000 (1)
Time of Day/ Day of Week	1.0000 (1)	1.0000 (1)	1.0000 (1)	1.0000 (1)	0.9167 (2)
Curvature	0.0814 (7)	0.0000 (NA)	0.4622 (5)	0.0000 (NA)	0.3680 (3)
ADT	0.9417 (2)	0.9449 (2)	0.7605 (2)	0.0000 (NA)	0.3237 (4)
Sidewalk	0.4700 (3)	0.0000 (NA)	0.2631 (7)	0.0000 (NA)	0.2651 (5)
Roadside Parking	0.1285 (6)	0.4043 (4)	0.0339 (8)	0.5276 (2)	0.0000 (NA)
T-factor	0.0763 (8)	0.0000 (NA)	0.2655 (6)	0.2296 (4)	0.0000 (NA)
Median Width	0.3826 (4)	0.6286 (3)	0.5684 (3)	0.3558 (3)	0.0000 (NA)

NA: Ranking not applicable since the variable has VIM = 0.000.

that the independent variables used in the models are not correlated with each other. Classification tree models used in this study, however, do not require any such underlying assumptions and even correlated independent variables can be included explicitly.

The results showed that more vehicles on the road (on sections with higher ADT during weekday peak hours) increase the likelihood of rear end crashes. Higher percentage of trucks increased the likelihood of lane change related crashes, indicating that on multilane arterial sections with higher T factor lane change restrictions might be needed. Pedestrian related as well as single vehicle/off road crashes were likely to occur on Friday/Saturday nights. It is worth noting that the parameters used in this study are non event specific in nature based on the practical considerations outlined by Milton et al. (2008). However, a comparison of classification performance of the four classification tree models (one for each crash type) calibrated in this study showed that while the performance of three of the models' was comparable to each other, the model for single vehicle/off road crashes performed poorly. It led to the inference that the occurrence of this group of crashes is not adequately explained based on the non event specific parameters that are used here and driver/vehicle characteristics need to be included in the analysis for at least this group of crashes.

The methodology to derive non crash cases, as a substitute for crash frequency analysis, may be easily implemented for freeway corridors as well. It is worth mentioning that this approach is limited in that it is not suitable for analyzing intersections' crash patterns. Assigning non crash cases to an intersection is not as simple as it is with the segments of the arterials. Comparisons between selected non crash cases with the signalized (or unsignlaized) intersection related crashes, for example, would yield information that would mostly reflect the characteristics belonging to *locations of the signalized intersection* and not much else. However, with segment crashes the classification tree based comparisons provide geometry/traffic related parameters that significantly relate with crash occurrence on the segments. Since individual intersections provide logical units for aggregating the crash data, a frequency approach is still best suited for analysis of intersection crashes. The small number of head on crashes on U.S. 19 also limited the analysis as these crashes could not be analyzed with an independent classification tree model.

Acknowledgement

The authors wish to thank the Florida Department of Transportation for supporting part of this research.

References

Abdel-Aty, M., & Keller, J. (2005). Exploring the overall and specific crash severity levels at signalized intersections. *Accident Analysis and Prevention*, *37*(3), 417 425.
Abdel-Aty, M., & Pande, A. (2007). Crash data analysis: Collective vs. individual crash level approach. *Journal of Safety Research*, *38*(5), 581 587.
Abdel-Aty, M., & Radwan, E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention*, *32*(5), 633 642.
Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth. Inc. 358.
Caliendo, C., Guida, M., & Parisi, A. (2007). A crash-prediction model for multilane roads. *Accident Analysis and Prevention*, *39*(4), 657 670.
Das, A., Abdel-Aty, M., & Pande, A. (2009). Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. *Journal of Safety Research*, *40*(4), 317 327.
Donnell, E., & Mason, J. (2006). Predicting the frequency of median barrier crashes on Pennsylvania interstate highways. *Accident Analysis and Prevention*, *38*(3), 590 599.
Florida Department of Transportation. (2001). *RCI Features and Characteristics Handbook.* Tallahassee, FL: Author.
Golob, T., Recker, W., & Alvarez, V. (2004). Freeway safety as a function of traffic flow. *Accident Analysis and Prevention*, *36*(6), 933 946.
Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining (adaptive computation and machine learning).* Cambridge, MA: Bradford Book.
Knuiman, M., Council, F., & Reinfurt, D. (1993). Association of median width and highway accident rates. *Transportation Research Record*, 70-70.
Milton, J., Shankar, V., & Mannering, F. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis and Prevention*, *40*(1), 260 266.
Pande, A., & Abdel-Aty, M. (2006). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis and Prevention*, *38*(5), 936 948.
Pande, A., & Abdel-Aty, M. (2009). A novel approach for analyzing severe crash patterns on multilane highways. *Accident Analysis and Prevention*, *41*(5), 985 994.
Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, *15*(2), 351 357.
SAS/STAT® 9.1 User's Guide. (2004). Cary, NC: Author.
Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 25.
Wang, X., & Abdel-Aty, M. (2006). Temporal and spatial analyses of rear-end crashes at signalized intersections. *Accident Analysis and Prevention*, *38*(6), 1137 1150.