Using conditional inference forests to identify the factors affecting crash severity on arterial corridors

Abhishek Das , Mohamed Abdel-Aty , Anurag Pande

A B S T R A C T

Introduction: The study aims at identifying traffic/highway design/driver vehicle information significantly related with fatal/severe crashes on urban arterials for different crash types. Since the data used in this study are observational (i.e., collected outside the purview of a designed experiment), an information discovery approach is adopted for this study. *Method:* Random Forests, which are ensembles of individual trees grown by CART (Classification and Regression Tree) algorithm, are applied in numerous applications for this purpose. Specifically, conditional inference forests have been implemented. In each tree of the conditional inference forest, splits are based on how good the association is. Chi square test statistics are used to measure the association. Apart from identifying the variables that improve classification accuracy, the methodology also clearly identifies the variables that are neutral to accuracy, and also those that decrease it. *Results:* The methodology is quite insightful in identifying the variables of interest in the database (e.g., alcohol/ drug use and higher posted speed limits contribute to severe crashes). Failure to use safety equipment by all passengers and presence of driver/passenger in the vulnerable age group (more than 55 years or less than 3 years) increased the severity of injuries given a crash had occurred. A new variable, 'element' has been used in this study, which assigns crashes to segments, intersections, or access points based on the information from site location, traffic control, and presence of signals. *Impact:* The authors were able to identify roadway locations where severe crashes tend to occur. For example, segments and access points were found to be riskier for single vehicle crashes. Higher skid resistance and k factor also contributed toward increased severity of injuries in crashes.

1. Introduction

Principal and minor arterial corridors with partially limited access experience a significant proportion of severe/fatal crashes. These corridors account for 43.4% of the fatal crashes in Florida (National Highway Traffic Safety Administration [NHTSA], 2007) resulting in 1,478 fatalities during 2006. The objective of the study is to identify contributing factors related to severe/fatal crashes occurring on the high speed (speed limit greater than 45 mph), multilane (more than one lane in each direction of travel) corridors in the state of Florida. Many safety studies identify contributing factors and use various modeling techniques for the same. Improvements in modeling methodology lead to better detection of causal factors. In this study the authors have not only introduced certain new variables (improve ment in data), but also have adopted new data mining methodology to better the understanding.

Approaches to safety on multilane corridors have traditionally been twofold. Brown and Tarko (1999), Abdel Aty and Radwan (2000), and Rees (2003) treated the corridors in totality; while Milton and Mannering (1998) and Miaou and Song (2005) divided the corridors into segments and intersections. Abdel Aty and Wang (2006) have shown a spatial correlation between crash patterns of successive signalized intersections, which may be attributed to the characteristics of the segments joining them.

Though both approaches have worked well for investigation purposes, the issue that still remains is how to assign crashes to the segments and the intersections. There is no uniformity in the influence area of an intersection among the states. For example, in Florida, all the crashes occurring within 250 ft. from the center of an intersection are categorized as intersection related crashes, as has been reported by Abdel Aty and Wang (2006) and Wang, Abdel Aty, and Brady (2006). Recently Das, Pande, Abdel Aty, and Santos (2008) showed that proximity only is not the best way to assign crashes. Wang, Abdel Aty, Nevarez, and Santos (2008) used frequency modeling for crashes with fixed as well as varying influence distance and found a different set of significant factors. Apart from the above research, it is also common knowledge that the way the crashes are

reported varies among different administrative units. The authors investigated several crash reports and came up with an innovative approach to assign crashes, the details of which are given in the next section which explains the data used in the study.

As previously mentioned, it is important not only to find the contributing factors but also to improve on the methodology adopted. Pande and Abdel Aty (2008) in their work on association rules point out that data mining techniques remain underutilized for analysis of crash. The underutilization is especially noteworthy since most studies use observational data collected outside the purview of an experimental design. Simple data mining tools like classification and regression trees have traditionally been used to identify variables of importance in safety studies (Pande & Abdel Aty, 2008). A decision tree, with all its simplicity and handling of missing values, can be very unstable. However, if instead of one tree, an ensemble of trees (commonly referred as forest) is used, the outputs become much more stable. The robustness of the forests makes them a better choice than the use of single trees. In this regard, Random Forests, developed using the Classification and Regression Trees (CART) algorithm, have been used by the authors (Abdel Aty, Pande, Das, & Knibbe, 2008) recently to identify variables of significance and then develop neural network classifiers. However, the method has been shown to have selection bias as shown by Strobl, Boulesteix, Zeileis, and Hothorn (2007). The selection bias is in favor of variables that are continuous or have higher number of categories. At the root of this selection bias is the application of 'Gini' index criterion to split a node (while building the tree) as well as for variable selection (generally based on the frequency a variable was chosen for the split). Details of the 'Gini' index criterion and the resulting bias have been provided in the 'Modeling Methodology' section. Hence, in this study conditional inference trees, developed by Hothorn, Hornik, and Zeileis (2006), and their forests have been used for the purpose of variable selection. The authors are of the belief that the application of this new methodology will improve traffic safety research. Details of how this algorithm is different (and better suited for the application at hand) than the CART have been given in the methodology section.

The authors included new variables like 'element,' in this study, which assigns crashes to segments, intersections, or access points based on the information from site location, traffic control, and presence of signals. The authors were able to identify roadway locations where severe crashes tend to occur. Failures to use safety equipment by all passengers and presence of driver/passenger in the vulnerable age group (more than 55 years or less than 3 years) were also other new variables that were included in the data. The details of how the inclusion helped in a better understanding of the severity aspect has been discussed in the 'Analysis and Results' section later on in the paper.

Crash data from the high speed multilane arterials with partial access control in Florida have been collected. These arterials have been divided into groups based on their lengths and roadway design standards (urban/suburban and rural). The following section will focus on the details of the data collection and aggregation. It is followed by the methodology section where conditional inference trees and forests will be discussed. The results and analysis section will explain the results from the conditional inference trees and the forests. While the random forests provide a more robust set of variables associated with severe/fatal crashes, individual tree helps in making relevant inferences about the relationship.

2. Data collection and preparation

2.1. Study area and available data

The crash data available were from the Crash Analysis and Reporting (CAR) system of the Florida Department of Transportation (FDOT). The Roadway Characteristics and Inventory (RCI) data were also made available to us through the FDOT. The data used are for the years 2004 through 2006 for all the state roads of Florida. The datasets have information regarding traffic, roadway geometric, and driver related factors. The datasets were merged and the parameters were modified to suit the data mining methodology being implemented in the study. The corridors, which were originally divided according to administrative units (i.e., roadway ids based on county boundaries), were logically combined to form continuous sections based on design standards. The details of the applied design standards are given in the next sub section.

2.2. Data Preparation

As mentioned earlier, the corridors available for the study were logically combined into continuous sections based on their design. Corridors with continuous urban/ sub urban design were grouped together and so also the ones with rural design. However, it should be noted that in the present study the authors focus only on the urban/ suburban corridors. Since the corridors are of variable lengths it was logical to cluster them based on the same parameter before further analysis on severity could take place. The optimum number of clusters was found based on the Partitioning around the Medoids (PAM) algorithm proposed by Kaufman and Rousseeuw (1990). In the PAM algorithm, which operates on the average dissimilarity, a 'medoid' is an object of the cluster whose average dissimilarity to all the objects in the cluster is minimal. Once the medoids are identified, all the objects are assigned to the nearest medoid. The objective function is the sum of the dissimilarities of all the objects to the nearest medoid. The algorithm terminates when the interchange of an unselected object with an already selected object no longer minimizes the objective function. The optimum number of clusters was found to be four. The following are the length of the corridors in each cluster: Cluster 1 (1.009 2.89 miles); Cluster 2 (2.898 5.729 miles); Cluster 3 (5.762 10.556 miles); Cluster 4 (10.644 78.293 miles).

Different types of crashes occur on the corridors and the contributing causes for the different types also vary. Even though the overall safety of the corridor is being analyzed, the approach to investigate different crash types separately would shed more light. The crashes were grouped into five major types as follows: (a) angle/ turning movement; (b) rear end; (c) head on; (d) sideswipe; and (e) crashes involving single vehicles.

The conditional inference trees used in this study helps us in identifying the contributing factors associated with the severity of the crashes that occurred along a corridor. However, too many parameters lessen the discriminating ability of the models as the overall degrees of freedom available for the model development decrease. Hence only a subset of the available factors should be chosen for model development. Milton, Shankar, and Mannering (2008) have also pointed out that event specific variables are least desirable in developing injury severity models. Hence, for the analysis a few variables were chosen based on engineering judgment and taking into consideration that event specific factors are not in use to a relatively large extent. The variables were broadly based on two different categories: (a) environmental and road geometric factors; (b) driver and vehicle related factors. The variables used in the study are described in Table 1. They have been derived directly from the datasets or a combination of parameters. Both these sets of parameters have their application values.

The variables illustrated in Table 1 are mostly derived from the RCI database. Many variables have too many categories, in the raw form, to start off with. Hence, level reduction in variables is not only critical but also simplifies the model and makes them more readily explainable. For example, vehicle movement, vehicle type, roadway conditions, vision obstruction, surface condition, surface type, and type of median are some of the variables with many categories. Also, the proposed methodology (conditional inference trees/forests) uses Chi square test statistic to identify the relationship between a particular parameter and

Table 1
Dependent / Independent Variables used in the analysis.

Variable Name	Variable Description	Urban / Sub-urban
Target or Dependent Variable		
Sev	Severity	Binary (1 = incapacitating injuries/ fatalities; 2 = possible/ non-incapacitating injuries)
Environmental and Roadway Geometric Parameters		
pavecond	Pavement condition	4 levels (poor, fair, good and very good)
surf_type	Type of surface	Binary (1 = black top surface; 2 = other)
surface_width	Surface width	Continuous
shld_t	Type of shoulder	Binary (1 = paved; 2 = unpaved)
max_speed	Maximum posted speed limit	Continuous
park	Presence of parking	Binary (1 = no; 2 = yes)
skid_f	Friction resistance	Skid <= 34 34<skid <= 38 Skid>38
median	Types of median	9 levels (0 = no median; 1 = painted; 2 = median curb <= 6"; 3 = median curb >6"; 4 = lawn; 5 = paved; 6 = curb <= 6" and lawn; 7 = curb>6" and lawn; 8 = other)
ACMANCLS_num	Type of median openings	7 levels (0 = no median opening; 2 = restrictive opening w/ service roads; 3 = restrictive median; 4 = non restrictive median; 5 = restrictive median with shorter directional openings; 6 = non restrictive median with shorter signal connection; 7 = both restrictive and non-restrictive median types)
road_cond	Road condition at time of crash	Binary (1 = no defects; 2 = defects)
vision	Vision obstruction	Binary (1 = no; 2 = yes)
shld_side	Shoulder + sidewalk width	Continuous
curvclass	Horizontal degree of curvature	6 levels (curve<4′; 4<= curve <= 5′; 5<curve <= 8′; 8<curve <= 13′; 13<curve <= 27′; curve>27′
surf_cond	Surface condition	Binary (1 = dry; 2 = other)
light	Daylight condition	Binary (1 = daylight; 2 = other)
ADT	Annual daily traffic	ADT <= 31000 31000<ADT <= 40000 40000<ADT <= 52500 ADT>52500
t_fact	Average truck factor	t_fact <= 4.05 4.05<t_fact <= 5.895 t_fact>5.895
k_fact	Average k - factor	k_fact <= 9.85 k_fact>9.85
dayandtime	Combination of the day of week and time of day	Afternoon Peak Weekday Morning Peak Weekday Friday or Saturday Night Off-peak
trfcway	Vertical curvature	Binary (1 = level; 2 = upgrade/ downgrade)
element/ element 1	Assignment of crashes to roadway elements	Ternary (1 = segment; 2 = intersections; 3 = access points) / Binary (1 = segments/ access points; 2 = intersections)
LIGHTCDE	Street lighting	Ternary (Y = full lighting; N = no lighting; P = partial lighting)
Driver and Vehicle related Parameters		
age_gr	Age group of the at fault driver	Age < = 25; 25 < age <= 35; 35 < age <= 45; 45 < age <= 55; 55<age <= 65; 65<age <= 75; Age>75
veh_type1	At-fault type of vehicle	4 levels (1 = automobiles; 2 = light trucks; 3 = heavy vehicles; 4 = light slow moving vehicles)
alcohol_use	Alcohol/ drug use of the at-fault driver	3 level (1 = no use; 2 = use; 3 = no information)
vuln_age	Presence of vulnerable age group passengers in the vehicle (age<5 or age>55)	Binary (1 = yes; 2 = no)
more	Presence of more than 5 passengers inside either of the involved vehicles	Binary (Y = yes; N = no)
sfty	Use of safety equipment in the vehicle by driver/passengers	Binary(1 = yes; 2 = no)
gender	Gender of the at-fault driver(s)	3 levels (1 = male; 2 = female; 3 = both)
veh_move1	Vehicle movement of the at-fault vehicle	4 levels (1 = straight ahead; 2 = turning movements; 3 = changing lanes; 4 = other)

target variable. Each category of the variable should have a sufficient number of observations in the contingency table for the Chi square to be evaluated as discussed by Das et al. (2008). Continuous variables like annual daily traffic (ADT), percentage of trucks, and K factor (design hour volume as a percentage of ADT) and skid (friction resistance multiplied by a factor of 100) were also divided into categories. Their relationships with severe/fatal crash occurrence may not be mono tonous in nature. Time of crash, along with day of week, were combined into one variable representing day of week *and* time of day. The weekend night times were not treated as off peak hours as there may be higher instances of alcohol impaired driving.

The authors have introduced some new variations to the traditional parameters. Traditionally the site location variable has been used by researchers to assign crashes to the three roadway elements (segments, intersections and access points). However a detailed review of several hundred crash reports suggested that the 'site location' variable by itself was a weak indicator for the same. For example, it was observed that it is possible for a crash to not be attributed to a signalized intersection even if it may have occurred very close to one. In fact, 'traffic control' in combination with the 'site location' along with the information of the presence or absence of signal, did a superior job in attributing crashes to one of the three roadway elements. Based on these three independent parameters, a variable 'element' was created to assign the crashes to the three roadway elements, namely segments, intersections, and access points. However it was also observed and verified through the study of crash reports that distributing crashes to the three roadway elements works fine with all types of crashes except for the angle / turning related crashes. Most of such crashes occur at the signalized intersections. The crashes that occur on the segments were observed to have occurred mostly on auxiliary lanes (right / left turning lanes). Hence these could be either way attributed to the segment or access points. Therefore for angle / turning related crashes the ternary variable 'element' takes the form of binary 'element1' where the crashes either belong to the signalized intersection or to segment/ access points. This new variable appears in certain tree results (developed along with conditional inference forests for relevant inference) and also positively contributes to model development in the forests.

Zhang, Lindsay, Clarke, Robbins, and Mao (2000) found that non use of seat belts increased the risk of severe injuries. In this study, the parameter for safety equipment in use is for all the passengers. This is different from the traditional approach as it is more useful to look at the overall safety of all the passengers rather than just focusing on the safety equipment use of the driver. The importance lies in the fact that there are a lot of crashes in which the drivers may not be injured at all. The vulnerable age group binary variable points out the presence of children or elderly passengers inside the vehicle. The physical fragility of the people belonging to these age groups described in Table 1 makes it an interesting variable and the results also show interesting pattern related to severity.

The median types were combined into nine levels. It does the two fold job of not only giving a sense of the median obstruction imposed but also gives an idea as to how far apart the opposing directional roads could be. The authors observed that the median width was a variable that is really dependent on the median type. Hence the median width was sufficiently represented within the variable median type. A new variable called 'shld_side' has been created that simply represents the total width of the outside shoulder and the sidewalk. This variable gives a more realistic idea of the side space available for the vehicles traveling in the outer lane, especially in the urban areas where the shoulder width sometimes is negligible as compared to those available in rural settings. Hence, the original information on shoulder width and the sidewalk width were replaced with this new variable.

The target variable of severity is binary. The first level represents fatalities and incapacitating injuries. They are combined into one level for two reasons; first, the relatively small frequency of fatal crashes compared to other injury severity levels. For example, the Chi square tests may not be valid due to low expected cell frequency. The second reason is that the crashes that involve incapacitating injury could easily have been fatal and vice versa possibly due to vulnerability of the subjects involved (Das et al., 2008). The second level includes crashes with possible injuries and non incapacitating injuries. The crashes with no injuries were not included as these are likely expected to be incomplete. This issue has been well investigated and documented by Abdel Aty and Keller (2005). Yamamoto, Hashiji, and Shankar (2008) also have discussed the issue of possible under reporting of such crashes and the bias resulting from it. Hence, in the present study the authors have included those crashes with injury severity level of at least a possible injury or higher.

It should be noted here that the conditional inference forests, which have been used to calculate the variable importance score, do not accept missing values. Hence, the data set has no missing data. Hence the introduction of random parameters to account for missing data, as done by Milton et al. (2008), is not required in this study. As mentioned earlier the crashes have been grouped into five types, namely: (a) angle/ turning movement; (b) rear end; (c) head on; (d) sideswipe; and (e) crashes involving single vehicles. The number of crashes in each of the crash categories are 6,231, 5,532, 1,261, 2,204, and 2,404, respectively, for the models developed for environmental and roadway geometric factors, whereas for the models developed for driver and vehicle related factors the number of crashes are 7,759, 6,775, 1,583, 2,612, and 2,879, respectively. As no missing data record could be used, the records deleted for the environmental and roadway geometric factors' models are 31,973. This accounts for 6.6% of the three years of Florida crash data used. Similarly, for the driver and vehicle related factors' models, the crash records that were deleted were 27,997, which accounts for 5.8% of the three years of Florida crash data used.

3. Modeling methodlogy

3.1. Conditional inference trees

The modeling approach adopted herein is the conditional inference trees and the forests developed there from. The focus of the study is to find out parameters that are related to the injury severity. The trees not only give the variables of importance, but also help us to better interpret the results. In severity analysis the advantage in using trees is that it helps us determine the values of parameters that contribute more to the severity of crashes. Hence from a safety aspect this is critical as it can help determine what changes need to be made in the design and/ or policies to improve the safety. Conventional classification and regression trees have always been used to select variables of importance. According to Strobl et al. (2007), the CART trees have a variable selection bias toward variables that are continuous or with higher number of categories. The most common splitting criterion in the CART tree is the Gini Index to find a favorable split. The Gini Index checks for the purity of the resulting "daughter" nodes in the tree. According to Breiman, Friedman, Olshen, and Stone (1984), for a given node't' with estimated class probabil ities 'p(j|t)', j = 1, ….., J, the node impurity 'i(t)' is given by:

$$i(t) = \Phi(p(1|t), \ldots\ldots, p(J|t)) \tag{1}$$

A search is made for the most favorable split, one that reduces the node or equivalently tree impurity. If the adopted form is Gini diversity index then 'i(t)' takes up the form:

$$i(t) = \sum_{j \neq i} p(j|t)p(i|t) \tag{2}$$

The Gini index considered as a function '$\Phi(p_1,\ldots\ldots,p_J)$' of the p_1, ……,p_J is a quadratic polynomial with nonnegative coefficients. Therefore for any split 's,' '$\delta(s, t) >= 0$.' Since the criteria looks for a favorable split, the chances to find a good split increases if the variable is continuous or has more categories. Therefore even if the variable is not informative, it could sit higher up on the tree's hierarchical structure. Hence in this study the researchers have used conditional inference trees (Hothorn et al., 2006) where the node split is selected based on how good the association is. The resulting node should have a higher association with the observed value of the dependent variable. The conditional inference tree uses a chi square test statistic to test the association. Therefore, it not only removes the bias due to categories but also chooses those variables that are informative.

The key to this recent algorithm is the separation of variable selection and splitting procedure. The recursive binary partitioning that is the basis of the framework is given below.

The response 'Y' comes from sample space 'Y,' which may be multivariate. The m dimensional covariate vector $X = (X_1,\ldots,X_m)$ is taken from a sample space $X = X_1,*\ldots\ldots*X_m$. Both the response variable and the dependent variables may be measured at any arbitrary scale. The conditional distribution of the response variable given the covariates depends on the function of the covariates.

$$D(Y|X) = D(Y|X_1, \ldots\ldots, X_m) = D(Y|f(X_1, \ldots\ldots, X_m) \tag{3}$$

For a given learning sample of 'n' iid observations a generic algorithm can be formulated using nonnegative integer valued case weights $w = (w_1,\ldots,w_n)$. Each node of a tree is represented by a vector of case weights having nonzero elements when the corre sponding observations are elements of the node and are zero otherwise. The generic algorithm is given below:

(1) For case weights w the global null hypothesis of independence between any of the covariates and the response is tested. The step terminates if the hypothesis cannot be rejected at a pre specified nominal level 'α.' Otherwise the j^{th} covariate X_j with the strongest association to the response variable is selected.
(2) Set $A \subset X_j$, is chosen to split X_j, into two disjoint sets. The case weights w_{left} and w_{right} determine the two subgroups with $w_{left,i} = w_i I(X_{ji} \in A)$ and $w_{right,i} = w_i I(X_{ji} \notin A)$ for all $i = 1,\ldots,n$ and I( ) denotes the indicator function, which indicates the membership of an element in a subset.
(3) Recursively repeat the steps 1 and 2 with modified case weights w_{left} and w_{right}, respectively.

The separation of variable selection and splitting procedure is essential for the development of trees with no tendency toward covariates with many possible splits. For more details of the algorithm the readers are directed to the paper by Hothorn et al. (2006).

3.2. Conditional inference forest

Forests that are a collection of multiple tree classifiers are used for variable selection. A decision tree, with all its simplicity and handling of missing values, can be very unstable. In other words, small changes in the input variables might result in large changes in the output. In this regard, forests are more robust variable selection tool. Random Forests' algorithm was developed by Breiman (2001), which works in the framework of the classification and regression trees, but instead of having one tree, they have multiple trees. The forests are most important in calculating the variable importance measure. Recent works in transportation by Abdel Aty et al. (2008) and Harb, Yan, Radwan, and Su (2009) used the random forests algorithm to determine the variables of importance. However Strobl et al. (2007) showed that the bootstrapping method (sampling with replacement) and the use of Gini index results in the biased selection of variables of importance. The Gini index shows a strong preference for variables with many categories or for the ones that are continuous. Variables with more potential cut off points are more likely to produce a good criterion value by chance. This variable selection bias that occurs in each individual tree also has an effect on the variable importance measure. In the previous sub section it was mentioned that the algorithm for recursive binary partitioning uses the association tests like chi square test to select informative variables. Therefore boot strap sampling with replacement induces bias because the cell counts in the contingency table are affected by observations that are either not included or are multiplied in the bootstrap sample. Hence the forests that we have used in this study comprise of the trees that have developed in the conditional inference framework. The next subsec tion describes the variable importance computation process.

3.3. Variable importance

The basis of the variable importance in forests is as follows. By first randomly permuting the predictor variable X_j, the original association with the response variable Y is broken. When the permuted variable along with other non permuted variables is used to predict the response for the out of bag observations the classification accuracy decreases substantially if the permuted variable is associated with the response. Hence the variable importance of a variable is the difference in the prediction accuracy before and after permutation of the variable X_j, averaged over all trees. Out of bag observations are those that the method excluded while developing the trees. They form an internal test data set and there is no need to allocate a test data set separately. Let $B^{(t)}$ be the out of bag sample for a tree't', with $t \in \{1,\ldots,ntree\}$. The variable importance of one tree is then given by the following:

$$VI^{(t)}\left(x_j\right) = \frac{\sum_{i \in B^{(t)}} I\left(y_i = \hat{y_i^{(t)}}\right)}{|B^{(t)}|} - \frac{\sum_{i \in B^{(t)}} I\left(y_i = \hat{y_{i,\pi_j}^{(t)}}\right)}{|B^{(t)}|} \quad (4)$$

Where $\hat{y}^{(t)} = f^{(t)}(x_i)$ is the predicted classes for observation 'i' before and $\hat{y}^{(t)}_{i,\pi_j} = f^{(t)}\left(x_{i,\pi_j}\right)$ is the predicted classes for observation 'i' after permuting its value of variable. The raw variable importance score for each variable is then computed as the mean importance over all trees and is given by:

$$VI\left(x_j\right) = \frac{\sum_{t=1}^{ntree} VI^{(t)}\left(x_j\right)}{ntree} \quad (5)$$

Since the individual importance scores $VI^{(t)}(x_j)$ are computed from 'ntree' independent bootstrap samples, a simple test for the relevance of variable X_j can be constructed based on the central limit theorem for the mean importance of $VI^{(t)}(x_j)$. If individual importance has a standard deviation σ, then the mean importance from 'ntree' replications has a standard error of $\sigma / \sqrt{ntree}$.

The next section emphasizes on the results of the random forests results for the various severity models developed on the urban/sub urban and rural corridors according to the various crash types.

Table 2 (a)
Conditional Inference Forest sample result for environmental and roadway geometric factors.

Variable Name	Variable Importance Score
Shoulder + Side	0.000358
Pavement condition	0.00026
Median Openings	0.000163
Median type	0.000163
Truck factor	0.00013
Vision obstruction	6.50E-05
Skid (friction resistance)	6.50E-05
Roadway condition	0
Horizontal Degree of Curvature	0
Surface condition	0
Parking type	0
Traffic-way character	0
Surface width	-9.76E-05
K factor	-6.50E-05
Day of the week and time of the day	-6.50E-05
Surface type	-3.25E-05
Daylight condition	-3.25E-05
Roadway element	-0.00013
Maximum posted speed limit	-0.00026
ADT	-0.00029
Shoulder type	-0.00036

4. Analysis and results

4.1. Conditional inference forest variable importance results

This section deals with the results of conditional inference forests that typically illustrate the variables of importance. In the present study the conditional inference forests generated for the models, with the binary severity variable as the target, gives the variable importance score for all the variables in the model. The sign (positive/ negative) of the importance score indicates whether the presence or absence of a variable in the model will improve or degrade the efficiency of the model. To exemplify the variable importance score the authors tabulate the results for a particular cluster (in this case Cluster 3 for angle/ turning movement crashes) in Tables 2a and 2b. As mentioned earlier in the section *Data Collection and Preparation,* the variables have been categorized into two. Hence for each cluster and crash type two models had been developed, one for the environmental and roadway geometric and the other for driver and vehicle related characteristics. Results in Table 2a are for the model with only environmental and roadway geometric factors and those in Table 2b are for the driver and vehicle related characteristics' model. As a reminder to the readers, Table 1 has the explanation of the variables.

It should be noted that Tables 2a and 2b are examples of the output of a condition inference forests. The variables with a positive variable

Table 2 (b)
Conditional Inference Forest sample result for driver and vehicle related factors.

Variable Name	Variable Importance Score
Alcohol usage	0.004544
Age group	0.004488
Vehicle movement	0.000309
Safety equipment use	0.00014
Vehicle type	5.61E-05
At fault driver gender	2.81E-05
Vulnerable age group	2.81E-05
Presence of more than 5 persons	0

Table 3 (a)
Severity models' conditional inference forests results for urban clusters with environmental and roadway geometric factors.

Variable	Cluster 1		Cluster 2		Cluster 1 & 2				Cluster 3					Cluster 4					
	angle	rearend	angle	rearend	headon	sideswipe	single	slow	angle	rearend	headon	sideswipe	single	slow	angle	rearend	headon	sideswipe	single	slow
surface_width	+	0	-	-	0	0	-	+	-	0	+	0	-	+	+	-	+	0	+	+
max_speed	-	0	+	-	-	0	-	-	-	+	-	0	-	-	+	0	0	0	-	+
LIGHTCDE	+	0			0	0	0	0							0	0	0	-	0	0
ACMANCLS_num	+	0	+	0	+	-	-	0	+	0	-	0	-	0	+	-	+	0	0	+
road_cond	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0	0	0	0
vision	0	0	0	0	0	0	0	0	+	0	+	0	0	+	+	0	-	-	0	0
shld_side	+	0	+	0	0	+	-	-	+	0	0	0	0	+	+	-	0	0	+	+
curvlass	-	0	0	0	0	0	0	0	0	0	0	0	-	0	+	0	0	0	0	0
surf_cond	+	0	-	0	-	0	-	-	0	0	-	0	-	+	+	0	0	+	0	-
light	-	0	0	0	0	0	+	0	-	0	+	0	-	+	0	-	0	-	0	+
ADT	+	0	+	+	-	0	-	+	-	+	+	0	-	+	+	+	0	0	-	+
t_fact	+	0	-	0	-	0	-	+	+	+	-	0	-	+	-	+	+	0	+	+
k_fact	+	0	+	0	-	-	0	0	-	0	-	0	+	-	+	0	0	0	-	+
dayandtime	-	0	+	0	0	-	+	0	-	-	-	0	-	0	+	-	0	0	+	+
trfcway	0	0	-	-	0	0	+	0	-	0	+	0	-	+	+	0	0	0	0	+
pavecond	+	0	-	0	+	0	+	+	+	0	0	0	+	0	+	-	+	0	+	-
park	+	0	0	0	0	0	0	0	0	0	0	0	0	-	0	-	0	0	0	+
surf_type	0	0	-	0	-	0	0	0	-	0	0	0	0	+	0	+	0	0	0	+
skid_f	+	0	+	0	+	0	+	+	+	-	+	0	0	+	+	0	+	0	-	+
median	+	0	+	-	+	0	0	+	+	0	+	0	-	+	+	+	+	-	+	+
element1	+	0	+	0	-	0	-	0	-	+	+	0	+	+	,	+	+	0	-	+
shld_t	+	0	,	0	,	0	+	0	,	0	,	0	,	+	+	0	0	,	,	+

importance score are the most important for the severity model developed here in the example. Their association with the target variable is the maximum and their absence would decrease the model performance. The variables with zero importance score are believed to have no effect on the model performance, while the ones with negative importance (as highlighted in Table 2a) are the ones decreasing the model performance. Readers may note that the variable *LIGHTCDE* has not been included in Table 2a; this is because it had only one level and can not be used for split during tree development. The same is the reason for *no information* on *LIGHTCDE* in some of the cells of Table 3a as well. It is critical to distinguish the significant from non significant. As the dataset change (i.e., a new model is being developed), the importance score may also change. A particular variable may improve the model efficiency in one group whereas it may decrease in another group, while being neutral in some other. All the conditional inference forests results were developed at 90% confidence level.

Tables 3a and 3b tabulate the conditional inference forests results developed for all severity models in the study. For certain types of crashes (i.e., head on, sideswipe, single vehicle involved, slow moving vehicles involved) the number of crashes in the urban clusters 1 and 2 were not sufficient for the trees to develop. Hence for these types of crashes the clusters 1 and 2 were combined. All the results were developed with the use of the statistical software package 'R.' The package 'party' developed by Hothorn et al. (2008) was used to generate the conditional trees and forests results. Key for Tables 3a and 3b is:

'+': variables which increase the model efficiency,
' ': variables which decrease the model efficiency,
'0': variables which are neutral to model efficiency.

It should be noted that in Tables 3a and 3b there could be certain blank cells (i.e., they do not have any of the three symbols mentioned above). For example, the variable *LIGHTCDE* does not appear in Table 3a in a number of cells. The reason for the exclusion is that the variable was not used for that particular model development, as it had only one level.

As mentioned earlier, the variables with "+" sign in the boxes are the variables with higher importance (i.e., they improve the model efficiency more than the other variables for the given model). The ones with "0" means they are neutral for the severity model. The variables with " " are the ones with least effect on the corresponding model. It must, however, be understood that the "+" sign need not necessarily mean that the variable is positively associated with severity. For better interpretation of the variable's influence on the severity, single conditional inference trees were developed for the models. And depending on how the variables split, the approach to severe/fatal crashes would be clearer. The next subsection deals with the individual conditional inference tree results.

5. Conditional inference trees results

5.1. Example of conditional inference trees and how to interpret them

The conditional inference trees are critical to observe which parameters are related more to severity and also how they are related. Before moving to the details of the results, the authors would like to exemplify certain individual conditional inference tree results through Fig. 1(a) and (b). The trees shown in the figures are for angle/ turning movement crashes in Cluster 1. Fig. 1(a) represents the tree model for environmental and roadway geometric factors, whereas Fig. 1(b) is the model for driver and vehicle related factors.

All the trees were developed at 90% confidence level. The *p* value in the nodes of Fig. 1(a) and (b) denotes the actual significance level at

Table 3 (b)
Severity models' conditional inference forests results for urban clusters with driver and vehicle related factors.

Variable	Cluster 1		Cluster 2		Cluster 1 & 2				Cluster 3						Cluster 4					
	angle	rearend	angle	rearend	headon	sideswipe	single	slow	angle	rearend	headon	sideswipe	single	slow	angle	rearend	headon	sideswipe	single	slow
age_gr	-	0	+	0	-	0	-	+	+	+	+	-	-	+	+	0	+	+	-	+
veh_typel	+	-	+	0	0	0	+	-	+	0	+	-	+	+	+	0	-	-	+	+
alcohol_use	-	-	+	0	0	0	+	0	+	-	+	0	+	+	+	0	+	0	+	-
more	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sfty	+	0	+	0	0	0	+	+	+	0	-	0	+	+	-	0	0	0	-	+
gender	+	0	+	0	+	0	-	0	0	+	+	0	+	-	-	0	-	+	+	0
veh_move1	-	0	+	0	-	0	+	+	-	+	0	-	-	+	+	0	-	-	+	+
vuln_age	0	0	+	0	0	0	0	0	-	0	0	0	-	0	+	0	0	0	0	-

which the split has taken place. All the nodes are shown in white oval shape, whereas all the terminal nodes (leaves) are shown in the rectangular boxes. The small square boxes with numbers on both the ovals and rectangles denote a unique numerical representation of the node or leaf. In the white oval shapes the variables mentioned is the split variable and the *p* value denotes the significance level. The numbers on the lines connecting the nodes to other nodes or leaves denotes the specific categories of variables or range of values of variables that lead to the extension of that particular branch of the tree. For example, in Fig. 1(b) the variable *alcohol use* splits the node and all the cases of the variable taking up the value 1 (denoting no alcohol/ drug use) leads to the leaf, which is uniquely numbered as '2.' For the other branch the variable either takes the value 2 or 3 (denoting alcohol use or pending test results) to reach the other leaf, uniquely numbered as '3.' The general direction of flow of the lines in any conditional inference tree is top to bottom. It goes from one node to other node/ leaf. As can be observed the leaf contains the information about the number of cases in the particular leaf, denoted by *n*. The proportion of non severe and severe crashes is also shown in the leaf, through the numbers given by *y*. To exemplify, the authors again refer to Fig. 1(b). The leaf, uniquely denoted by '2' has $n = 1{,}849$ cases, whereas the proportion of non severe crashes was 0.851 while that of severe crashes was 0.149 (denoted by $y = (0.851, 0.149)$). As can be observed from Fig. 1(a) and (b), there are red ovals covering certain leaves. These leaves have higher proportion of severe crashes than the proportion of severe crashes in the particular dataset from which the model had been developed. The path taken from the original parent node to the particular leaf thus gives us the conditions that lead to higher severity. The variables on the path, on which the splits have been done, reflect which variables are associated with severity.

From here on the results will be based on crash type and the relevant results from different clusters will be grouped together. The explanation will include both the categories of models developed, namely: (a) environmental and roadway geometric and (b) driver and vehicle related. The order will be adhered to for the most part of the explanation.

5.2. Angle/ turning movement crashes

As mentioned earlier the corridors in Cluster 1 (1.009 2.89 miles) are the smallest in length. According to the environmental and roadway geometric model for angle/ turning movement crashes occurring in this particular cluster's corridors, the severity is higher where the shoulders are paved and the k factor is higher. Even though paved shoulders leading to higher severity seems counterintuitive, the only reason could be that better shoulders may be misused as additional lanes for dangerous maneuvers. The higher k factor indicates that the higher the peak hour volume, the higher risk it involves for angle/turning movement crashes. With lower k factor but restrictive medians (with longer distance between openings), the severity of the crashes is found to be higher. Since angle/ turning movement crashes mostly occur at intersections, it is interesting to note that Levinson (2000) pointed out that even though restrictive medians provided better separation of traffic and better pedestrian safety, however adequate provisions have to be made for left and U turns to avoid an overwhelming increase in movements at the intersections. Lack of adequate left or U turns could be one of the reasons why this result was observed. For the same cluster alcohol/drug use is also found to be associated with severe/fatal crashes in the model for driver and vehicle related factors. The authors in a previous study (Das et al., 2008) found similar results for alcohol/ drug use. Wang and Abdel Aty (2008) found an increasing effect of alcohol/ drug use in severity of crashes. In Cluster 2 (2.898 5.729 miles) for the environmental and roadway geometric model, posted speeds greater than 45 mph are found to be riskier. In a recent study by Malyshkina and Mannering (2008), they found higher posted speed limit to be associated with higher severity of injuries. For corridors where the posted speed limits are less than 45 mph and high k factor, conditions are suitable for

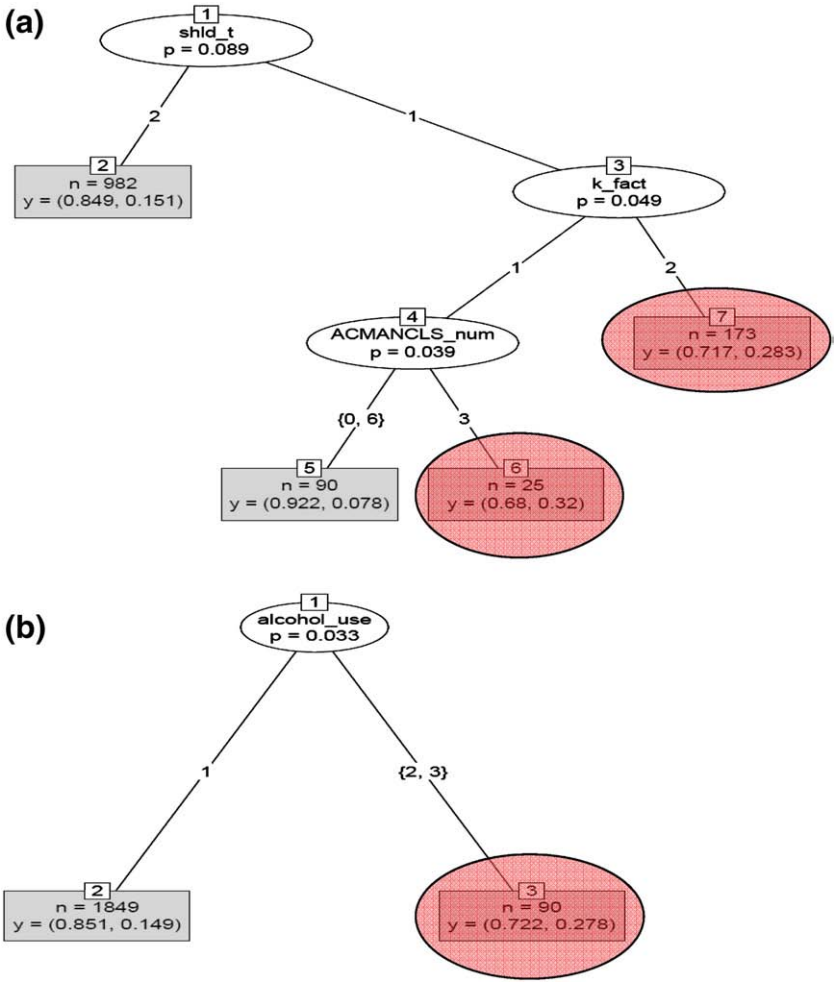

Fig. 1. (a): Conditional Inference Tree sample result for environmental and roadway geometric factors. Fig. 1(b): Conditional Inference Tree sample result for driver and vehicle related factors.

crashes with higher injury severity. In the driver and vehicle related factors' model, failure to use safety equipment and alcohol/drug use also lead to severe/fatal crashes. Though much research highlights the seat belt use and its obvious benefits (Evans, 1996; Derrig, Segui Gomez, Abtahi, & Liu, 2000; Eluru & Bhat, 2007), very few discuss the effects of other safety equipment in use inside the vehicle in general. Likewise for Cluster 3 (5.762 10.556 miles) corridors, the model for environmental and roadway geometric reflects that posted speeds of greater than 50 mph leads to higher severity, while the model for driver and vehicle related factors show that the non use of safety equipment and alcohol/ drug use again lead to crashes that are more at threat to be severe. However, for Cluster 4 (10.644 78.293 miles) corridors, the two models (environmental and roadway geometric factors' model; and driver and vehicle related factors' model) were developed at only 70% and 75% levels of confidence, respectively. Hence, the results are not reported here. Summarizing the results reflects that angle/turning movement crashes are more severe under high speeds, when no safety equipment is in use, and while driving under the influence. The results are consistent with the common perception.

5.3. Rear end crashes

The environmental and roadway geometric factors' model for the rear end crashes in Cluster 1 suggests that higher friction resistance (*skid*>38) leads to higher severity of injuries given the crash has occurred. This is counterintuitive as higher friction should be better at preventing severe crashes. The result could provide insight to the phenomenon that when the friction is higher and the vehicles can brake within shorter distances, the internal movement could be sudden and any internal/secondary collision (i.e. passengers hitting something inside the vehicle) could lead to a severe injury. In the model for driver and vehicle related factors it was observed that the severe/fatal crashes are linked to light, slow moving vehicles like cycles and mopeds. The higher severity level is intuitive, as any crash with light vehicles will generally be severe. Huang, Chor, and Haque (2008) found similar results in their investigation of traffic crashes at intersections. The environmental and roadway geometric model for Cluster 2 corridors indicate that the posted speed limit of greater than 50 mph leads to severe rear end crashes. Similarly, in a recent technical report developed for NHTSA by Liu and Chen (2009) it was observed that severe crashes are more likely to occur at corridors with posted speed limits of 50 mph or greater. On the other hand when the speeds are less than 50 mph, crashes will be severe/ fatal when the k factor is high. For the same Cluster 2 corridors, alcohol/ drug use leads to crashes that are severe/ fatal as shown by the driver and vehicle related factors' model. It is observed that when there is no alcohol/ drug use by the responsible driver, the presence of a person in the vulnerable age group (>55 yrs or <3 yrs) makes the crash more severe in general. While alcohol/ drug use is a case of irresponsible driving behavior, the presence of a person in the vulnerable age group is a clear case of physical fragility. People in the vulnerable age group always tend to experience severe injuries resulting out of a crash. The authors would like to reference a particular case reported by Batra and Kumar (2008) in which an 84 year old man succumbs to injuries

resulted in a low velocity collision. In this particular case the injury was a subaxial cervical spinal cord injury that was triggered by the airbag deployment, and interestingly the driver was not wearing a seat belt. The authors cite this particular example as it was observed that under relatively slower speeds (<50 mph) severe injuries can occur if the safety equipment is not properly used and it also confirms the observation that the presence of a person in the vulnerable age group will succumb to injuries that become more apparent due to physical fragility. On Cluster 3 corridors, lower ADT leads to higher severity crashes while the driver and vehicle related factors' model indicate alcohol/ drug use leads to severe/fatal crashes. Lower ADT could mean higher speeds that more often lead to severe/ fatal crashes. For even longer corridor groups (i.e., Cluster 4), higher friction resistance (skid>34) leads to severe rear end crashes by the environmental and roadway geometric factors' model. The explana tion was given at the beginning of this subsection. For lower friction resistance, greater surface widths (corresponding to 3 or more lanes per direction) and the presence of median curb increase the severity level of crashes. The increase in surface width should traditionally reduce severity (Petritsch, Challa, Huang, & Mussa, 2007), however this result might seem counter intuitive. This could be explained in the following way. Higher surface width may result in higher speeds and more driver comfort, which might cause some drivers to be less cautious. Hence the increase in speeds and less attention by the drivers could lead to crashes with severe injuries. Das et al. (2008) had found similar results. On the same corridor group, older drivers (>55 yrs) also are involved in severe rear end crashes. The longer the corridors, the more the exposure of the driver and the older the driver the more prone is he/she to make an error. Marshall (2008) states that prevailing medical conditions and impairments associated with old age leads to deteriorating fitness and hence to higher crash risk for the older driver.

5.4. Head on crashes

For head on type of crashes on corridors belonging to Clusters 1 and 2 combined, crashes on dry surface condition were found to be more severe/ fatal from the environmental and roadway geometric model. However, the model for driver and vehicle related factors was developed at a lower confidence level of 70%, hence the results are not reported here. Dry surface conditions probably indicate fine weather and more vehicles on the road. Hence improper maneuvers could result in head on collisions, especially when the highways are undivided, resulting in severe crashes. In a related study by Yan, Harb, and Radwan (2008) it is shown that slippery road conditions lead to a higher probability of crash avoidance maneuvers as drivers will drive more cautiously during unfavorable conditions. Hence, the results in this study indicate that drivers could be less attentive when driving in good weather and road conditions. In Clusters 3 and 4, alcohol/drug use is the primary reason for severe head on crashes.

Table 4 (a)
Significant factors for Angle/ turning movement crashes.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Environmental and Roadway Geometric Factors	Paved shoulders and k factor>9.85; Paved shoulders and k factor<9.85 and restrictive median	Posted speed limit>45 mph; posted speed limit<45 mph and k factor>9.85	Posted speed limit>50 mph	No significant results
Driver and Vehicle Related Factors	Alcohol/ drug use	Non-use of safety equipment and alcohol/ drug use	Alcohol/ drug use	No significant results

Table 4 (b)
Significant factors for Rear-end crashes.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Environmental and Roadway Geometric Factors	Skid resistance>38	Posted speed limit>50 mph; posted speed limit<50 mph and k factor>9.85	Lower ADT (<31,000)	Skid resistance >34; Skid resistance <34 and surface width >32 ft and presence of median curb
Driver and Vehicle Related Factors	Light slow moving vehicles	Alcohol/drug use; No Alcohol/ drug use and presence of person in the vulnerable age group (>55 yrs or <3 yrs)	Alcohol/ drug use	Older drivers >55 yrs

5.5. Sideswipe crashes

In sideswipe crashes, restrictive medians are more threatening on shorter corridors (Cluster 1) as shown by the environmental and roadway geometric model. While on longer corridors (Cluster 3), straight ahead movement is crucial as observed from the driver and vehicle related factors' model. For all other types of movements, severe sideswipe crashes occur when slow moving vehicle type and light trucks are involved. Research by Anderson (2008) indicates that the increase in the light truck traffic increases the number of fatalities on the road. In the same work it was indicated that up to 80% of the increased deaths can be assigned to occupants in other vehicles and pedestrians. For severe/fatal sideswipe crashes involving slow moving vehicles, turning movements along with changing lanes are the significant parameters on Cluster 3 corridors. The more the lane changing maneuvers, the higher the probability of crash severity as many of the maneuvers will be risky.

5.6. Single vehicle crashes

For crashes involving single vehicles, higher friction factor leads to increased severity in crashes on shorter length corridors (Cluster 1 and 2 combined) according to the environmental and roadway geometric factors' model. On the other hand, the driver and vehicle related factors' model for the same corridors indicate straight vehicle movement related crashes are found to be more severe. For the single vehicle type of crashes occurring on Cluster 3 corridors that are related to segments or access points, the crashes tend to be more severe at stretches where the posted speed limits are 45 mph or greater. The driver and vehicle related factors' model shows that failure to use safety equipment in slow moving vehicles also leads to severe injuries in crashes. In Cluster 4 the crashes are more at risk to be severe when the posted speed limit is greater than 50 mph. The driver and vehicle related factors' model for this cluster indicate that slow moving vehicles (e.g., cycles, mopeds) tend to be involved in severe crashes. This could be explained by the fact that on corridors with 50 mph posted speed, slow moving vehicles pose a risk as they will create speed variance on the roadways. Collisions with slow vehicles would likely be severe.

Table 4 (c)
Significant factors for Head-on crashes.

	Clusters 1 and 2	Cluster 3	Cluster 4
Environmental and Roadway Geometric Factors	Dry surface condition	No significant results	No significant results
Driver and Vehicle Related Factors	No significant results	Alcohol/ drug use	Alcohol/ drug use

Table 4 (d)
Significant factors for Sideswipe crashes.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Environmental and Roadway Geometric Factors	Restrictive medians	No significant results	No significant results	No significant results
Driver and Vehicle Related Factors	No significant results	No significant results	Straight ahead movement of the vehicle; turning movements along with changing lanes and slow moving vehicles	No significant results

5.7. Results summary

The results discussed in the preceding subsections are summarized in Tables 4a 4e. The variables in the cell represent those that increase severity along with the range or categories. The blank cells indicate that the results could not be developed with the 90% confidence level. These tables will help the reader to have a comparative understanding of the variables entering a particular tree model and how they affect safety. Tabulating the results helps to better understand the results; particularly in this study where the results are brought together and compared across crash types and corridor clusters.

6. Conclusions

The application of conditional inference trees and forests leads to the identification of an unbiased set of variables significantly related with severity. The advantage of the new algorithm of tree/forest develop ment over the traditional CART tree/forest is that it prevents the uninformative variables from being identified as significant just by the virtue of having a higher number of categories or being continuous in nature. The novel way of separating the split criteria from the variable importance selection while developing a tree is what makes the conditional inference trees unique. The chi square test is used to determine the strength of association with the target variable, in the present application it is the binary severity variable. Once a variable is selected at a particular tree level for split, the split can then be decided based on any criteria, including those used in the CART algorithm. The conditional inference forests on the other hand calculates individual variable importance of each variable for every tree by first breaking the association with permutation and then testing the tree with out of bag estimates. In the forests, the variable importance is based on the result from multiple trees, thus avoiding the instability of individual trees.

Among the results from the analysis, alcohol/ drug use is asso ciated with increased severity of crashes irrespective of the length of the corridors or the type of crashes. Since the drivers are less likely to be in control, it invariably leads to severe crashes. Failure to use safety equipment has lead to increased severity of single vehicle as well as angle/turning movement related crashes. In this regard, conclusions drawn by Abdel Aty and As Saidi (2000), by analyzing the zip codes of the offenders for better targeting the education programs, may be of renewed interest. Older at fault drivers are found to be more at risk of getting involved in a severe crash, especially in a rear end collision on longer corridors. On similar corridors, a crash is more likely to have a severe injury where there is a person in the vulnerable age group (more than 55 years or less than 3 years).

Table 4 (e)
Significant factors for Single vehicle crashes.

	Clusters 1 and 2	Cluster 3	Cluster 4
Environmental and Roadway Geometric Factors	Skid resistance >38	Crashes related to segments and/ or access points and posted speed limit>45 mph	Posted speed limit >50 mph
Driver and Vehicle Related Factors	Straight ahead movement of the vehicle	Non-use of safety equipment and slow moving vehicles	slow moving vehicles

Slow moving vehicles like cycles and mopeds have been observed to be involved in severe injury crashes. Many of these severe crashes occur at signalized intersections. It indicates that the designs of the intersections need to improve with respect to the slow moving vehicle and possibly even pedestrians. For shorter length corridors, higher k factor is a significant parameter for increased severity crashes. Higher k factor essentially means that the corridor is designed for handling higher volume during peak hour. It in turn has the potential not only to reduce rear end crashes during the peak hour (due to improved congestion situation), but also to increase speeds due to better design during off peak periods. Since rear end crashes tend to be less severe, higher k factor leads to increased likelihood of severe crashes.

On longer corridors, like those in Cluster 3, severity of rear end crashes increases when the posted speed limit is greater than 50 mph. Lowering the posted speed limit may not be the best strategy from an operations point of view, but it may lead to reduction in severity of crashes. Lower ADT also leads to severe rear end crashes on Cluster 3 corridors, especially for rear end crashes. Severe/fatal crashes involving single vehicles are more likely to be associated with access points on longer corridors. Reducing the number of access points may not always be feasible; however, design changes such as improved merging may be adopted for these issues.

Corridors of smaller lengths (generally less than 5 miles) have been observed to have problems of increased severity if crashes occur on corridors with high skid resistance values. Shorter corridors also have problems when the posted speed limit is greater than 45 mph. Since most of these small urban/ suburban corridors are located between longer stretches of rural corridors, they have lower speed limits compared to adjacent sections. However, since the congestion is not high on the rural sections, some drivers will tend to speed and thus create a larger variation in prevailing speeds. This variation could lead to more severe crashes on shorter length corridors. Restrictive median openings on shorter corridors have also been found to be problematic. The variable indicating the presence of vulnerable age group also came out significant on shorter corridors rather than on longer corridors. On longer length (greater than 5 miles) corridors, speed limits of greater than 50 mph are a cause of concern. Non use of safety equipment is also more pronounced in contributing toward severity on longer corridors. In a recent paper by Eluru and Bhat (2007) the question of the endogenous relationship between seat belt use and injury severity is raised. There is a possibility of intrinsically unsafe drivers not wearing the seat belt and being the ones to be likely involved in high injury severity crashes because of their unsafe driving habits. In the present study, however, the researchers observe the overall safety equipment in use in the vehicle. Results also show that the failure to use the safety belt in single vehicle crashes and crashes involving a slow vehicle lead to higher severity crashes. Thus the present study is not only in line with concurrent research but also goes a step further in identifying the type of crashes that are more likely to be affected by the underlying endogenous relationship.

Due to these observed differences, the decision to cluster the corridors has been justified. The subtle differences are highlighted when the groups are logically made. The clusters that were originally made based on the length actually shed light on the factors and a lot of new significant variables come into the picture.

The results from the forest and the trees are intuitive and their association with severity may be explained. Certain known results about severity of crashes have been confirmed, while some new information is discovered about others. Alcohol/ drug use along with higher speed limits tend to result in more severe/fatal crashes. The new variable "element," which uses information from site location, signal type in formation, and traffic control was also insightful in identifying locations

that are more critical from the severity aspect. Drivers of vehicles with passengers in the vulnerable age group ranges must also be more careful while driving, as the physical fragility of these subjects tends to make the injuries more severe. The authors also used the safety information for all passengers seated in the car. That particular variable also was significantly associated with severity of crashes. Hence, it is critical that internal safety should be a concern for the law enforcement agencies if they are intended to reduce the occurrences of severe/fatal crashes on the arterials of Florida.

Acknowledgements

The authors wish to thank the Florida Department of Transportation for funding this research. They would also like to thank Mr. Ali Darwiche for his help in the data extraction and preparation.

References

Abdel-Aty, M., & As-Saidi, A. H. (2000). Using GIS to locate the high risk driver population. *Swedish National Road and Transport Research Institute*, 111 126.

Abdel-Aty, M., & Keller, J. (2005). Exploring the overall and specific crash severity levels at signalized intersections. *Accident Analysis & Prevention*, *37*(3), 417 425.

Abdel-Aty, M., Pande, A., Das, A., & Knibbe, W. J. (2008). Analysis of infrastructure based ITS data for assessing safety on freeways in Netherlands. *Journal of the Transportation Research Board*, *2083*, 153 161.

Abdel-Aty, M., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, *32*(5), 633 642.

Abdel-Aty, M., & Wang, X. (2006). Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors. *Journal of the Transportation Research Board*, *1953*, 98 111.

Anderson, M. L. (2008). Safety for Whom? The Effects of Light Trucks on Traffic Fatalities. *Journal of Health Economics*, *27*(4), 973 989.

Batra, S., & Kumar, S. (2008). Airbag-Induced Fatal Subaxial Cervical Spinal Cord Injury in a Low-Velocity Collision. *European Journal of Emergency Medicine*, *15*(2), 52 55.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5 32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International Group.

Brown, H.C., & Tarko, A.P. (1999). *Effects of access control on safety on urban arterial streets. Transportation Research Record, 1665.* Washington DC: Transportation Research Board, National Research Council.

Das, A., Pande, A., Abdel-Aty, M., & Santos, J. B. (2008). Urban arterial crash characteristics related with proximity to intersections and injury severity. *Journal of the Transportation Research Board*, *2083*, 137 144.

Derrig, R. A., Segui-Gomez, M., Abtahi, A., & Liu, L. L. (2000). The effect of population safety belt usage rates on motor vehicle-related fatalities. *Accident Analysis & Prevention*, *34*(1), 101 110.

Eluru, N., & Bhat, C. R. (2007). A joint econometric analysis of seat belt use and crash – related injury severity. *Accident Analysis & Prevention*, *39*(5), 1037 1049.

Evans, L. (1996). Safety-belt effectiveness: the influence of crash severity and selective recruitment. *Accident Analysis & Prevention*, *28*(4), 423 433.

Harb, R., Yan, X., Radwan, A. E., & Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests. *Accident Analysis & Prevention*, *41*(1), 98 107.

Huang, H., Chor, C. H., & Haque, M. M. (2008). Severity of driver injury and vehicle damage in traffic crashes at intersections: A Bayesian hierarchical analysis. *Accident Analysis & Prevention*, *40*(1), 45 54.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651 674.

Hothorn, T., Hornik, K., & Zeileis, A. (2008). *A laboratory for recursive partitioning.* http://cran.r-project.org/web/packages/party/party.pdf Accessed February 28, 2008.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis.* New York: Wiley.

Levinson, H. (2000). Restrictive medians and two-way left turn lanes: some observations. *Third National Access Management Conference* (pp. 243 245). Federal Highway Administration: Washington, DC: Federal Highway Administration.

Liu, C., & Chen, C. L. (2009). *An Analysis of Speeding-Related Crashes: Definitions and the Effects of Road Environments. NHTSA Technical Report, 2009.* Washington, DC: National Center for Statistics and Analysis, NHTSA.

Malyshkina, N., & Mannering, F. L. (2008). Effect of Increases in Speed Limits on Severities of Injuries in Accidents. *Journal of the Transportation Research Board*, *2083*, 122 127.

Marshall, S. C. (2008). The Role of Reduced Fitness to Drive Due to Medical Impairments in Explaining Crashes Involving Older Drivers. *Traffic Injury Prevention*, *9*(4), 291 298.

Miaou, S. P., & Song, J. J. (2005). Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis & Prevention*, *37*(4), 699 720.

Milton, J. C., & Mannering, F. L. (1998). The relationship among highway geometrics, traffic related elements and motor-vehicle accident frequencies. *Transportation*, *25*(4), 395 413.

Milton, J. C., Shankar, V. N., & Mannering, F. L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis & Prevention*, *40*(1), 260 266.

National Highway Traffic Safety Administration [NHTSA]. (2007). *Traffic safety facts 2006: A compilation of motor vehicle crash data from the Fatality Analysis Reporting System and the General Estimate System.* Washington, DC: Author.

Pande, A., & Abdel-Aty, M. (2008). Discovering indirect associations in crash data using probe attributes. *Journal of the Transportation Research Board*, *2083*, 170 179.

Petritsch, T.A., Challa, S., Huang, H., & Mussa, R. (2007). *Evaluation of Geometric and Operational Characteristics Affecting the Safety of Six-Lane Divided Roadways.* Final Report. Sprinkle Consulting Inc., FDOT.

Rees, J. (2003). *Corridor management: identifying corridors with access problems and applying access management treatments, a U.S. 20 study.* Retrieved February 10, 2007, from http://www.ctre.iastate.edu/mtc/papers/2003/JREES.pdf.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*.

Wang, X., & Abdel-Aty, M. (2008). Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. *Accident Analysis and Prevention*, *40*(5), 1674 1682.

Wang, X., Abdel-Aty, M., & Brady, P. (2006). Crash Estimation at Signalized Intersections: Significant Factors and Temporal Effect. *Journal of the Transportation Research Board*, *1953*, 10 20.

Wang, X., Abdel-Aty, M., Nevarez, A., & Santos, J. B. (2008). Investigation of safety influence area for four-legged signalized intersections: nationwide survey and empirical inquiry. *Journal of the Transportation Research Board*, *2083*, 86 95.

Yamamoto, T., Hashiji, J., & Shankar, V. N. (2008). Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. *Accident Analysis & Prevention*, *40*(4), 1320 1329.

Yan, X., Harb, R., & Radwan, E. (2008). Analyses of Factors of Crash Avoidance Maneuvers Using the General Estimates System. *Traffic Injury Prevention*, *9*(2), 173 180.

Zhang, J., Lindsay, J., Clarke, K., Robbins, G., & Mao, Y. (2000). Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario. *Accident Analysis & Prevention*, *32*(1), 117 125.