ELSEVIER

Accident Analysis and Prevention 37 (2005) 417–425

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Exploring the overall and specific crash severity levels at signalized intersections

Mohamed Abdel-Aty*, Joanne Keller

Department of Civil and Environmental Engineering, University of Central Florida, Orlando, FL 32816-2450, USA

Received 19 September 2004; received in revised form 24 November 2004; accepted 29 November 2004

Abstract

Many studies have shown that intersections are among the most dangerous locations of a roadway network. Therefore, there is a need to understand the factors that contribute to injuries at such locations. This paper addresses the different factors that affect crash injury severity at signalized intersections. It also looks into the quality and completeness of the crash data and the effect that incomplete data has on the final results. Data from multiple sources have been cross-checked to ensure the completeness of all crashes including minor crashes that are usually unreported or not coded into crash databases. The ordered probit modeling technique has been adopted in this study to account for the fact that injury levels are naturally ordered variables. The tree-based regression methodology has also been adopted in this study to explore the factors that affect each severity level. The probit model results showed that a combination of crash-specific information and intersection characteristics result in the highest prediction rate of injury level. More specifically, having a divided minor roadway or a higher speed limit on the minor roadway decreased the level of injury while crashes involving a pedestrian/bicyclist and left turn crashes had the highest probability of a more severe crash. Several regression tree models showed a difference in the significant factors that affect the different severity types. Completing the data with minor non injury crashes improved the modeling results and depicted differences when modeling the no injury crashes.

Keywords: Signalized intersections; Crash severity; Ordered probit models; Tree-based regression; Traffic safety data

1. Introduction

1.1. Background

Intersections are a common place for crashes, which may be due to the fact that there are several conflicting movements as well as a myriad of different intersection design characteristics. Intersections also tend to experience severe crashes due to the fact that some of the injurious crashes such as angle and left turn collisions commonly occur at intersection. During 1999, there were 243,409 crashes recorded in the Florida Crash Database. Of these, 98,756 crashes (about 40%) occurred at or were influenced by a signalized intersection. About 9.6 crashes occur at signalized intersections per year compared to 2 per year where stop or yield signs control traffic. The factors affecting injury levels of crashes occurring at signalized intersections are not well understood. Therefore, there is a need to identify the effects that certain geometric and crash-specific aspects have on the injury level of crashes occurring at signalized intersection.

Furthermore, when a crash occurs and the local police department is notified, the responding officer will determine whether to fill out a long- or short-form crash report. For instance, if a crash involves an injury or a felony (e.g., hit and run), the crash must be filed on a long-form. If a crash involved only property damage (a minor crash with no injuries), usually it is up to the officer to report it on a long- or a short-form. Crash forms are then forwarded to the respective counties. From here, only the crashes reported on long-forms are forwarded onto the Florida Department of Transportation (FDOT) and the Department of Highway Safety and Motor

* Corresponding author. Tel.: +1 407 823 5657; fax: +1 407 823 3315.
E-mail address: mabdel@mail.ucf.edu (M. Abdel-Aty).

0001-4575/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2004.11.002

Vehicles (DHSMV), which maintain electronic records based on only crashes reported on long-forms. By focusing on state-maintained databases, most non injury crashes are neglected. Ignoring these non injury crashes will bias the distribution of crash injury severity levels. Therefore, for this study, two databases were considered. The first database consisted of crashes from the state agencies (reported only on long-forms) and was considered the restricted dataset. The second database consisted of all crashes (reported on both long- and short-forms) and was completed by obtaining all crashes reported on short-forms (we here refer to it as the complete dataset). Furthermore, multiple databases were cross-checked to ensure that the crashes reported on long-forms are as complete as possible (it is worth noting that even the complete data set does not include non reported crashes).

This study explores the hypothesis that crash injury levels are affected by both crash-specific and intersection-specific variables. Furthermore, the authors investigate the significant differences in the important crash-related factors between models based solely on crashes reported on long-forms and models based on crashes reported on both long- and short-forms (i.e., models based on restricted and complete datasets). Additionally, several databases were cross-checked to ensure the completeness of our data. The authors anticipate that these results will provide a significant contribution to the area of safety at signalized intersections as well as consider the possible consequences of the common practice of analyzing restricted datasets. Separate tree regression models for crashes of every severity level were also estimated to identify the significant factors that affect each. Injury levels are categorized as follows: no injury (property damage only), possible injury (no visible signs of injury), non-incapacitating injury (any visible injuries, e.g., bruises or limping), incapacitating injury (any visible signs of injury and the person is carried from the scene) and fatal injury (an injury sustained in a motor vehicle crash that results in death within 90 days).

2. Relevant studies

Many previous studies have used the ordered probit modeling methodology to study injury severity at different roadway locations. For example, Abdel-Aty (2003) applied the ordered probit models to predict crash injury severity on roadway sections, signalized intersections and toll plazas. Jianming and Kockelman (2004) used the ordered probit technique to predict injury severity based on factors including traffic, roadway and occupant characteristics and weather conditions at the time of a crash and type of vehicle. Kweon and Kockelman (2003) showed that wearing seatbelts decreases the risk of injury in crashes on highways.

Khattak and Targa (2004), Khattak et al. (2002, 2003) used ordered probit models to predict the injury level for crashes occurring at construction zones and involving trucks, to predict injury severity for single-vehicle truck rollovers, and to determine vehicle, roadway, driver, crash, and environmental characteristics that influence the severity level of older drivers involved in crashes, respectively.

Duncan et al. (1998) determined the specific variables that influenced the injury levels in two-vehicle (truck and car) rear-end crashes on divided roadways. Zajac and Ivan (2003) determined the effect of roadway and area features on the severity of pedestrian crashes in rural areas. Renski et al. (1999) revealed the effect that increases in the speed limit on interstate highways has on injury severity. Again, the ordered probit modeling methodology was the chosen technique in all of these studies. O'Donnell and Connor (1996) created two ordered probit models to predict the injury levels for crashes in Australia. Increases in both the age of injured person and the speed of vehicle caused a greater injury level.

It could be seen from this section that many studies have adopted the ordered probit modeling methodology to account for the natural ordering of the severity levels. In this study, not only the ordered probit will be adopted to depict the factors that affect the overall severity level, but the regression tree approach will also be attempted to show the importance of the significant factors for each severity level.

3. Data collection

Data collection began in early 2003 when several counties across the midsection of the State of Florida were contacted for cooperation. Four agencies that comprise a majority of Central Florida were identified and contacted: Brevard County, Seminole County, City of Orlando and Hillsborough County. Each jurisdiction provided drawings for several hundred intersections and each drawing was then individually examined and identified by the authors. Information obtained from each drawing was the number of through lanes on each roadway, the number of left turn lanes and whether they were exclusive, the presence of medians on each approach, whether any of the right turns were channelized and the speed limits. Each county also provided a database of crashes reported on both long- and short-forms for 3 recent years. In the meanwhile, crashes reported on long-forms were also downloaded from FDOT and DHSMV databases and cross-referenced against the crashes reported on long-forms provided by the counties. This process served as a check to ensure that each county's database was accurate. It was found that no database by itself was complete and each was missing crashes that another database included. Finally, crashes reported on long-forms from county, FDOT and DHSMV databases were combined with crashes reported on short-forms from the counties' databases to ensure that the dataset for this analysis was complete as much as possible. The master database created for this analysis includes 33,592 crashes from 832 intersections. Table 1 summarizes the data that was collected. Only data from years 2000 and 2001 would be used in model estimation (those were the only consistent years among the four jurisdictions). Table 1 shows that all counties have more crashes reported on short-forms

Table 1
Summary of data collected

County	Number of intersections	1999		2000		2001		2002		Total
		Long-forms	Short-forms	Long-forms	Short-forms	Long-forms	Short-forms	Long-forms	Short-forms	
Brevard County	151	–	–	490	1009	506	1015	561	1090	4671
City of Orlando	296	–	–	1793	2789	1745	2636	1690	2485	13138
Hillsborough County	190	1531	1052	1554	1262	1585	1333	–	–	8317
Seminole County	195	879	1556	799	1706	905	1621	–	–	7466
Combined total	832	2410	2608	4636	6766	4741	6605	2251	3575	33592

than long-forms except Hillsborough County. This shows that the police agencies in Hillsborough County are providing more crash records to the state agencies, which only maintain databases for crashes reported on long-forms. After identifying the years available for analysis, all possible independent and dependent variables were determined. The final database included 5 dependent variables and 40 possible independent variables. Table 2 lists the variable names, definitions and whether each was a dependent or independent variable in the models.

Fig. 1 presents the frequency of crashes by injury level for 2000 and 2001 for the four jurisdictions. The figure illustrates that about 80% of non injury crashes are reported on short-forms and therefore, are usually excluded from crash databases (for each other type of injury, crash databases are complete because crashes involving an injury are always

Table 2
Variables included in the final database

Variable	Variable definition	Variable role
int id	Intersection identification number	Identification
MJ Ln	Number of through lanes on major road	Independent
MN Ln	Number of through lanes on minor road	Independent
Tot LTMJ1 or 2	Number of left turn lanes (LTL's) on major road approach #1 or 2, respectively	Independent
Tot LTMN1 or 2	Number of left turn lanes (LTL's) on minor road approach #1 or 2, respectively	Independent
Tot LTLMJ	Total number of LTL's on major road	Independent
Tot LTLMN	Total number of LTL's on minor road	Independent
LTProt MJ1 or 2	Total number of exclusive LTL's on major road approach #1 or 2, respectively	Independent
LTProt MN1 or 2	Total number of exclusive LTL's on minor road approach #1 or 2, respectively	Independent
LTProt MJ	Total number of exclusive LTL's on major road	Independent
LTProt MN	Total number of exclusive LTL's on minor road	Independent
RTChMJ1 or 2	Whether right turn lanes are channelized on major road approach #1or 2, respectively	Independent
RTChMN1 or 2	Whether right turn lanes are channelized on minor road approach #1 or 2, respectively	Independent
RTChMJ	Whether any or all right turns are channelized on major, yes: 1 and no: 0	Independent
RTChMN	Whether any or all right turns are channelized on minor, yes: 1 and no: 0	Independent
DivMJ1 or 2	Whether major road is divided (by median or two-way LTL) on approach #1 or 2, respectively	Independent
DivMN1 or 2	Whether minor road is divided (by median or two-way LTL) on approach #1 or 2, respectively	Independent
DivMJ	Whether any or both approaches on major road are divided, yes: 1 and no: 0	Independent
DivMN	Whether any or both approaches on minor road are divided, yes: 1 and no: 0	Independent
SL MJ	Speed limit on major road	Independent
SL MN	Speed limit on minor road	Independent
ADT MJ	Average daily traffic on major road	Independent
ADT MN	Average daily traffic on minor road	Independent
Brevard Co.	Whether crash occurred in Brevard County, yes: 1 and no: 0	Independent
City of Orlando	Whether crash occurred in City of Orlando, yes: 1 and no: 0	Independent
Hillsborough Co.	Whether crash occurred in Hillsborough County, yes: 1 and no: 0	Independent
Angle crash	Whether crash was an angle crash, yes: 1 and no: 0	Independent
Head-on crash	Whether crash was an head-on crash, yes: 1 and no: 0	Independent
Left turn crash	Whether crash was an left turn crash, yes: 1 and no: 0	Independent
Pedestrian/bicycle crash	Whether crash was an pedestrian/bicyclist crash, yes: 1 and no: 0	Independent
Rear-end crash	Whether crash was an rear-end crash, yes: 1 and no: 0	Independent
Right turn crash	Whether crash was an right turn crash, yes: 1 and no: 0	Independent
Sideswipe crash	Whether crash was an sideswipe crash, yes: 1 and no: 0	Independent
No injury	Coded as 0 in ordered probit model	Dependent
Possible injury	Coded as 1 in ordered probit model	Dependent
Non-incapacitating injury	Coded as 2 in ordered probit model	Dependent
Incapacitating injury	Coded as 3 in ordered probit model	Dependent
Fatal injury	Coded as 4 in ordered probit model	Dependent

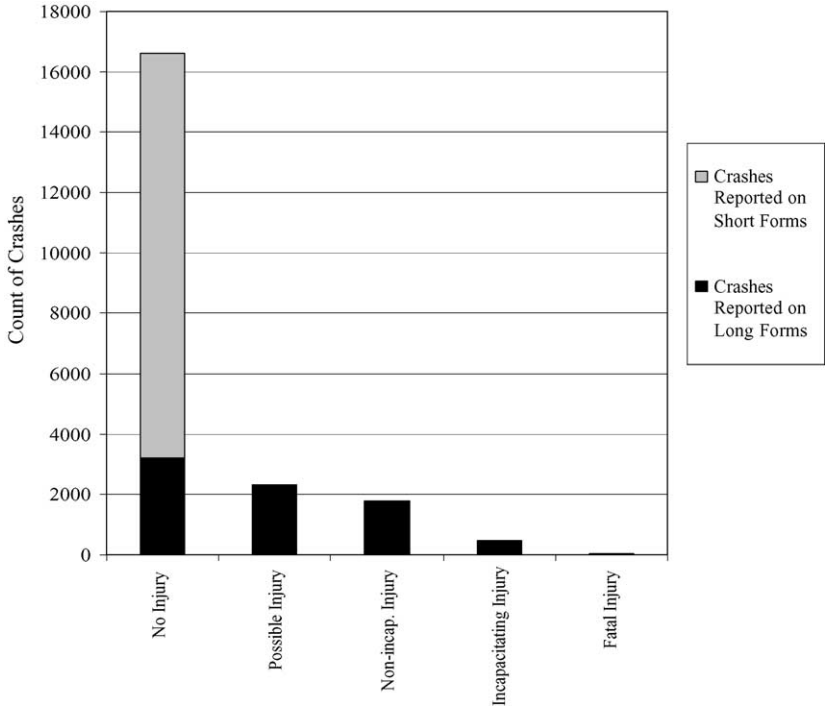

Fig. 1. Frequency of injury severity levels.

reported on long-forms). These crashes were also excluded in the restricted dataset for this research to show the consequences of analyzing an incomplete dataset.

4. Ordered probit modelling

4.1. Model definition

Due to the fact that some variables are naturally ordered, such as the severity level in a motor vehicle crash, various types of models can be specified for these types of data. The data for this research included crash-specific information such as the injury type, which was categorized into one of five groups: no injury, possible injury, non-incapacitating injury, incapacitating injury and fatal injury. These groups were then ranked from 0 to 4 with no-injury corresponding to the lowest level. Ordered probit models have gained popularity for this type of data mainly because they can account for the dependent variable's ordinal nature. The ordered multiple-choice model is as follows:

$$\sum_{j=1}^{j} P_n(j) = F(\alpha_j - \beta_j X_n, \theta), \quad j = 1, \ldots, J-1$$

$$P_n(J) = 1 - \sum_{J=1}^{J-1} P_n(j)$$

where $P_n(j)$ is the probability that subject n belongs to category j, α_j the alternative specific constant, X_n a vector of measurable characteristics, β_j a vector of estimable coefficients and θ is a parameter that controls the shape of the probability distribution F (Abdel-Aty, 2003). By assuming a standard normal distribution for F, the ordered probit model has the following form:

$$P_n(1) = \phi(\alpha_1 - \beta_j X_n)$$

$$P_n(j) = \phi(\alpha_1 - \beta_j X_n) - \phi(\alpha_{j-1} - \beta_j X_n), \; j = 2, \ldots, j-1$$

$$P_n(1) = 1 - \sum_{j=1}^{J-1} P_n(j)$$

where ϕ is the cumulative standard normal distribution function. The predicted outcome is the j-value with the largest probability (Abdel-Aty, 2003). Ordered probit models were created for this analysis using the econometric software LIMDEP. The ordered probit models created were based on crashes from the four counties/city for the years 2000 and 2001 as seen in Table 1. Only crashes where the exact injury severity was known were used for analysis. There were 7833 crashes reported on long-forms and 13,371 crashes reported on short-forms, making a total of 21,204 crashes used for these ordered probit models. The main objectives for this analysis were to determine the factors affecting crash severity as well as determine if there is a difference when models are based on the completeness of the data. Three types of ordered probit models were created; a model where crash types were the only independent variables, model where

intersection characteristics were the only independent variables and a final model that combined the variables in the first two models. Dummy variables representing the county that a crash occurred were included in all models to account for county-specific factors.

4.2. Severity models for crash type

The first ordered probit models created were arranged so that seven of the independent variables were dummies each representing different crash types: angle, head-on, left turn, pedestrian/bicycle, rear-end, right turn, sideswipe and other/unknown crashes (although there are eight types of crashes, only seven were used). Two models were created, one for the restricted dataset (based only on crashes reported on long-forms) and the other for the complete dataset (based on crashes reported on both long- and short-forms). Table 3 shows the variables used in each model. In Table 3, the values in the restricted model were calculated using only crashes reported on long-forms, the second set of values were calculated using crashes reported on both forms. In the restricted model, right turn crashes were found to be insignificant in the prediction of injury severity, however, in the complete model right turn crashes were found to reduce the likelihood of injury levels. The restricted dataset, which included 7833 crashes reported on long-forms, had a fairly low classification rate of 43.6%. However, when crashes reported on short-forms were added to the dataset, the classification accuracy increased to 78.4%.

Since the classification rate of the model based on the complete dataset was much higher, interpretations were based on this model. These coefficients show that the crash type likely to have the highest injury level is a crash involving a pedestrian or bicyclist ($\beta = 1.4982$). Of the motor vehicle crashes, left turn, angle and head-on crashes cause the most severe injury levels. On the other hand, it was found that right turn and sideswipe crashes tend to result in a lower crash injury level due to the negative coefficients in the model. Also, the model shows that Hillsborough County is more likely to be the location of a severe crash. One issue with this type of analysis is that a crash must first occur in order to predict the severity level. Therefore, this analysis is only useful when trying to identify the types of crashes that cause more severe injuries. To cope with this issue, models involving only intersection characteristics were estimated.

4.3. Severity models for intersection characteristics

In an effort to predict the expected crash severity levels for, say, a newly constructed intersection where no crashes have previously occurred, there was a need to create models based on other measurable variables related to the intersection characteristics. The independent variables available for these models are the variables listed in Table 2 that relate to the actual intersections. Again, two models were created, one for the restricted dataset and the other for the complete dataset. For these models, backward selection was utilized for simplification so that only variables found to be significant were included in the model. Table 4 shows the results of these ordered probit models. The model based on the restricted dataset had a relatively low prediction rate of 41.1%, while the model based on the complete dataset maintained the prediction rate of 78.4%. In addition to the variables shown in Table 2, a log transformation on the variable for ADT on the major road was tested. It was found that the variable log (ADT) was insignificant and caused a decrease in the prediction power

Table 3
Ordered probit model results based on crash types

Variable	Restricted dataset			Complete dataset		
	Coefficient	*t*-Statistic	*P*-Value	Coefficient	*t*-Statistic	*P*-Value
Constant	−0.5130	−11.552	0.0000	−1.3490	−35.815	0.0000
Angle	0.4596	10.358	0.0000	0.3846	10.220	0.0000
Head-on	0.6677	5.532	0.0000	0.2469	2.891	0.0038
Left turn	0.5983	12.879	0.0000	0.5617	14.148	0.0000
Ped/bike	1.1905	14.819	0.0000	1.4982	19.123	0.0000
Rear-end	0.2456	6.031	0.0000	0.1480	4.267	0.0000
Right turn				−0.3702	−4.110	0.0000
Sideswipe	−0.3507	−5.475	0.0000	−0.4006	−7.946	0.0000
Brevard Co.	0.6755	13.582	0.0000	0.4050	11.918	0.0000
City of Orlando	0.5946	15.067	0.0000	0.2995	10.555	0.0000
Hillsborough Co.	0.4717	14.007	0.0000	0.6477	23.283	0.0000
α_1	0.8189	54.793	0.0000	0.4883	51.150	0.0000
α_2	1.8428	72.067	0.0000	1.3039	61.542	0.0000
α_3	3.0419	45.530	0.0000	2.4079	40.256	0.0000
Sample size		7833			21204	
Degrees of freedom		9			10	
Chi-squared		783.450			1656.981	
log likelihood function		−9423.603			−14783.55	
Restricted log likelihood		−9815.328			−15612.04	

Table 4
Ordered probit model results based on intersection characteristics

Variable	Restricted dataset			Complete dataset		
	Coefficient	*t*-Statistic	*P*-Value	Coefficient	*t*-Statistic	*P*-Value
Constant	−0.0177	−5.950	0.0000	0.082980	2.608	0.0091
Major no. of lanes				−0.009710	−2.978	0.0029
MJ left turn lanes				0.022178	2.528	0.0115
RT channel. on MJ				−0.075410	−2.881	0.0040
Division on MN				−0.006740	−6.647	0.0000
Speed limit on MN	0.0001	3.866	0.0001	−0.000013	−9.361	0.0000
ADT on MJ				0.000001	2.050	0.0404
Brevard Co.	0.7835	14.297	0.0000	0.481081	11.164	0.0000
City of Orlando	0.5202	13.016	0.0000	0.516818	15.321	0.0000
Hillsborough Co.	0.4904	14.675	0.0000	0.207549	9.127	0.0000
α_1	0.7890	54.781	0.0000	0.4551	50.100	0.0000
α_2	1.7768	71.405	0.0000	1.2045	59.925	0.0000
α_3	2.9327	46.333	0.0000	2.2156	29.555	0.0000
Sample size		7833			21204	
Degrees of freedom		4			–	
Chi-squared		315.1828			–	
log likelihood function		−9657.736			−21280.62	
Restricted log likelihood		−9815.328			–	

of the model. Therefore, no transformations were used in the model.

Since the model based on complete data was proven to have a better classification accuracy, interpretations were again based on this model. Increases in the number of lanes and speed limit on the minor road, right turn channelization on the major road and division on the minor road were found to decrease the expected level of injury. Meanwhile, increases in the number of left turning lanes as well as traffic volume on the major road were found to increase the crash severity level. For this model, City of Orlando intersections were found to have a higher crash severity risk than Brevard and Hillsborough Counties because of the relatively larger coefficient associated with City of Orlando.

4.4. Severity models based on combined variables

Due to the fact that the models based on the complete dataset consistently had a higher prediction rate, a model was attempted with both intersection characteristics and crash type variables for the complete dataset. The variables available for this model were shown in Table 2 and backward selection was used to determine the most significant variables at a 5% level. Table 5 shows the final ordered probit model. The results indicate that crashes involving pedestrians or bicyclists have the greatest risk of a severe injury. Of the motor vehicle crashes, left turn, angle and head-on were most likely to result in a higher level of injury. Right turn and sideswipe crashes were found to be insignificant in the determination of severity and were dropped from the final model. With the exception of a median (or other type of division) and the speed limit on the minor road, all other intersection characteristics were found to be insignificant at the 5% level. Specifically, the presence of a median (or other type of division) on the minor road and a higher speed limit of the minor road were found to lower the risk of a serious injury. It is commonly believed that higher speeds in general cause higher severity. However, the results from the model shows a lower injury if the speed limit on the minor road is high, possibly indicating a better design for the intersecting minor roadway and more phases for the signal cycle. Finally, the City of Orlando was found to have a higher expected level of injury for crashes at signalized intersections amongst Brevard and Hillsborough Counties. The model achieved a classification accuracy of 79.1%.

Table 5
Final ordered probit model results based on combination of crash type and intersection characteristics

Variable	Complete dataset		
	Coefficient	*t*-Statistic	*P*-Value
Constant	−1.4780	−42.246	0.0000
Angle	0.5414	15.249	0.0000
Head-on	0.4859	5.310	0.0000
Left turn	0.7590	20.175	0.0000
Ped/bike	1.4757	20.352	0.0000
Rear-end	0.3150	10.062	0.0000
Division on MN	−0.0891	−3.731	0.0002
Speed limit on MN	−0.0001	−6.728	0.0000
Brevard Co.	1.2950	25.224	0.0000
City of Orlando	1.3367	34.216	0.0000
Hillsborough Co.	0.6615	25.189	0.0000
α_1	0.6472	53.430	0.0000
α_2	1.5576	71.915	0.0000
α_3	2.6848	40.991	0.0000
Sample size		21204	
Degrees of freedom		10	
Chi-squared		7578.639	
log likelihood function		−11822.72	
Restricted log likelihood		−15612.04	

5. Hierarchical tree-based regression

5.1. Approach

Hierarchical tree-based regression (HTBR) was used to estimate the expected number of crashes for each injury severity level. While the ordered probit models illustrated the significant factors that affect the overall severity levels whether accounting for collision types, intersection characteristics, or both, the HTBR models will focus on identifying the factors that affect each severity level, including no injury, one at a time. This method involves splitting the data into branches on a tree diagram based upon the given information and the average or expected value at each node. One of the most important benefits of this type of model is that, since it is based on crash frequencies under different conditions, the model does not require any assumptions or knowledge of the population's functional form in advance. HTBR is also robust against multicollinearity between the variables, which is commonly a problem in crash studies. Additionally, the model is capable of handling missing observations by treating a missing value as a valid response. Finally, outliers can easily be detected using tree-based regression because if an observation is an outlier, it will be on a branch alone. The model is essentially binary because the tree begins with one parent node that can split into exactly two child nodes. From here, each child node can either split into zero, one or two more child nodes. Nodes are split based upon the deviance of the sample and the splitting value is chosen such that the deviance in each of the two child nodes is minimized. Karlaftis and Golias (2002) defined the deviance as:

$$D = \sum_{i=1}^{L} (Y_{ia} - X_a)^2$$

where D is the deviance (also the sum of squared error) of Y at node a, Y_{ia} the observation at node a, and X_a is the average of L observations in node a. The analysis was conducted using SAS where stepwise variable selection as well as a splitting criterion based on an F-test was engaged.

5.2. Analysis

Tree-based regressions were conducted for crashes belonging to each severity level, for the complete dataset, and a model for the non injury crashes for the restricted dataset, resulting in a total of six regressions (four for each injury level, and two for non injury, since the difference between the restricted and complete data is only for this case, because all injury crashes are reported on long-forms). To visually identify the difference between models based on the restricted and complete datasets, a total of six tree diagrams were produced. Fig. 2 was chosen as an example of the regression trees for the prediction of the number of possible injury crashes per intersection for 2 years. The second box from the top reflects the number of observations in the dataset as well as the average number of possible crashes reported per intersection for the dataset. The third box contains the name of the independent variable that the data was split by to cause the largest decrease in deviance. For Fig. 2, the number of lanes on the minor road

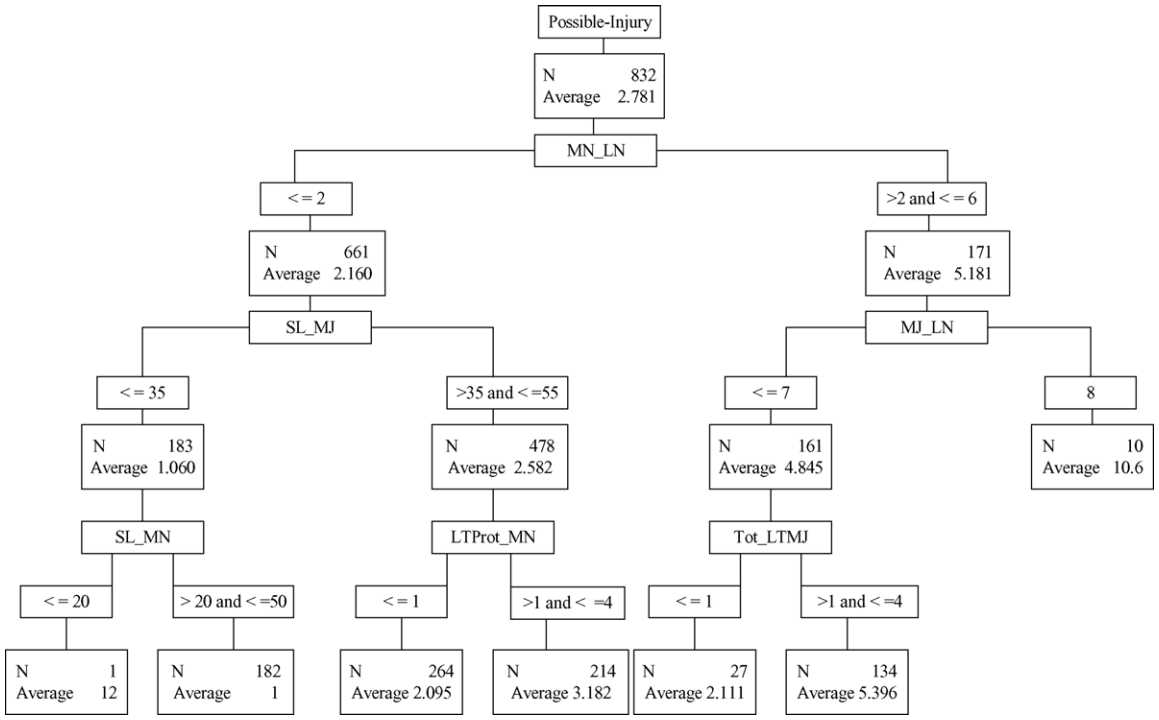

Fig. 2. Regression tree for the expected number of possible injury crashes per intersection for 2 years.

Table 6
Relative importance of the factors affecting every severity level based on the HTBR models

Variables	No injury		Possible injury	Non-incapacitating injury	Incapacitating injury	Fatal injury
	Complete	Restricted				
Daily traffic volume on MJ	1.0000	1.0000	0.6042	0.1022	0.7861	0.0000
Speed limit on MJ	0.5296	0.6133	0.5259	0.0000	0.4493	0.0000
Median present on MJ	0.5290	0.5000	0.0000	0.3132	0.3735	1.0000
Exclusive left turn lanes on MN	0.3856	0.6240	0.4134	1.0000	0.8114	0.0000
Number of lanes on MN	0.3225	0.5318	1.0000	0.1567	0.0000	0.0000
Speed limit on MN	0.1520	0.2756	0.3115	0.2204	0.3685	0.0000
Exclusive left turn lanes on MJ	0.1457	0.3018	0.0000	0.0000	0.0000	0.4858
Daily traffic volume on MN	0.1266	0.5961	0.1976	0.3916	1.0000	0.0000
Number of lanes on MJ	0.0853	0.5270	0.5512	0.4867	0.3417	0.7687
Right turns channelized on MJ	0.0000	0.0166	0.0000	0.6193	0.2076	0.0000
Median present on MN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Right turns channelized on MN	0.0000	0.1016	0.0000	0.1526	0.1044	0.0000
Total left turn lanes on MJ	0.0000	0.0000	0.4422	0.0000	0.0000	0.0000
Total left turn lanes on MN	0.0000	0.1800	0.0000	0.0000	0.0000	0.0000

was found to minimize the deviance most. The result is that the tree diagram breaks into two branches and then has the opportunity to branch again. In this case, the left branch is further divided by the speed limit on the major road and the right branch is divided by the number of lanes on the major road. Each of the new branches again splits one more time before stating the final expectation for the number of crashes. Furthermore, by examining the various tree levels, the effect that a certain variable has on the frequency of crashes can be determined.

5.3. Importance of factors

In addition to creating tree diagrams to visually describe the difference in models, lists of variables that entered into each model and their relative importance were also produced. Variables found to be significant were identified as “input” variables, and “rejected” variables were those that did not enter the particular model. To determine each variable’s importance, the improvement in the reduction of deviance that can be attributed to each variable for the first split in the tree is rated. These values are then summed and scaled, showing the variable that reduced the deviation most to have an importance value of 1.00. Table 6 illustrates the significant factors and their level of importance for the six estimated models. It is clear that the significant factors are different among the models. Their level of importance is also different. For example, the most important factor that affects fatal crashes is whether the major road is divided followed by the number of lanes on the major road. For no injury crashes, the traffic volume of the major road is the most significant factor. The number of lanes on the minor road, the number of exclusive right turn lanes, and the ADT on the minor road, are the most significant factors for possible injury, non-incapacitating injury and incapacitating injuries, respectively. Also Table 6 illustrates that there are also differences in estimating the number of non injury crashes when using the restricted versus the complete datasets. While both have the ADT as the most significant factor, the following variables are not consistent. The speed limit and median on the major road were significant for the model based on the complete data, while the exclusive left turning lanes, speed limit on major road and ADT on minor road were significant for the model based on the restricted dataset. Again confirming the results from the ordered probit modeling section that the completeness of the data is very important and that the dataset that misses crashes would produce different results.

6. Conclusion

This paper explored the severity levels of crashes at signalized intersections. The ordered probit modeling methodology has been adopted to analyze the overall severity levels, while regression tree models looked into each severity type and the factors that affect each of them.

Ordered probit models were created in this study for three different types of variables; one based on collision types, another for intersection characteristics and the last for a combination of significant variables. Both the restricted and complete datasets were used to create these models and the output was compared. It was determined that the models based on the complete dataset were more accurate.

The final ordered probit model based on the complete dataset included 10 significant variables and a constant term. The first five variables referred to the type of crash, the next two dealt with intersection characteristics on the minor roadway and the last three accounted for county-specific factors. Division of the minor road, as well as a higher speed limit on the minor road, was found to lower the expected injury level. A median on the minor road may prevent more head-on crashes, which were found to be more severe crashes. A higher speed limit on the minor road may cause the differential speeds between vehicles on intersecting roads to be smaller and might indicate a better roadway design, likely resulting in a decrease in the crash severity level.In this analysis,

the vehicle crash type that has the highest risk for a severe injury is the left turn crash, followed by angle and head-on crashes. The fact that crash types were found to be significant for predicting injuries is similar to results of O'Donnell and Connor (1996) whose ordered probit models also showed that collision type to have a significant effect on crash injury levels. Similar to the results shown in Table 5, the study conducted by Renski et al. (1999) found that increasing speed limits (up to 65 mph) increased the probability of less severe injuries. Jianming and Kockelman (2004) found that sideswipe and rear-end crashes to be less severe while Table 5, shows that rear-end crashes do cause less severe injuries but sideswipe crashes were found to be insignificant.

Regression tree models depicted the important factors and the level of their importance that influence each crash severity level. The results summarized in Table 6 illustrated substantial differences among the factors affecting each severity level. For example, Table 6 shows that the most important factor in predicting a non injury crash is the daily traffic volume on the major road, which was found to be insignificant in predicting fatal crashes. Quite contrary, in the ordered probit models, the traffic volume on the major road was only significant in the complete model and it predicted higher severity levels for higher volumes. One similarity between the tree-based regression and the ordered probit complete model was that speed limits on the minor road significantly affected lower injury severity levels. The presence of a median caused lower crash severity levels in the complete ordered probit model but was found to be insignificant in the tree-based regression models. Only three variables (presence of a median on the major road, number of lanes on the major road and number of exclusive left turn lanes on the major road) were found to be important in the tree-based regression model for the number of fatalities. The complete ordered probit model did not show the presence of a median on the major road to be significant, however, the presence of a median on the minor road was found to lower the injury level. Finally, right turn channelization on the minor road was found to cause lower injury crashes in the complete ordered probit models but was essentially insignificant in the non injury and possible injury models created from tree-based regressions.

In summary, the results of this research show that when attempting to forecast the number of expected crashes of different severity levels, it is imperative that models are developed for each level of collision instead of aggregating crash types to predict the overall severity level. While we have adopted the ordered probit model approach, as did many previous researchers, using the tree-based regression for each severity level improved our understanding of the specific factors and their importance for each severity level. Furthermore, the results showed that crashes reported on short-forms are important and should therefore be retained and included in crash databases. Ignoring this data could lead to biasing the results by under reporting crashes of certain severity or type that could be related to specific explanatory factors. Other crash types or severities might appear to have higher percentages, and therefore, their effect could be artificially exaggerated.

Acknowledgement

The authors wish to thank the Florida Department of Transportation for funding this research. All opinions and results are those of the authors.

References

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. J. Saf. Res. 34 (5), 597–603.

Duncan, C., Khattak, A., Council, F., 1998. Applying the Ordered Probit Model to Injury Severity in Truck-passenger Car Rear-end Collisions. Transportation Research Record 1635. Transportation Research Board, National Research Council, pp. 63–71.

Jianming, M., Kockelman, K., 2004. Anticipating injury & death: controlling for new variables on Southern California highways. In: Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, DC.

Karlaftis, M., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accid. Anal. Prev. 34 (3), 357–365.

Khattak, A., Targa, F., 2004. Injury severity and total harm in truck-involved work zone crashes. In: Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, DC.

Khattak, A., Schneider, R., Targa, F., 2003. Risk factors in large truck rollovers and injury severity: analysis of single-vehicle collisions. In: Presented at the 82nd Annual Meeting of the Transportation Research Board, Washington, DC.

Khattak, A., Pawlovich, M., Souleyrette, R., Hallmark, S., 2002. Factors related to more severe older driver traffic crash injuries. J. Transportation Eng. 128 (3), 243–249.

Kweon, Y., Kockelman, K., 2003. Driver attitudes and choices: seatbelt use, speed limits, alcohol consumption, and crash histories. In: Presented at the 82nd Annual Meeting of the Transportation Research Board, Washington, DC.

O'Donnell, C., Connor, D., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choices. Accid. Anal. Prev. 18 (6), 739–753.

Renski, H., Khattak, A., Council, F., 1999. Effect of Speed Limit Increases On Crash Injury Severity: Analysis Of Single-Vehicle Crashes On North Carolina Interstate Highways. Transportation Research Record 1665. Transportation Research Board, National Research Council, pp. 100–108.

Zajac, S., Ivan, J., 2003. Factors influencing injury severity of motor vehicle crossing pedestrian crashes in rural Connecticut. Accid. Anal. Prev. 35 (3), 369–379.