Accident Analysis and Prevention 130 (2019) 91–98

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Evaluation of surrogate measures for pedestrian trips at intersections and crash modeling

Jaeyoung Lee*, Mohamed Abdel-Aty, Imran Shah

Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, Florida, 32816-2450, United States

ARTICLE INFO

Keywords:
Pedestrian exposure
Pedestrian safety
Pedestrian crash analysis
Surrogate measures
Zero inflated negative binomial model
Risk factors

ABSTRACT

Pedestrians are considered the most vulnerable road users who are directly exposed to traffic crashes. With a view to addressing the growing concern of pedestrian safety, Federal and local governments aim at reducing pedestrian-involved crashes. Nevertheless, pedestrian volume data are rarely available even though they among the most important factors to identify pedestrian safety. Thus, this study aims at identifying surrogate measures for pedestrian exposure at intersections. A two-step process is implemented: the first step is the development of Tobit and generalized linear models for predicting pedestrian trips (i.e., exposure models). In the second step, negative binomial and zero inflated negative binomial models were developed for pedestrian crashes using the predicted pedestrian trips. The results indicate that among various exposure models the Tobit model performs the best in describing pedestrian exposure. The identified exposure-relevant factors are the presence of schools, car-ownership, pavement condition, sidewalk width, bus ridership, intersection control type and presence of sidewalk barrier. It was also found that the negative binomial model with the predicted pedestrian trips and that with the observed pedestrian trips perform equally well for estimating pedestrian crashes. Also, the difference between the observed and the predicted pedestrian trips does not appear as statistically significant, according to the results of the *t*-test and Wilcoxon signed-rank test. It is expected that the methodologies using predicted pedestrian trips or directly including pedestrian surrogate exposure variables can estimate safety performance functions for pedestrian crashes even though when pedestrian trip data is not available.

1. Introduction

Being classified as vulnerable road users, pedestrians are recognized as the worst victims of traffic crashes. The injury severity levels of pedestrians are relatively high compared to those in motor vehicle crashes. The reason behind this is unlike the passengers or drivers in motor vehicle crashes, pedestrians are directly exposed to the impact of traffic crashes. Nowadays it is one of the most challenging and growing concerns for transportation engineers to ensure safe movement of pedestrians. With a view to developing a sustainable, safe and dynamic transportation system, proper attention must be given to the safety of pedestrians. Walking also has significant contribution to public health since it can reduce rates of chronic disease and ameliorate rising health care costs (Lee and Buchner, 2008). The crash statistics demonstrate the relative risk level of people walking on the road. In 2014, there were 4,884 pedestrians killed and 65,000 injured on roads with a rate of 1.53 person per 100,000 population in the United States (National Highway Traffic Safety Administration, 2018). Among them there were 923 (18.9%) that died at or near intersections. The proportion of pedestrian fatalities has steadily increased from 11% to 14% over the past decade. The number describes the importance of addressing the safety of pedestrians and raising awareness among the population about safe walking.

With a view to addressing the safety of pedestrians, the study aims at identifying the exposure measure to pedestrian crashes at intersections. It is obvious that the number of people walking on the road (i.e., pedestrian trips) is one of the best measures of exposure for pedestrians (Davis and Braaksma, 1988; Qin and Ivan, 2001; Lam et al., 2014). However, it is difficult to continuously measure the pedestrian trips at all locations as it involves using significant amount of resources. This study aims at addressing the situation when it is difficult to collect pedestrian trip data. The objective of this study is to analyze surrogate measures for pedestrian exposure to traffic at intersections including the use of socio-demographic, land-use and geometric characteristics of the surrounding environment. The two-step process implemented in the study involves developing the exposure models first and then the crash models. The exposure models were developed using Tobit and generalized linear modeling methods that predict the pedestrian trips. After

* Corresponding author.
E-mail address: jaeyoung@knights.ucf.edu (J. Lee).

https://doi.org/10.1016/j.aap.2018.05.015
Received 2 February 2017; Received in revised form 13 April 2018; Accepted 19 May 2018
Available online 31 May 2018
0001-4575/

identifying the best exposure model, negative binomial and zero inflated negative binomial crash models were developed using the predicted pedestrian trips as an exposure variable. The method can be described as an integrated pedestrian safety analysis around intersections (micro-level) with macro-level data from census block groups.

The study area is located in Orange and Seminole Counties of Central Florida where pedestrian fatality rate is nearly double (3.05 per 100,000 population) compared to the nationwide average fatality rate (1.53 per 100,000 population). Although the two-step process (i.e. first estimating pedestrian trips and then predicting crash) has relatively larger modeling error than one-step model (predicting pedestrian crash directly), it still can be used for better understanding of safety by analyzing pedestrian volume and crash data at the same time. The study contributes in the research area of pedestrian safety through identifying best exposure for pedestrian crashes at intersections and developing a process of safety analysis for locations where pedestrian data is not available. Proper knowledge of exposure factors can help transportation officials to develop safer roads for pedestrians through implementing the right safety interventions.

2. Literature review

Since pedestrian safety is a recently growing concern, there has been extensive research studies conducted to ensure the safe movement of pedestrians. Researchers have attempted to identify the factors responsible for fatalities of people walking on the road. Before proceeding to the actual study a brief review of the previous studies has been carried out to better understand the underlying procedure of pedestrian safety analysis.

Lee et al. (2015) applied different exposure variables for pedestrians and found that the product of 'Log of population' and 'Log of Vehicle-Miles-Traveled (VMT)' is the best exposure for 'Pedestrian crashes per crash location ZIP code area', whereas 'Log of population' is the best exposure variable for 'Crash-involved pedestrians per residence ZIP'. The authors combined hot zones from where vulnerable pedestrians originated with hot zones where many pedestrian crashes occur. It was expected that the proposed screening method would be helpful for the practitioners to suggest appropriate safety treatments for pedestrian crashes.

Ukkusuri et al. (2011) developed a random-parameter negative binomial model of pedestrian crash frequencies for New York City at the census-tract level. The model found that the proportion of uneducated population, Black or Hispanic neighborhood areas, commercial areas, school areas, intersection operation characteristics, type of access control in the roads and number of lanes have positive impacts on pedestrian crashes. Based on the results the authors emphasized the focus on improved policy framework to improve pedestrian safety.

Lee and Abdel-Aty (2005) made a comprehensive study on vehicle-pedestrian crashes at intersections in Florida. The study followed Keall's method (Keall, 1995) to develop a logical expression of pedestrian exposure to crash risk using the individual walking trip data collected from the household travel survey. The proposed exposure reflected different walking patterns by different age groups of pedestrians. In spite of applying certain assumptions and adjustments it was quite hard for the authors to identify pedestrian exposure.

Miranda-Moreno et al. (2011) analyzed two important relationships between land development and pedestrian which are: (a) between the land-use and pedestrian activities, and (b) between the risk exposure (pedestrian and vehicle activities) and the pedestrian crash frequency. The authors concluded that the land-use pattern affects the pedestrian activity level with limited direct effect on pedestrian safety. Land-use affects the pedestrian volume, which is an important component of exposure to risk. The authors also found a non-linear relationship between the exposure and the crash count. These outcomes and the difficulty in identifying the pedestrian exposure data mentioned by Lee and Abdel-Aty (2005) helped in justifying models that include exposure-related variables.

Abdel-Aty et al. (2007) analyzed the safety of students around schools and found that middle and high school children are involved in crashes more frequently than younger children. It confirmed that school-aged children are exposed to high crash risk near schools. The authors figured out that driver's age, gender, and alcohol use, pedestrian's/bicyclist's age, number of lanes, median type, speed limits, and speed ratio are correlated with the frequency of crashes. These pedestrians and bicyclists' demographic factors and geometric characteristics of the roads adjacent to schools are expected to be considered in determining safety interventions of school districts. The study presented an example of combining two approaches to safety improvement including identification of locations with pedestrian safety problems and evaluating specific safety interventions.

It has been observed in recent studies that the application of zero-inflated model in traffic safety analysis is questionable to many researchers (Lord et al., 2007; Lord et al., 2005; Son et al., 2011). The basic dual-state assumption for crash occurrence has been criticized specifically for micro-level analysis. Although the criticism may be acknowledged for micro-level analysis of vehicle crash count, it may not be applicable for the cases of pedestrian crashes. Since there may be cases where no pedestrian activity is observed due to the absence of walking infrastructure and for the same reason expected number of zero pedestrian crash is possible. In such circumstances the dual state representation can describe the excess zero cases in terms of exogenous variables. Hence, the present study considers the application of both single-state (negative binomial) and dual-state models (zero inflated negative binomial) for analyzing pedestrian crashes at the micro-level. It is obvious that developing negative binomial models without considering the excess zeros may result in biased estimates.

In order to select the variables to be used in the suggested exposure model several previous studies have been analyzed. Previous researchers have shown that the volume of pedestrians is a significant exposure measure that has a positive impact on the occurrence of vehicle-pedestrian collisions (Davis and Braaksma, 1988; Qin and Ivan, 2001; Lam et al., 2014; Lam et al., 2013; Lee et al., 2018a). Another significant exposure measure is vehicular traffic that has significant impact on vehicle-pedestrian collisions (Lee and Abdel-Aty, 2005; Van den Bossche et al., 2005; Wier et al., 2009). The effects of land-use pattern have long been examined by researchers (Wier et al., 2009; Cervero, 1996; Graham and Stephens, 2008; Lee and Abdel-Aty, 2017; Lee et al., 2017; Lee et al., 2018b). It was found by Wier et al. Wier et al. (2009) that the frequency of pedestrian crashes is relatively larger in commercial and residential areas. Lam et al. (2014) found that public transport facilities are significant in pedestrian collisions because of the fact that pedestrians are most often in a hurry to board buses or cross roads immediately after getting off. There exist quite a lot of studies that describe the impact of demographic and socio-economic characteristics on pedestrian safety (Graham and Stephens, 2008). Most of the studies found that pedestrian safety is a greater concern in socially deprived areas.

Although previous researchers put their effort to explain pedestrian exposure to risk, there are very few studies that identified the exact pedestrian exposure factors specifically at intersections. Another issue is the reliability of pedestrian volume data. There are many cases where pedestrian volume data is not available or accurate enough to do a safety analysis. A reliable process of identifying surrogate measures for pedestrian exposure needs to be established in such cases. Apart from these issues, the application of Zero Inflated Negative Binomial model at micro-level safety analysis has been questionable to many authors (Lord et al., 2007; Lord et al., 2005; Son et al., 2011) although some authors adopted it in macro-level analysis (Cai et al., 2016). The current study is inspired by these aforementioned research questions. It investigates them with respect to socio-demographic, land-use and geometric characteristics of the ambient environment. The study included similar variables that have been used in previous studies, and also new

variables have been included that could possibly affect pedestrian safety analysis.

3. Methodologies

The study followed several statistical procedures to identify the pedestrian exposure and analyzed the crash frequency based on different model building techniques. The reason behind this is to identify the most appropriate modeling technique that works best with pedestrian exposure determination. Below are brief descriptions of the modeling techniques that are adopted in this study:

3.1. Methods used for exposure models

3.1.1. Generalized linear model (GLM)

Generalized Linear Models (GLM) is a general class of statistical models that includes many commonly used models as special cases. The equation of GLM is given by:

$$Y = \sum_{i=1}^{m} \beta_i x_i + \varepsilon_i \quad (1)$$

Where $\sum_{i=1}^{m} \beta_i x_i = \eta\ (say)$ is the linear predictor and ε_i is the error term. Generalized linear models are characterized in the following ways:

a) In the generalized linear model, the assumptions of independent and normal distribution of the components of y are relaxed. It allows the distribution to be any distribution that belongs to the exponential family of distributions. This includes distributions such as Normal, Poisson, gamma and binomial distributions.
b) Instead of modeling the mean, $\mu = E(y)$ directly as a function of the linear predictor η, some function g (μ) of μ is used. Thus, the model becomes $g(\mu) = \eta = \sum_{i=1}^{m} \beta_i x_i$. Where, the function g (·) is called a link function (Olsson, 2002).

The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

The current study followed linear regression of generalized linear model to identify the exposure factors which does not assume that the distribution is normal.

3.1.2. Tobit model

In order to handle any negative prediction of pedestrian trips, the Tobit model was used to identify the exposure. The Tobit model is a statistical model used to describe the relationship between a non-negative dependent variable (censored dependent variables) y_i and an independent variable (or vector) x_i. The Tobit model which is introduced by James Tobin (1958) takes the form:

$$y_i^* = \beta x_i + \varepsilon_i,\ i = 1, 2, \ldots, N \quad (2)$$

$$\text{and, } yi = \begin{cases} y_i^* \text{ if } y_i^* > 0 \\ 0 \text{ if } y_i^* \le 0 \end{cases} \quad (3)$$

where y_i^* is a latent variable that is observed only when positive, N is the number of observations, y_i is the dependent variable, x_i is a vector of explanatory variables, β is a vector of estimable parameters and ε_i is a normally and independently distributed error term with zero mean and constant variance σ^2 (Washington et al., 2010).

3.1.3. Variable importance for exposure model using random Forest

The important explanatory variables in the exposure model were determined using Random forest procedure. The first step of this process is to fit a random forest to the data. During the fitting process the out-of-bag error (a method of measuring the prediction error of random forests) for each data point is recorded and averaged over the forest. In order to measure the importance of the *j*-th feature after training, the values of the *j*-th feature are permuted among the training data and the out-of-bag error is again estimated on this perturbed data set. The difference in out-of-bag error before and after the permutation is averaged over all trees to determine the importance score for the *j*-th feature. Finally, the score is normalized by the standard deviation of these differences. Features with higher score values are ranked as more important than the others (Breiman, 2001).

3.1.4. Variable importance for exposure model using principal component analysis (PCA)

Principal Component Analysis (PCA) aims at compressing the size of the data set by extracting the most important information from the data table. It analyzes the structure of the observations and the variables, and then computes new variables called principal components which are obtained as linear combinations of the original variables (Washington et al., 2010). The first principal component is required to have the largest possible variance. The second component that is uncorrelated with the first component captures most information not captured by the first component. PCA maximizes the variance of the elements of z = xu, such that uu' = 1 where $z = [z_1, z_2, \ldots z_n]$, $x = [x_1, x_2, \ldots x_n]$ and $u = [u_1, u_2, \ldots u_n]'$. The solution is obtained by solving the equation (R-λI) u = 0, where R is the sample correlation matrix of the original variables x, λ is the eigen-value and u is the eigen-vector.

3.2. Statistical test to compare observed and predicted pedestrian trips

3.2.1. Paired sample t-test

Paired sample *t*-test is a statistical technique that is used to compare two population means in the case of two samples that are correlated. It is a parametric test procedure that assumes the population to be normally distributed. Paired sample *t*-test is used in 'before-after' studies, or when the samples are the matched pairs, or when it is a case-control study. The null hypothesis states that the mean of two paired samples are equal. The following formula is used to calculate the statistics of the test:

$$t = \frac{\bar{d}}{\sqrt{s^2/n}}$$

Where d bar is the mean difference between two samples, s^2 is the sample variance, n is the sample size and t is the test statistics with n-1 degrees of freedom.

3.2.2. Wilcoxon signed-rank test

The study compared the observed and predicted pedestrian trips using the Wilcoxon Signed-Rank test. It is a nonparametric test procedure that can analyze matched-pair data based on differences to assess whether their population mean ranks differ. The method does not require assuming the population to be normally distributed. The null hypothesis is that the difference between the pairs follows a symmetric distribution around zero. The absolute values are ranked and the test statistic is calculated by adding the ranks for either the positive or the negative values (Woolson, 2008).

3.3. Methods used for crash models

3.3.1. Negative binomial (NB) model

Since the beginning of crash frequency analysis the Poisson model has been the most accepted by the researchers (Lord and Mannering, 2010). The basic assumption of Poisson model is equal mean and variance of the distribution. However, crash data are often overdispersed. Because the Poisson model is not capable of dealing with the overdispersed crash data, the Poisson models are not popularly used in traffic safety field nowadays. The NB model relaxes the equal mean variance assumption of Poisson model. The NB model can generally

account overdispersion resulting from unobserved heterogeneity and temporal dependency, but may be improper for accounting for the overdispersion caused by excess zero counts (Rose et al., 2006). The negative binomial distribution has an extra parameter ε_i than the Poisson regression that adjusts the variance independently from the mean. The error term ε_i that considers overdispersion parameter to the mean of the Poisson model according to the following equation:

$$\lambda_i = \exp(\beta x_i)\exp(\varepsilon_i) \quad (4)$$

Where λ_i is the expected number of Poisson distribution for entity i, x_i is a set of explanatory variables, and β_i is the corresponding parameter. Here the distribution of the error term $\exp(\varepsilon_i)$ is normally assumed to be gamma-distributed with mean 1 and variance α. It makes the variance of the crash frequency distribution $\lambda_i(1 + \alpha\lambda_i)$ which is not equal to the mean λ_i. The NB model for the crash count y_i of entity is given by:

$$P(yi) = \frac{\Gamma\left(y_i + \frac{1}{a}\right)}{\Gamma(y_i + 1)\,\Gamma\left(\frac{1}{a}\right)}\left(\frac{\alpha\lambda_i}{1 + \alpha\lambda_i}\right)^{y_i}\left(\frac{1}{1 + \alpha\lambda_i}\right)^{\frac{1}{\alpha}} \quad (5)$$

where y_i is the number of crashes y_i of entity i and Γ(•) refers to the gamma function.

3.3.2. Zero-Inflated negative binomial (ZINB) model

The zero-inflated models assume that the data have a mixture with a degenerate distribution whose mass is concentrated at zero (Lambert, 1992). The first part of the mixture is the extra zero counts and the second part is for the usual single state model conditional on the excess zeros (Cai et al., 2016). The zero-inflated NB model can be regarded as an extension of the traditional NB specification as:

$$yi \sim \begin{cases} 0, & \textit{with probability } p_i \\ NB, & \textit{with probability } 1-p_i \end{cases} \quad (6)$$

The logistic regression model is employed to estimate p_i,

$$pi = \frac{\exp(\beta_i' x_i)}{1 + \exp(\beta_i' x_i)} \quad (7)$$

where β_i is the corresponding parameter. Substituting Eqs. (5) into (6) we can define ZINB model for crash counts y_i of entity i as:

$$P(yi) = \begin{cases} p_i + (1-p_i)\left(\frac{1}{1+\alpha\lambda_i}\right)^{\frac{1}{\alpha}}, & y_i = 0 \\ (1-p_i)\frac{\Gamma\left(y_i + \frac{1}{a}\right)}{\Gamma(y_i+1)\,\Gamma\left(\frac{1}{a}\right)}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}\left(\frac{1}{1+\alpha\lambda_i}\right)^{\frac{1}{\alpha}}, & y_i > 0 \end{cases} \quad (8)$$

Initially, all the explanatory variables were used directly to predict the pedestrian trips GLM and Tobit models. Next the Random Forests process has been applied before running the GLM and Tobit models to identify the important variables in the dataset that should be given priority. Finally, Principal Component Analysis has been applied before running the GLM and Tobit models to the explanatory variables to find out the components that are related to specific group of variables. After the exposure model has been identified the corresponding predicted pedestrian trips were calculated. Finally using the predicted pedestrian trips, crash models were developed following the negative binomial and zero-inflated negative binomial modeling techniques.

4. Data preparation

The explanatory variables were classified into three categories namely 'Demographic and Socioeconomic', 'Land-use', and 'Traffic and Geometric'. There were in total 134 intersections in Orange and Seminole Counties of Central Florida that have been used for the analysis. The data were collected for each intersection from various data sources. Geographic Information Systems (GIS) and SAS software were used to extract and process the data. In order to extract the data, a circular area (referred as to "buffer" in this paper) around the intersection as center was defined to extract the data for each specific variable category.

4.1. U.S. Census Bureau

American Community Survey data that is a 5 year estimates was used in this study. A buffer size of 0.25-mile radius was defined to extract the census data from the American Fact Finder database. The buffer size was chosen with 0.25 miles radius since over the past 2 decades, 0.25 miles (400 m or a 5-minute walk) has been assumed to be the distance that "the average American will walk rather than drive" (Boer et al., 2007; Yang and Diez-Roux, 2012).

4.2. Florida department of revenue property tax oversight program

The database provides land-use pattern of District 5, Florida from which the land-use of the study area was extracted from the department of revenue land-use code. Similar to the census variables a buffer size of 0.25-mile radius was used for extracting the data.

4.3. LYNX

In order to find the bus ridership and number of bus stops around intersections the LYNX GIS database was used. LYNX is the company that operates the public transport system in the study area. The buffer size is same as FDOT GIS data.

4.4. FDOT roadway characteristics inventory (RCI) data

FDOT Roadway Characteristics Inventory is a useful data source of traffic crashes, road infrastructure and traffic characteristics. The crash data was extracted from the 2013–2015 crash database produced by the Florida Department of Transportation (FDOT) Safety Office. Since crashes are not reported exactly at the intersection a circular buffer size of 50 ft. radius from the stop bar (the pavement marking line behind which vehicles stop at intersections) of the intersection is defined to extract the number of crashes. The traffic and geometric characteristics data were extracted from FDOT RCI (Roadway Characteristics Inventory). In order to extract traffic and geometric data a buffer size of 100 ft. radius was defined.

Fig. 1 shows the pedestrian trips distribution of the 134 intersections in the dataset. It can be seen from the distribution that most of the intersections have a volume ranging 20–60 trips (more than 50%). The average number of pedestrian trips of an intersection is around 56 trips. The distribution is skewed to the right. The pedestrian volume count was taken from eight-hour turning movement of a typical weekday. The hours of pedestrian counting operations are 7 A M -9. A.M, 11 A.M-1 P.M and 4 P.M-8 P.M. These counts were prepared for Florida Department of Transportation District 5 that also includes traffic volume along with pedestrian trips. It was considered that if a pedestrian crossed any of the one legs of the intersection in any direction that will be taken as one pedestrian count. The total number of pedestrian over the eight hours was taken as the dependent variable of the exposure model.

Fig. 2 shows the pedestrian crash frequency distribution of the intersections selected for the analysis. It can be seen that most of the intersections have zero pedestrian crashes (around 66%) which shows pedestrian crashes are rare events. The average number of crashes of an intersection is 0.563. Again, the distribution is skewed to the right. The crashes within the 50 ft distance of the intersection are considered as the crash frequency of that particular intersection since the crash may not exactly be reported at the intersection.

The correlations among the explanatory variables used in the Tobit model were checked before the modeling process. Any variable pairs that have higher than correlation coefficient of 0.4 were not included

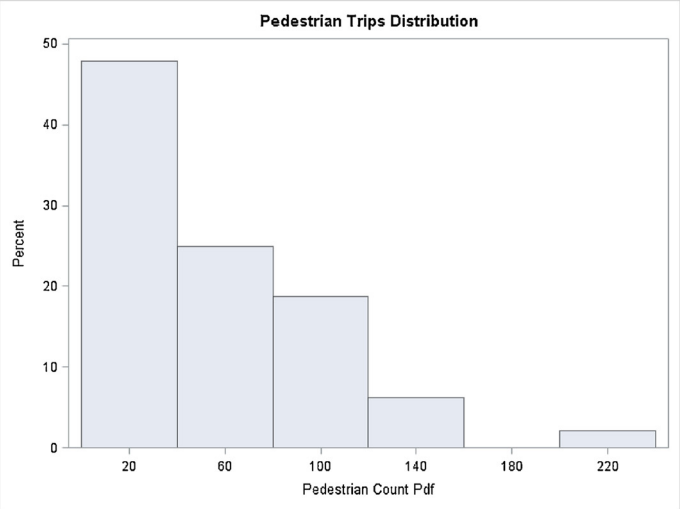

Fig. 1. Pedestrian Trips Distribution.

simultaneously in the models. In the modeling process the explanatory variable with the smallest correlation value was included first in the model and the variables with relatively smaller correlation value were given preference in the model (Table 1).

5. Modeling results

There were six exposure models in total developed in this study. Since the modeling technique is different (GLM vs Tobit) it is unsuitable to compare the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) to find the best model. In order to compare the models, the Mean Absolute Deviation (MAD) and Root Mean Square Error (RMSE) were calculated for each model using the following formulas:

$$MAD = \frac{\sum_{i=1}^{n} |y_{pred} - y_{obs}|}{n} \quad (9)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_{pred} - y_{obs})^2}{n}} \quad (10)$$

MAD and RMSE can measure the differences between values predicted by a model and the values actually observed. The model with lower MAD and RMSE value is considered to be relatively more accurate than the other models. It was found that the models developed by using significant variables from the random forests process and principal component analysis did not perform well. It was also found that compared to GLM, Tobit models performed better in every case. Table 2 below provides list of all the models developed and compared in terms of MAD and RMSE.

Table 2 shows that the Tobit model using all variables performs best with the lowest MAD and RMSE values. The Tobit model also handles any negative predicted pedestrian trips with the value zero since the lower bound is set at 0. In this study the significance level of identifying exposure is relaxed to 10% since the factor identification is the primary objective rather than the value of the parameters. Table 3 presents the significant factors obtained from the best Tobit model to predict the pedestrian trips.

The significant variables in the model are discussed below:

5.1. Presence of school near intersection (1 = yes, 0 = no)

There are more pedestrian activity near schools due to parent drop off and pick up, children walking along the route in groups, etc. Intersections near schools are expected to have higher pedestrian exposure especially at the start and ending times of schools. In this study the school is used as a dummy variable and the positive coefficient in

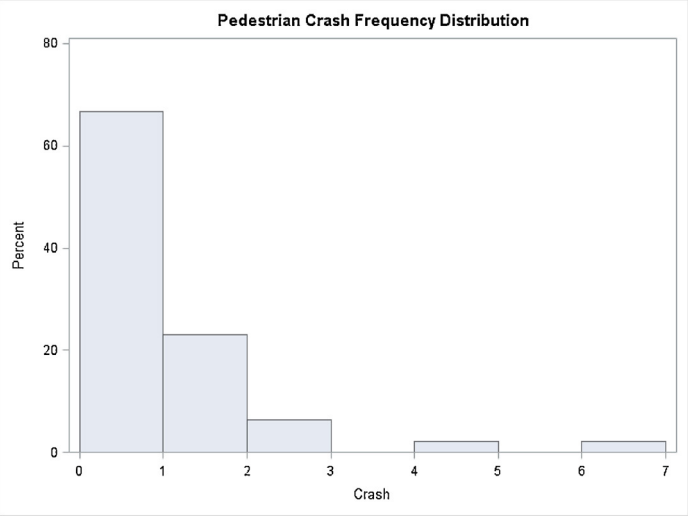

Fig. 2. Pedestrian Crash Frequency Distribution.

Table 1
List of Variables and Data Sources.

Category	Variable name	Data Source
Demographic & Socioeconomic	Population	U.S. Census Bureau
	Age: Below 15 years	
	Age: 15 to 30 years	
	Age: 31 to 45 years	
	Age: 46 to 60 years	
	Age: 61 to 75 years	
	Age: Above 75 years	
	Education: Less than or equal to high school	
	Education: Greater than high school	
	Household size: Less than or equal to 4 persons	
	Household size: Greater than 4 persons	
	Commuters: Walking	
	Commuters: Public transit	
	Household car ownership: Less than two vehicles	
	Household car ownership: Greater than or equal to two vehicles	
	Household below poverty line	
	Proportion of employed people	
Land-use Pattern	Residential Area	Florida Department of Revenue Property Tax Oversight Program
	Commercial Area	
	Industrial Area	
	Agricultural Area	
	School Area	
	Bar Area	
	Hotel Area	
Geometric & Traffic	Number of Intersection legs	Florida Department of Transportation (FDOT) Roadway Characteristics Inventory
	Intersection control type	
	No of lanes on the major road	
	No of lanes on the minor road	
	AADT	
	Maximum speed limit around intersection	
	Average median width	
	Average pavement condition	
	Average sidewalk width	
	Presence of sidewalk barrier	
	Pedestrian crash counts (2013-2015)	
	Number of bus riders	Lynx GIS Data
	Number of bus stops around intersection	

Table 2
Comparison of the exposure models developed.

Model Type (Exposure)	MAD	RMSE
Exposure Model (GLM) Using All Variables	28.91	42.12
Exposure Model (GLM) Using PCA Variables	34.21	46.26
Exposure Model (GLM) Using Random Forests Variables	32.06	45.40
Exposure Model (Tobit) Using All Variables	**27.69**	**41.99**
Exposure Model (Tobit) Using RF Variables	30.76	45.43
Exposure Model (Tobit) Using PCA Variables	32.61	46.24

Table 3
Tobit model result (Best exposure model obtained).

Parameter	Estimate	Std Err	p-value
Intercept	−128.2880	36.3614	0.0004
Presence of school near intersection (1 = yes, 0 = no)	22.3864	8.6643	0.0098
Household car ownership (Number of households with less than two vehicles within buffer of 0.25 miles radius)	81.0957	22.8838	0.0004
Pavement condition (1 = very poor, …, 5 = very good)	11.1645	6.6966	0.0955
Sidewalk width (Average value of all legs)	7.3550	1.9896	0.0002
Bus ridership (Number of bus user within buffer of 0.25 miles radius)	0.1825	0.0683	0.0075
Intersection control type (1 = signal, 0 = stop)	39.4279	11.4672	0.0006
Presence of sidewalk barrier (1 = yes, 0 = no)	26.1963	9.9762	0.0086
σ	44.8521	2.9050	< .0001

the model demonstrates that the presence of schools around the intersection contributes to higher exposure of pedestrians at intersections.

5.2. Household car ownership: (number of households with less than two vehicles within the buffer of 0.25-mile radius)

Households with less than two vehicles (0 or 1 vehicle) are another significant source of pedestrian activities. Car ownership is directly related to household income level that reflects the socio-economic impact on pedestrian activity. It is obvious that household members with no vehicles satisfy their transportation needs by means of public transportation or walking. Also for households with only one vehicle it may not be possible to accommodate all the members with one vehicle due to different travel schedules and trip purposes. The result shows that the more frequent the number of such households there will be more pedestrians exposed at the intersections.

5.3. Pavement condition (1 = very poor, …, 5 = very good)

Better pavement condition around the intersection provides more accessibility for pedestrian walking. The model in this study includes the variable in the scale of 1 to 5 where 1 indicates poor pavement and 5 indicates best pavement. It has been found higher value of pavement condition yields more pedestrians. In the previous studies (Said et al., 2016; Stradling et al., 2007; Suh et al., 2017; Weinberger and Sweet, 2012), the pavement condition is an important factors for walkability. Thus, it might have effects on pedestrian demands. However, p-value of the pavement condition is close to 0.10, and it is possible that it could be insignificant if data from other regions are used.

5.4. Sidewalk width (average value of all legs)

If the pedestrian activity is higher, it is expected to have wider sidewalks to be installed at the particular intersection. In other words, it can be said that if the sidewalk has larger width more pedestrians are walking around the intersection. The model showed that the large sidewalk width can be attributed to more pedestrian activity at the intersection.

5.5. Bus ridership (number of bus users within the buffer of 0.25-mile radius)

Daily Bus users walk through the route and intersections to reach the bus stops. The distance between home and the bus stops leads the bus rider to walk along the route and cross intersections. Most of the bus stops are at a certain distance from the intersection (100–200 ft.) which may lead the pedestrian to be exposed at the intersections. It is likely to be true that these bus users are exposed to intersections two times a day during ride –on and ride-off.

5.6. Intersection control type (1 = signal, 0 = stop)

Signalized intersection control type leads to increasing pedestrian activity since it is relatively safer than the other control types. The study included intersection control type as a dummy variable and the positive coefficient indicates that if the control type is signal it is more likely to have higher pedestrian exposure.

5.7. Presence of sidewalk barrier (1 = yes, 0 = no)

The study incorporated the physical sidewalk barrier like guardrail, traffic barrier, etc. as a dummy variable. Sidewalk barriers are installed in places where there is a probability of more pedestrian activity. In other words, more pedestrians are exposed when crossing even where sidewalk barriers are present. Pedestrians are completely separated from traffic that makes the walking safe. Intersections that have sidewalk barriers are more likely to have more pedestrians.

Most of the models identified almost similar significant exposure-related variables that have been described in the previous section. The predicted pedestrian trips were calculated using the best exposure model. In order to check the significance of the difference between observed pedestrian trips and predicted pedestrian trips, both *t*-test and Wilcoxon signed-rank test were performed. The results of the tests are shown in Table 4.

The difference between the observed and the predicted pedestrian trips does not appear as statistically significant, since the results of both tests yielded a p-value greater than 0.05. It implies that the predicted pedestrian trips could be used in lieu of the actual pedestrian trips if there were no pedestrian counts available in the study area.

Using the predicted pedestrian trips from the Tobit model, negative binomial and ZINB crash models were developed. In these models pedestrian trips have been used as exposure to predict the number of crash at a particular intersection. In order to evaluate how the predicted pedestrian trips perform relative to observed pedestrian trips a similar type of crash model was developed using the original observed pedestrian trips as exposure. The Negative Binomial crash models are described in Tables 5 and 6,.

Finally, the authors developed a Negative Binomial model using the variables from Table 4 that can predict the number of pedestrian crashes at intersections. The factors in Table 3 are exposure factors that does not have a causal effect relationship with the crash. The model results in Table 7 describe the relative effect of the pedestrian exposure to traffic crashes.

All the developed models are summarized in Table 8 based on AIC and BIC values.

Table 8 shows that the crash models from observed and predicted pedestrian trips have close AIC and BIC values for both models. It indicates that the predicted pedestrian trips performed almost the same as the observed pedestrian trips in the crash models. Thus, the approach of surrogate measures for pedestrian exposure in this study can be justified. Although the negative binomial crash model with the variables directly from Table 4 performed best among the developed models but the difference in AIC and BIC values are not very large. It was also found that the NB crash model performed slightly better than the ZINB model in terms of AIC and BIC values although the dataset contained almost 66% of intersections with no pedestrian crashes. In this two-step safety analysis not only the predicted crashes but also the reliable

Table 4
Statistical test to compare observed and predicted pedestrian trips.

Test	Statistic	p-Value
Student's (t)	−1.59558	0.1173
Signed Rank (S)	−160	0.1013

Table 5
NB crash model using predicted trips from exposure model.

Parameter	Estimate	Standard Error	p-value
Intercept	−0.961831	0.254413	0.0002
Predicted Pedestrian Trips	0.016512	0.003870	< .0001
Dispersion	0.851405	0.276449	.

Table 6
NB crash model using original pedestrian trips as exposure.

Parameter	Estimate	Standard Error	p-value
Intercept	−0.667031	0.179431	0.0002
Observed Pedestrian Trips	0.010164	0.002113	< .0001
Dispersion	0.728411	0.259032	.

Table 7
NB crash model using the variables from Table 3.

Parameter	Estimate	Standard Error	Pr > ChiSq
Intercept	−2.488378	0.734974	0.0007
Bus ridership (Number of bus user within buffer of 0.25 miles radius)	0.005430	0.001589	0.0006
Intersection control type (1 = signal, 0 = stop)	2.201391	0.742612	0.0030
Presence of sidewalk barrier (1 = yes, 0 = no)	0.517032	0.255692	0.0432
Dispersion	0.633611	0.238670	.

Table 8
Comparison of developed crash models.

Negative Binomial (NB) Crash Models	AIC	BIC
NB using predicted trips from pedestrian exposure model (Tobit model using all variables)	355.4595	364.1530
NB using observed pedestrian trips	351.2452	359.9387
NB crash model using the variables from the optimal tobit model (Table 4)	346.3255	360.8147

Zero-Inflated Negative Binomial (ZINB) Crash Models	AIC	BIC
ZINB using predicted trips from pedestrian exposure model (Tobit model using all variables)	356.0369	367.6282
ZINB using observed pedestrian trips	352.7862	367.2754
ZINB crash model using the variables from Table 4	348.3255	365.7125

predicted pedestrian trips can be obtained.

6. Summary and conclusions

Over the past decades, practitioners have adopted various approaches to develop knowledge and guidelines for pedestrian safety. In recent, many researchers have focused on pedestrians' safety by identifying risk factors. However, depending only on crash counts without considering the underlying risk factors may provide limited and even biased information about where safety measures are needed and what measures would be best to be used to improve pedestrian safety.

In this study, a systemic approach has been developed that uses pedestrian surrogate measures in terms of exposure information. The two-steps procedure described in the study can be utilized in cases where it is difficult to collect the pedestrian volume data. The study recommends using the two-steps procedure as it provides important exposure information to better understand the pedestrian activity and crash frequency relation. The study applied negative binomial and zero-inflated negative binomial models in micro-level pedestrian safety analysis and found that the negative binomial model performed slightly better. Apart from these, the identification of the measures affecting pedestrian exposure such as the presence of school, car-ownership, pavement condition, sidewalk width, bus ridership, intersection control type, and presence of sidewalk barrier was another important contribution of the study to the pedestrian safety research approach. The study emphasizes focused on these factors while determining the safety measures for pedestrians. For example, if there is a school near an intersection, strict enforcement of speeding and traffic calming is required, and also pedestrian education programs should be promoted for the school children (Abdel-Aty et al., 2007).

There are several possible extensions to this study that could be made in the future. The proposed two-steps procedure in this study involves two consecutive modeling processes. The first model estimates the number of pedestrians; and the second model estimates pedestrian crash counts using the predicted pedestrian counts from the first model. Nevertheless, the procedure has a limitation as the result can be biased due to the accumulated errors from the first step. It is possible that the issue can be overcome by adopting simultaneous modeling approach. The influential area around each intersection (buffer size) considered in the study for extracting socio-economic and land-use variables was determined based on the average walking distance (Boer et al., 2007; Yang and Diez-Roux, 2012). However, it is common to have a longer trip distance than the average walking distance. It may be possible that the variables found to be insignificant in the current study may become significant using other buffer sizes. The study can be done using different buffer sizes to identify other significant factors. In such case mixed models or random variables would be better approach for the analysis.

It is expected that the study can be a good platform for further analysis of pedestrian exposure not only at intersections but also on segments with surrounding environments. The methodologies implemented in the study including surrogate exposure variables can estimate safety performance functions for pedestrian crashes even though pedestrian trip data is not available. With developed the pedestrian safety performance functions, it would be possible to identify crash hotspots for pedestrians and provide appropriate countermeasures to prevent pedestrian crashes.

References

Abdel-Aty, M., Chundi, S.S., Lee, C., 2007. Geo-spatial and Log-linear analysis of pedestrian and bicyclist crashes involving school-aged children. J. Saf. Res. 38 (No. 5), 571–579.

Boer, R., Zheng, Y., Overton, A., Ridgeway, G.K., Cohen, D.A., 2007. Neighborhood design and walking trips in Ten us metropolitan areas. Am. J. Prev. Med. 32 (No. 4), 298–304.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (No. 1), 5–32.

Cai, Q., Lee, J., Eluru, N., Abdel-Aty, M., 2016. Macro-level pedestrian and bicycle crash analysis: incorporating spatial spillover effects in dual State count models. Accid. Anal. Prev. 93, 14–22.

Cervero, R., 1996. Mixed Land-uses and commuting: evidence from the American housing survey. Trans. Res. Part A: Policy Pract. 30 (No. 5), 361–377.

Davis, D.G., Braaksma, J.P., 1988. Adjusting for luggage-laden pedestrians in Airport terminals. Trans. Res. Part A: Gen. 22 (No. 5), 375–388.

Graham, D.J., Stephens, D.A., 2008. Decomposing the impact of deprivation on child pedestrian casualties in England. Accid. Anal. Prev. 40 (No. 4), 1351–1364.

Keall, M.D., 1995. Pedestrian exposure to risk of Road accident in New Zealand. Accid. Anal. Prev. 27 (No. 5), 729–740.

Lam, W.W.Y., Loo, B.P.Y., Yao, S., 2013. Towards exposure-based time-space pedestrian crash analysis in facing the challenges of ageing societies in Asia. Asian Geogr. 30 (No. 2), 105–125.

Lam, W.W., Yao, S., Loo, B.P., 2014. Pedestrian exposure measures: a time-space framework. Travel Behav. Soc. 1 (No. 1), 22–30.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34 (No. 1), 1–14.

Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida. Accid. Anal. Prev. 37 (No. 4), 775–786.

Lee, J., Abdel-Aty, M., 2017. Macro-level analysis of bicycle safety: focusing on the characteristics of both crash location and residence. Int. J. Sustain. Trans. 1–8.

Lee, I.-M., Buchner, D.M., 2008. The importance of walking to public health. Med. Sci. Sports Exerc. 40 (No. 7 Suppl), S512–518.

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

Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. Accid. Anal. Prev. 102, 213–226.

Lee, J., Abdel-Aty, M., Cai, Q., Wang, L., Huang, H., 2018a. Integrated modeling approach for non-motorized mode trips and fatal crashes in the framework of transportation safety planning. Trans. Res. Rec 0361198118772704.

Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M., Cai, Q., 2018b. Analysis of crash proportion by vehicle type at traffic analysis zone level: a mixed fractional Split multinomial logit modeling approach with spatial effects. Accid. Anal. Prev. 111, 12–22.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Trans. Res. Part A: Policy Pract. 44 (No. 5), 291–305.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37 (No. 1), 35–46.

Lord, D., Washington, S., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accid. Anal. Prev. 39 (No. 1), 53–57.

Miranda-Moreno, L.F., Morency, P., El-Geneidy, A.M., 2011. The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections. Accid. Anal. Prev. 43 (No. 5), 1624–1634.

National Highway Traffic Safety Administration, 2018. Pedestrians: 2014 Data (Traffic Safety Facts. Report No. Dot Hs 812 270). Accessed 5 July, 2016 01:18 UTC. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812270.

Olsson, U., 2002. Generalized Linear Models. An Applied Approach 18. pp. pp Studentlitteratur, Lund.

Qin, X., Ivan, J., 2001. Estimating pedestrian exposure prediction model in rural areas. Trans. Res. Rec.: J. Trans. Res. Board (No. 1773), 89–96.

Rose, C.E., Martin, S.W., Wannemuehler, K.A., Plikaytis, B.D., 2006. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. J. Biopharm. Stat. 16 (No. 4), 463–481.

Said, M., Abou-Zeid, M., Kaysi, I., 2016. Modeling satisfaction with the walking environment: the case of an Urban university neighborhood in a developing Country. J. Urban Plan. Dev. 143 (No. 1) pp. 05016009.

Son, H.D., Kweon, Y.-J., Park, B.B., 2011. Development of crash prediction models with individual vehicular data. Trans. Res. Part C: Emerg. Technol. 19 (No. 6), 1353–1363.

Stradling, S.G., Anable, J., Carreno, M., 2007. Performance, importance and user disgruntlement: a six-step method for measuring satisfaction with travel modes. Trans. Res. Part. A: Policy Pract. 41 (No. 1), 98–106.

Suh, W., Kim, B.S., Yurk, Y., 2017. Walkability assessment for elderly citizens and people with disabilities. Int. J. Eng. Technol. 9 (No. 4), pp.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica: J. Econometric Soc. 24–36 Chicago.

Ukkusuri, S., Hasan, S., Aziz, H., 2011. Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. Trans. Res. Rec.: J. Trans. Res. Board (No. 2237), 98–106.

Van den Bossche, F., Wets, G., Brijs, T., 2005. Role of exposure in analysis of Road accidents: a Belgian case study. Trans. Res. Rec.: J. Trans. Res. Board (No. 1908), 96–103.

Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. CRC press. Statistical and Econometric Methods for Transportation Data Analysis.

Weinberger, R., Sweet, M., 2012. Integrating walkability into planning practice. Trans. Res. Rec.: J. Trans. Res. Board (No. 2322), 20–30.

Wier, M., Weintraub, J., Humphreys, E.H., Seto, E., Bhatia, R., 2009. An Area-level model of vehicle-pedestrian injury collisions with implications for Land use and transportation planning. Accid. Anal. Prev. 41 (No. 1), 137–145.

Woolson, R., 2008. Wilcoxon Signed-Rank Test. Wiley Encyclopedia of Clinical Trials pp.

Yang, Y., Diez-Roux, A.V., 2012. Walking distance by trip purpose and population subgroups. Am. J. Prev. Med. 43 (No. 1), 11–19.