A stop safety index to address pedestrian safety around bus stops

Mehmet Baran Ulak a, b, Ayberk Kocatepe b, Anil Yazici c, Eren Erman Ozguven d, Ashutosh Kumar b

a Department of Civil Engineering, University of Twente, Enschede 7522 NB, Netherlands
b Connects Transportation Group, 3111 N University Drive, Suite 725, Coral Springs, FL 33065, United States
c Department of Civil Engineering, Stony Brook University, Stony Brook, NY 11794, United States
d Department of Civil and Environmental Engineering, FSU College of Engineering, Tallahassee, FL 32310, United States

ARTICLE INFO

Keywords:
Bus stop safety index
Pedestrian safety
Public transportation
Spatio-statistical analysis
Decision tree modeling

ABSTRACT

Despite the decline in the numbers of crashes and fatalities in the U.S. since 1990, pedestrian crashes have been steadily increasing and reached its 28-year peak in 2018. This increase led to initiatives such as Vision-Zero in response to this deterioration in pedestrian safety. In spite of the severe outcomes of pedestrian crashes, guidelines are still not fully capable of alleviating pedestrian safety issues and to formulate safety performance functions; mainly due to the scarcity of pedestrian data, particularly the pedestrian counts. However, pedestrian safety is a critical concern; hence safety of pedestrian facilities is also needed to be quantified. With this need in mind, this study proposes a safety index for public transportation bus stops which are facilities that are heavily utilized by the pedestrians. For this purpose, this paper first shows that there is a significant spatio-statistical correlation between the bus stop locations and pedestrian-involved crashes. Then, a bus stop safety index (SSI) is proposed in order to quantify and assess pedestrian safety around bus stops. Finally, a regression tree model is also developed for SSI scores (in a fashion similar to safety performance functions) in order to make the SSI available to practitioners who do not have access to relevant software and pedestrian crash data. Overall, the developed SSI measure can be used as a screening metric which can rank the pedestrian safety around the bus stops, and help identify high-risk locations in a proactive manner before the pedestrians become crash statistics.

1. Introduction

The number of crashes in the U.S., especially fatal crashes, has been steadily declining since 1990 (Wagner, 2018). However, a similar safety improvement is not observed for pedestrian crashes. In 2018, the number of pedestrian fatality crashes reached its 28-year peak since 1990 (LeBeau, 2019), and triggered new proactive initiatives such as Vision-Zero (Vision Zero, 2019). Vision-Zero and similar initiatives aim to decrease the severity of the unavoidable crashes, assuming that the roadway users will make mistakes regardless of the precautions. The recommended strategies include those rather related to generic traffic flow and roadway design aspects (e.g., lowering speed limits) and do not target specific pedestrian facilities (e.g., crosswalk, sidewalk, overpass). Moreover, there is not a detailed safety manual focusing on pedestrian safety (similar to Highway Safety Manual - HSM) that can help screen the roadway network with respect to facility type. Thus, pedestrian safety improvements are mostly performed either with specific after-the-fact evaluations (e.g., physical design changes around crash hotspots), or with generic precautions such as reducing speed limits.

The lack of quantitative guidelines/manual for pedestrian safety is mainly a result of the scarcity of pedestrian data, particularly the pedestrian counts, needed to formulate safety performance functions. Especially with the widespread ITS deployments, researchers have access to comprehensive traffic flow and crash datasets which help develop accurate safety assessment and performance functions. In almost all equations given in HSM, traffic flow data play an important role to reflect the utilization level of the roadways and function as a normalizing variable for safety measures, e.g. crashes per 100 million vehicle miles traveled. The availability of data leads to SPFs that are custom-formulated for all facility types, ranging from highways to rural roadways. Unfortunately, pedestrian flow sensor deployments and data are at miniscule levels compared to those from the vehicle domain. Most of the existing pedestrian flow/count data are either for very specific localities that limit generalization, or they are extrapolated based on a
limited number of counts that reduces reliability. Consequently, it is difficult to perform network screening of various pedestrian facilities with standardized measures, as it is done for highways. Thus, there is a need for screening metrics that will rank the pedestrian facilities with respect to safety, and help identify high-risk locations in a proactive manner before the pedestrians become crash statistics.

Tracking pedestrian movements are rather difficult compared to vehicles. It is still not common to measure pedestrian movements using sensors in a similar fashion to motor vehicles, which makes it difficult to measure the pedestrian “exposure” on the roadways. In addition, pedestrian movements are not necessarily constrained by the roadway geometry and markings, e.g. irregular actions such as jay-walking. However, pedestrian safety is a critical concern; hence safety of pedestrian facilities such as crosswalks or public transportation stops is needed to be quantified. With this need in mind, this study proposes a safety index for public transportation bus stops as facilities that are heavily utilized by the pedestrians. Similar to the safety performance functions that quantify safety of a roadway facility (e.g., four-way intersections) by aggregating crashes in proximity of that facility, pedestrian crashes around bus stops are utilized to quantify the safety of bus stops. The choice of bus stops as the target pedestrian facility is premised on one conceptual consideration and one statistical finding:

1) The pedestrians around bus stops can be the transit users (i.e., waiting at, arriving to, or leaving bus stops) or random passersby. Nevertheless, the bus stops are, by design, located at heavy pedestrian activity locations such as residential or commercial areas (Mistretta et al., 2009). Considering the lack of pedestrian count data (and hence no other way to screen for risky roadway sections), screening high risk bus stops to improve their safety would definitely enhance the overall pedestrian safety. In other words, the safety improvement around bus stops does not affect only the transit riders, but all pedestrians.

2) As presented in this study as a finding, there exists a statistically significant spatial relationship between the location of bus stops and the pedestrian-involved crashes, i.e., the pedestrian involved crashes are more likely to happen closer to the bus stops.

In this context, the formulated bus Stop Safety Index (SSI) assigns scores to the bus stops based on the severity of pedestrian-involved crashes in proximity. SSI can help transportation agencies screen the urban roadway network for pedestrian safety and identify the high risk bus stops for further safety assessment. In order to make the SSI available to practitioners, a regression tree model is also developed to tabulate the SSI scores and made available to practitioners in a fashion similar to safety performance functions. The decision tree results can easily be implemented in a spreadsheet format and a SSI value for any given bus stop can be determined based on generally available data input, such as U.S. Census demographics and socio-economics, transit and traffic operations data, and facilities in the study region.

The paper outline is as follows: First, a literature review is provided, followed by the details of the data utilized for the study. Second, the statistical tests are performed to show the spatio-statistical correlation between the bus stop locations and pedestrian-involved crashes. Third, the bus stop safety index () is formulated and a regression tree model is estimated to tabulate the SSI scores with specifically practitioner use in mind. Last, the conclusion and future research directions are discussed.

2. Literature review

The safety of transit riders and pedestrians has been analyzed using spatial and statistical approaches (Hess et al., 2004; Ulak et al., 2019, 2018; Vogel and Pettinari, 2002; Volinski and Tucker, 2003; Weiner and Singa, 2006). For instance, Hess et al. (2004) investigated pedestrian crashes on the bus transit corridors. Authors identified an association between pedestrian crashes and bus stop usage. Similarly, Pulugurtha and Vanapalli (2008) identified hazardous bus stops using pedestrian crashes through a GIS-based visual analysis. However, authors did not conduct a statistical analysis to verify their findings. Another GIS-based study to identify unsafe bus stops was conducted by Tuong and Somenahalli (2011). In their study, authors conducted a spatial autocorrelation analysis using Moran’s I statistics and Getis-Ord GI* statistics to identify severity-based pedestrian crash hotspots. Then, bus stops which fall inside of these hotspots were determined and a severity index was calculated based on the severity-weighted pedestrian crashes within those hotspots.

Predicting or modeling the level of safety of bus stops through various factors is also an important research direction. For example, a level of safety modeling for bus stops in China was proposed by Ye et al. (2016). The authors used several predictors such as number of boarding, bus frequency, traffic-related factors (e.g., pavement condition, signs, and markings), and lighting conditions to quantify the safety of bus stops using traffic conflict data. However, their sample size of 46 was relatively small. In addition to the factors associated with traffic and roadway geometry, it was shown that pedestrian crashes were also associated with the presence of facilities such as hospitals, supermarkets, and religious facilities (Ulak et al., 2018). Furthermore, Zhao et al. (2013) showed that public transportation ridership was significantly associated with the locations of education buildings and shopping centers. This finding combined with study of Ulak et al. (2018) indicates that pedestrians around public transportation bus stops and stations might be exposed to an elevated crash risk.

To evaluate the safety of bus stops, Amadori and Bonino (2012) developed a software for Italian bus stops using a risk assessment based on traffic and geometric features. In their study, they used a survey to collect bus stop features and accidents within the 30 m of these stops. The software has been used to improve bus stops which impose safety risks. The effect of bus stop location (e.g., distance to intersection) on the traffic safety was also investigated and results indicate a relationship between the bus stop location and pedestrian crashes (Eom et al., 2014). Authors used this information to develop a model in order to optimize the bus stop locations on a given roadway. It is reasonable to expect more pedestrian traffic around bus stops which have high ridership; hence bus ridership might have an impact on the pedestrian crashes due to this increased exposure. Therefore, it is important to identify factors driving the ridership. For instance, Chakour and Eluru (2016) found that ridership increased with a higher frequency of buses while residential areas had a higher ridership than that of commercial zones. The study of Jun et al. (2015) also found that population and employment densities as well as availability of intermodal connectivity positively affected the ridership. Moreover, they identified that 600 m was an appropriate radius for pedestrian catchment areas with a focus on the subway ridership.

Statistical approaches have also been commonly used to analyze the safety of transit riders and pedestrians. Linear regression is one of the most common techniques used to understand the relationship between bus stops and the surrounding area (Ye et al., 2016; Zhao et al., 2013). Recently, methods addressing the unobserved heterogeneity issue are increasingly adopted by safety studies due to the effect of unobserved characteristics on statistical estimation (Manning et al., 2016). Later studies focusing on the pedestrian safety showed the necessity of accounting for the unobserved heterogeneity issue and adopted mixed models and random parameters models for statistical estimations (Behnood and Manning, 2016; Sarwar et al., 2017b, 2017a; Xin et al., 2017). Nevertheless, Manning et al., (2020) argued potential complexity and dimensionality challenges posed by heterogeneity...
Another recently popular approach in transportation studies is the decision trees. A recently popular approach in transportation studies is the decision trees (Breiman et al., 1984). Decision trees (i.e. classification or regression trees) are nonparametric models that predict the outcome of an event through a flowchart structure built based on the values of the predictors. In the transportation domain, decision trees were used to model and analyze temporal variations in traffic patterns (Kamga and Yazici, 2014), travel time reliability based on day of week and time of day (Yazici et al., 2012), incident-induced capacity reductions (Almotahari et al., 2019), traffic crashes (Olutayo and Eludire, 2014; de Ona et al., 2013; Zheng et al., 2016) and crash severity (Abellan et al., 2013).

Despite the severe outcomes of pedestrian crashes, studies that focus on the safety of individuals using public transportation are limited. One recent example is the study of Yoon et al. (2017), which focuses on the bus-involved crashes. Yoon et al. (2017) investigated the injury severity associated with the local bus crashes using a hierarchical ordered model. They used individual and regional level variables such as vehicle speed, vehicle age, roadway geometry, traffic rate, healthcare employee per residents, and ratio of seniors in the whole population.

The approach adopted in our research is similar to Truong and Somenahalli (2011) in terms of developing a network distance-based severity index for bus stops. However, to the authors’ knowledge, none of the studies in the literature delved into the spatio-statistical correlation between pedestrian-involved crashes and bus stop locations, and developed a KABCO scale and decay function-based safety index for bus stops.

3. Approach

The overall approach consists of three steps and it is illustrated in Fig. 1.
1) First, the significant spatio-statistical relationship between the bus stop locations and pedestrian-involved crashes is shown for the study area, namely Palm Beach County, FL. For this purpose, Global Cross K-function method (Okabe et al., 2006; Okabe and Sugihara, 2012) is used. The pedestrian crash data and bus stop locations are obtained from Florida Department of Transportation (FDOT) (FDOT, 2015) and Southeast Florida (SEFL) STOPS Planning Model, respectively. Fig. 2 shows the study area and the spatial distribution of the pedestrian-involved crashes. The STOPS was developed in 2016 to supplement the need for ridership forecasts for transit projects in South Florida, and uses readily-available input data from Census, ACS, Metropolitan Planning Organization (MPO), and General Transit Feed Specification (GTFS).

2) Based on the verified significance, SSI values were calculated through integrating 1) network-based distances between bus stops and pedestrian-involved crashes, 2) KABCO severity levels (Herbel Fig. 3. Examples of interpolated demographic variables through Kriging.
et al., 2010) to weight the severity of pedestrian-involved crashes, and 3) a decay function to weight the effect of distances between these bus stops and crashes.

3) The SSI scores assign numerical values on the spatio-statistical relationship between the bus stop locations and pedestrian-involved crashes for the study area. In order to provide a more generalized, practitioner friendly measure that can be applied elsewhere, a regression tree was developed to estimate the SSI measure of a bus stop using a set of factors including:

- Bus Frequency and Daily Boarding (STOPS, 2016): Bus stop metric data showing stop utilization was gathered from the STOPS model. STOPS is a limited implementation of the conventional “4-step” travel demand model, which is developed by the Federal Transit Administration (U.S. DOT, 2019). For this study, the majority of data was obtained from the SEFL STOPS Planning Model, which is a part of the Florida Statewide and Regional Planning Model (SERPM 7.061).
- Total population around bus stops as a measure of urban density.
- Total employment around bus stops as a measure of pedestrian attraction.
- Socio-demographic factors (i.e., income and age groups) provide the characteristics of potential transit users that may have an impact on the crash involvement.
- Facilities (i.e., supermarket, hospital, school) that lead to an increase in the pedestrian activity.
- Traffic indicators (i.e., volume, speed, roadway type) as general characteristics affecting crash severity.

In essence, the decision tree utilizes the available “proxy” data and yields tabulated scores. The tabulated scores can help practitioners estimate SSI values for their localities and scan their network for bus stops that have higher probability of pedestrian-involved crashes in proximity. The summary and definitions of the variables used in the analysis are provided in Table 1.

Please note that census units for demographic variables are demarcated by roadways. The bus stops are also points on those roadways that form the census unit boundaries. Hence, a bus stop cannot be assigned to a particular census unit. Therefore, a Kriging interpolation method is used to assign the demographics data to a bus stop location by interpolating the values of census units surrounding a bus stop (Fig. 3). This Kriging method utilizes geostatistical techniques to create the interpolation surface from scattered point data (Oliver and Webster, 1990). An ordinary spherical semivariogram and 4 nearest neighbors are used as analysis parameters (ArcGIS, 2016).

4. Spatio-statistical correlation between the bus stops and pedestrian-involved crashes

One of the main premises of this study is the identification of the spatial correlation between the bus stop locations and pedestrian-involved crashes. In order to determine whether there is a significant spatio-statistical correlation, the Global Cross K-function method (available in the SANET tool) is used (Okabe et al., 2006; Okabe and Sugihara, 2012). The Cross K-function method is based on the Ripley’s K-function developed for testing the spatial randomness of a set of points distributed over a plane or a network (Lamb et al., 2016; Okabe and Yamada, 2001). As an extension, the Cross K-function determines if the spatial distribution of a set of points are statistically independent from the spatial distribution of another set of points. In this study, the following null hypothesis is tested:

\[ H_0: \text{The spatial distribution of pedestrian-involved crashes along the network is correlated with the bus stop locations.} \]

\[ H_1: \text{The alternative hypothesis is that the pedestrian-involved crashes are distributed via a complete spatial randomness (CSR) along the network and that they are independent of bus stop locations.} \]

The Cross K-function approach tests the null hypothesis by comparing the distances between the bus stops and actual pedestrian crashes, with the bus stop distances of randomly generated crash locations along the roadways. The Cross K-function is calculated by the formula below (Okabe and Sugihara, 2012):

![Fig. 4. Global Cross K-function plot.](image)
KABCO classification and severity weights (Herbel et al., 2010).

<table>
<thead>
<tr>
<th>Injury Severity Level</th>
<th>Crash Costs</th>
<th>Severity Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatality (K)</td>
<td>$4,008,900</td>
<td>541.74</td>
</tr>
<tr>
<td>Incapacitating Injury (A)</td>
<td>$216,000</td>
<td>29.19</td>
</tr>
<tr>
<td>Non-incapacitating Injury (B)</td>
<td>$79,000</td>
<td>10.68</td>
</tr>
<tr>
<td>Possible Injury (C)</td>
<td>$44,900</td>
<td>6.07</td>
</tr>
<tr>
<td>No Injury (O)</td>
<td>$7,400</td>
<td>1.00</td>
</tr>
</tbody>
</table>

$K_{ij}(d) = \frac{1}{\rho_j} \sum_{l=1}^{n_j} n_l(d_{pl})$  \hspace{1cm} (1)

where $d$ is the network-based distance; $n_l(d_{pl})$ is the number of pedestrian crashes ($l$) within the $L(d_{pl})$; $L$ is the subnetwork that includes network sections stretching $d$ distance from the bus stop ($J_l$); $\rho_j = n_j/L$ is the density of pedestrian crashes; $L$ is the total length of all subnetworks $L(d_{pl})$; and $n_j$ is the total number of bus stops. The Cross K-function $K(d)$ is calculated for all $d$ values from 0 to 8 miles (a predefined large distance for the study area and topic), and tested against complete spatial randomness hypothesis for each distance bin.

Cross K-function analysis (Fig. 4) indicates that bus stop locations and pedestrian-involved crashes are spatially-correlated with each other at a 95% confidence level. That is, the observed Cross K-function curve is above the expected Cross K-function at the 95% confidence level for every distance bin. This result indicates that the null hypothesis ($H_0$) cannot be rejected, therefore the pedestrian-involved crashes are spatially correlated with (rather than independent of) bus stops. This finding further implies that bus stop proximity has pedestrian safety implications; thus justifies the subsequent effort to develop the bus stop safety index for pedestrian safety around bus stops.

5. An index for pedestrian safety around bus stops

The bus Stop Safety Index (SSI) is developed to identify the bus stops with high pedestrian-involved crash probability in proximity. For this purpose, each bus stop is assigned a SSI score based on injury severities and spatial decay function as follows:

$$SSI_j = \sum Sev_i \times \text{Decay}(d_{ij})$$  \hspace{1cm} (2)

where $SSI_j$ is the stop safety index of bus stop $j$, $Sev_i$ is the severity weight of pedestrian $i$ calculated based on KABCO scale, $d_{ij}$ is the network distance between bus stop $j$ and pedestrian $i$, and $\text{Decay}(d_{ij})$ is the decay function value calculated based on the distance between $i$ and $j$.

5.1. Distance between bus stops and pedestrian crashes

First, network distance-based origin (bus stops) – destination (crash locations) matrix is formed. For this purpose, three data components are obtained from the Florida Standard Urban Transportation Model Structure (FSUTMS) provided by the Florida Department of Transportation (FSUTMS, 2017): (a) bus stop locations – origins, (b) pedestrian-involved crash locations – destinations, and (c) the statewide roadway network that connects these origins and destinations. Shortest path network distance is calculated for each origin–destination (O-D) pair using ArcGIS software (ESRI, 2014). For this purpose, a “closest facility” analysis is performed and the corresponding distances for each pair are recorded in miles.

5.2. KABCO severity level of pedestrian crashes

Developing a severity index requires the standardization of different crash and/or severity categories via weights, e.g., how many non-incapacitating injuries are equivalent to a single fatality. In order to create the necessary equivalencies, KABCO weights of each pedestrian crash are identified based on the classification provided by FHWA (Herbel et al., 2010). KABCO is an injury classification scale used to determine the monetary value of each crash outcome (K: fatality; A: incapacitating injury; B: non-incapacitating injury; C: possible injury, e.g., no visible injury but complains of pain; O: no injury). The severity weights are calculated by normalizing the monetary values of each severity level with the monetary values of the property damage-only crashes (Table 2). KABCO-based weights ensure that pedestrians who were severely injured have more impact on the stop safety score than less severe crashes. For example, a pedestrian who suffered an incapacitating injury was given a weight of 29.19 whereas a pedestrian who did not sustain any visible injury was given a weight of 6.07.

5.3. Decay function

A crash closer to a bus stop would be more likely to have impact on the bus stop safety compared to another crash which is further away. As such, decay functions are shown to be valid approaches to weight and diminish the effects of events based on a cost such as distance (Iacono et al., 2008; Kocatepe et al., 2017; Pulugurtha and Agurla, 2012). Therefore, a decay function is created to weight the effect of each pedestrian-involved crash, instead of matching bus stops and crashes irrespective of the distance. The formulated decay function ensures that the effect of each pedestrian-involved crash on the safety index of bus stops decreases with increasing distance.

Developing a decay function for an event requires information on how that event (or effect of the event) diminishes based on a given cost. In this study, we use the information on walking habits of the people

![Decay function](image-url)
who are potentially using the bus stops (i.e., people who are subjected to the crash risk). Hence, the likeliness of a pedestrian-involved crash being associated with a bus stop is assumed to diminish with a trend similar to walking distances for the bus stops. Literature presents some statistics on the pedestrian walking distances from public transportation stops to their homes and offices, and vice versa (Iacono et al., 2008). For instance, it was shown that average walking distance to a bus stop was equal to 0.35 miles where 25% of the passengers walked less than 0.15 miles and 75% of passengers walked less than 0.5 miles (Daniels and Mulley, 2013). Moreover, an analysis of walking patterns to bus stop locations showed that 10.5%, 22.5%, 12%, 34%, 14.5% of people walked, respectively: 0–400 m (0.25 mi), 400–800 m (0.50 mi), 800–1000 m (0.63 mi), 1000–2000 m (1.25 mi), and 2000–3000 m (1.88 mi) in order to reach the bus stops (Ker and Ginn, 2003). A survey study, which includes 328 pedestrians walking to bus stops on weekday mornings, found out that the average walking distance of these pedestrians was equal to 0.5 miles (Agrawal et al., 2008). In light of these studies, an empirical decay function is developed in order to estimate the distribution of people walking to and from the bus stops. To achieve this, an exponential function is used (Iacono et al., 2008) as shown in Eq. (1).

\[
\text{Decay}(d_i) = a \exp \left( - \left( \frac{d_i - b}{c} \right)^2 \right)
\]  

(3)
where \( d_{ij} \) is the network distance between bus stop \( j \) and pedestrian crash \( i \). \( a = 1.024 \), \( b = -0.093 \), and \( c = 0.685 \) are function parameters. The resulting function and parameter values of decay function are shown in Fig. 5.

6. SSI case study

In order to showcase the practical implementation of the SSI, the pedestrian safety around the Palm Beach area bus stops is analyzed. The calculated bus SSI values are shown in Fig. 6. The histogram and cumulative distribution function along with descriptive statistics of the SSI values are provided in Fig. 7. Note that SSI values range from 0 to 1,650. Considering that KABCO value for a fatality is 541.74, a SSI value closer to 1,650 is equivalent to three pedestrian fatalities in the vicinity (less than 0.05 miles based on the decay function) of a bus stop location. This is a very alarming finding from a pedestrian-focused safety perspective. Fig. 6 shows that bus stops with high SSI values are clustered in some regions of the study area rather than being spatially dispersed. In particular, three regions in the study area exhibit substantially high SSI values:

1) Cross county plaza region is located at the intersection of Okeechobee Blvd and Military Trl. This region is one of the priority investment areas based on the Palm Beach Transit Development plan (PalmTran, 2016) and is very close to the Palm Beach International Airport. Additionally, the Okeechobee Boulevard has been identified as a transit-oriented corridor due to the high shares of low-income and zero-vehicle households in the area.

2) Lake Worth Corridor is at the intersection of Military Trail and Lake Worth Rd. This intersection is also one of the high transit-oriented areas due to high population density, heavy labor force density (PalmTran, 2016) and high percentage of zero-vehicle households. Accordingly, the corridor is assigned a high priority for further bus service improvements.

3) The intersection of US-1 and Forest Hill Blvd is the third major location given high SSI values. Note that the area between the US-1 and Turnpike to the west, at which the intersection is located, has a significantly higher population than other locations.

Overall, the identified regions are generally low income residential areas with a high population. Note that using public transportation is common among lower income groups; hence pedestrian safety deteriorates around bus stops located in lower income neighborhoods. This indicates serious social equity issues. SSI scores successfully identify such regions with heavy pedestrian activity and transit use, which are more likely to have elevated crash risks for pedestrians. Nevertheless, the SSI’s novelty is to rank the bus stops located at those high risk regions; thus agencies can efficiently direct their efforts at specific locations to analyze and enhance pedestrian safety.

7. Estimating and tabulating the SSI scores through a regression tree

SSI methodology provides a practical tool to rank the bus stops in a region based on the pedestrian safety, and can help agencies in planning appropriate measures to improve safety. Nonetheless, SSI calculation compels the use of rather specialized tools (such as network analysis) and crash data that may not be available depending on the locality. In order to make SSI available to researchers who do not have access to the
necessary technical tools and data resources, a regression tree model is developed to tabulate the SSI values with respect to commonly available data, such as U.S. Census demographics and socio-economics, transit and traffic operations data, and facilities in the study region.

Regression trees are nonparametric models (i.e., no need for distributional assumptions), and they are not affected by outliers, collinearities, or heteroscedasticity (Breiman et al., 1984). One of the important features of regression trees is the easiness to interpret them and tabulate the results. A decision tree constructs nodes that partition the data to create homogeneous groups of observations. Once constructed, the regression estimations can be obtained by answering a yes/no question at each node. For example, Fig. 8 shows a very simple regression tree on estimating the traffic volume of a roadway segment. In this example, if the answer is "≥ 3" to "Number of Lanes," "≥ 45" to "Speed Limit," and "Urban" to "Region," then the volume estimated should be 60,000. However, if the answer to the last question is "rural," then the volume estimate should be 25,000.

Regression tree algorithms are well-established in the literature and there are several available software tools such as Matlab, R, or Python that is capable of conducting regression tree analysis. In this study, we used fitrtree function of Matlab’s statistics toolbox with “interaction-curvature” predictor variable selection algorithm. Regression trees are non-parametric approaches; thus parametric testing methods (e.g., t-test statistics) cannot be directly used to assess the statistical significance of predictor variables. Nonetheless, the predictor selection in tree models is achieved by alternative approaches such as “tree pruning” which prevents overfitting problem associated with tree models. The option of “interaction-curvature” prioritizes the predictor importance that help the interpretation of results (Mathworks, 2019) and fits the purposes of this study. Interaction-curvature utilizes chi-square test to select predictors splitting the tree based on the independence between each predictor and the response. In the final estimated tree, the number of the nodes of the tree depends on the data and the criteria to prune the tree (i.e., reducing number of nodes). That is, the tree can expand fully to represent the response variable perfectly by developing a very complex structure tailored exactly for the training data. However, the fully expanded tree may fail to achieve a successful prediction for another dataset. On the other hand, increasing complexity leads to interpretation difficulties. This overfitting problem can be alleviated by optimizing the number of nodes through a cross-validation analysis. For this purpose, cross-validation errors (cost) are calculated for all trees starting from one node to full tree (i.e., pruning levels). Then, optimal tree can be found by identifying the pruning level (number of nodes) which is 1 standard deviation away from the minimum-cost tree, as suggested by Breiman et al. (1984). In this study, the optimal tree is identified through a 1000-fold cross-validation. Fig. 9 shows the cross-validated errors and residual errors that are used to identify the optimal tree size of 298 nodes.

### Table 3

<table>
<thead>
<tr>
<th>SSI Scores</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>10</th>
<th>1500</th>
<th>1400</th>
<th>1300</th>
<th>1180</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median Income</td>
<td>&lt; $42,000</td>
<td>&lt; $31,100</td>
<td>&gt; $67,800</td>
<td>&gt; $42,000</td>
<td>&lt; $42,000</td>
<td>&lt; $42,000</td>
<td>&lt; $42,000</td>
<td>&gt; $42,000</td>
</tr>
<tr>
<td>Roadway Function Class</td>
<td>11, 16, 17, 18, 19</td>
<td>-</td>
<td>14, 16, 18, 19</td>
<td>14, 16, 18, 19</td>
<td>-</td>
<td>14</td>
<td>14</td>
<td>-</td>
</tr>
<tr>
<td>Supermarket</td>
<td>&lt; 4.5</td>
<td>&lt; 3.5</td>
<td>-</td>
<td>&lt; 3.5</td>
<td>&gt; 14.5</td>
<td>&lt; 14.5</td>
<td>&lt; 14.5</td>
<td>&gt; 3.5</td>
</tr>
<tr>
<td>Religious Facility</td>
<td>&lt; 17</td>
<td>&lt; 22.5</td>
<td>-</td>
<td>&lt; 15</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&gt; 7.5</td>
</tr>
<tr>
<td>Health Facility</td>
<td>&lt; 20</td>
<td>&lt; 22.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&lt; 23.5</td>
<td>&gt; 8.5</td>
</tr>
<tr>
<td>School Facility</td>
<td>-</td>
<td>-</td>
<td>&lt; 2.5</td>
<td>&gt; 2.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Law Enforcement</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&lt; 0.5</td>
</tr>
<tr>
<td>Total Employment</td>
<td>&gt; 890</td>
<td>&lt; 460</td>
<td>&gt; 55</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&gt; 460</td>
<td>&lt; 435</td>
</tr>
<tr>
<td>Total Population</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&lt; 1,200</td>
<td>&gt; 925</td>
<td>-</td>
<td>&gt; 1,400</td>
<td>-</td>
</tr>
<tr>
<td>Young Population</td>
<td>&lt; 330</td>
<td>-</td>
<td>-</td>
<td>&lt; 330</td>
<td>&gt; 30</td>
<td>&lt; 330</td>
<td>&gt; 30</td>
<td>&lt; 330</td>
</tr>
<tr>
<td>Aging Population</td>
<td>-</td>
<td>-</td>
<td>&gt; 160</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&lt; 120</td>
<td>-</td>
</tr>
<tr>
<td>Bus Frequency</td>
<td>&gt; 12.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&gt; 24.5</td>
</tr>
<tr>
<td>Total Traffic Volume</td>
<td>&lt; 10,500</td>
<td>&gt; 4,400</td>
<td>-</td>
<td>-</td>
<td>&gt; 6,450</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Speed</td>
<td>-</td>
<td>-</td>
<td>&gt; 38.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>


7.1. Application of regression tree findings

The calculated optimal regression tree for the SSI cannot be presented to the reader since it is almost impossible to comprehend the whole structure and results. Hence, Table 3 is created which shows the tabulated version of the decision tree results for the bus stops with the highest 1% (most unsafe) and lowest 1% (safest) SSI values. The full optimal tree that leads to highest SSI values (i.e., exhibits severe safety concerns) is provided in the Appendix. For this purpose, the threshold of 541.74 is adopted considering that the SSI value of 541.74 will...
correspond to one pedestrian fatality in the proximity of the bus stop with respect to the KABCO weights and decay function.

The difficulty of visually representing the full decision tree may seemingly defy the simplicity of utilizing decision trees. However, the binary structure of the decision tree can easily be implemented within a spreadsheet format without rigorous coding skills. Then, the practitioner can type in the variable values to calculate the SSI for specific bus stops. Overall, the regression tree model can provide easily implementable SSI tabulations using commonly available datasets and help practitioners estimate SSI scores of bus stops. In that sense, the results of the regression tree can be utilized in a similar fashion with safety performance functions (Persaud and Lyon and Felsburg Holt & Ullevig, 2009) while predicting crash numbers using predictors such as average annual daily traffic (AADT).

8. Conclusions

This paper shows that there is a significant spatio-statistical correlation between the bus stop locations and pedestrian-involved crashes. That is, the pedestrian-involved crashes do not happen randomly in proximity to the bus stops. On the contrary, they are spatially correlated with bus stop locations. This finding further implies that bus stop proximity has significant pedestrian safety implications. Given that bus stops are generally located at high pedestrian activity locations (Mistretta et al., 2009), a metric that can measure pedestrian safety around bus stops can improve the overall pedestrian safety. Accordingly, a bus stop safety index (SSI) is developed based on the severity (with respect to KABCO scale) and the location of the pedestrian-involved crashes. The SSI metric can be used by practitioners and researchers by ranking the bus stops as high pedestrian activity areas and choosing the ones with the high SSI scores for further scrutiny and treatment. Due to the utilization of KABCO severity weights in its formulation, SSI values can also be categorized as low, medium, and high risk, e.g. higher than 540 (KABCO equivalent of fatality) as “high risk”, between 30 (KABCO equivalent of incapacitating injury) and 540 as “medium risk”, and lower than 30 as “low risk”. These risk level categorizations can be adjusted to reflect local conditions and subjective assessment of the policy makers.

That being said, the requirement of detailed crash data to calculate SSI values makes them not readily available tools for practitioners. In order to make the metric available for practitioners, a regression tree model is estimated to tabulate SSI scores by using socio-demographic factors (i.e., income, population), traffic indicators (i.e., volume, speed), proximity of facilities (i.e., supermarket, hospital, school), and bus stop metrics (daily boarding and frequency). A regression tree is composed of binary decision nodes (e.g., speed limit lower or higher than 35 mph). Thus, the regression trees can easily be implemented in a spreadsheet, and can provide tabulated SSI values. Hence, the regression tree results can be used by practitioners to scan the urban bus stops for safety concerns in a similar fashion where a roadway network is scanned with safety performance functions. However, it should be noted that the presented decision tree tabulations may require re-estimations for localities that are substantially different from the study area of this paper.

8.1. Limitations and future research

A wide range of socio-demographic, land use characteristics, traffic, and roadway-related factors were included in the regression tree; however, there are other roadway geometrics-related influential characteristics which are not utilized in the model. These geometric variables such as presence of a crosswalk or pedestrian signs in the vicinity, parking characteristics of the roadways can be expected to have an impact on the pedestrian safety and the safety of the bus stops. Unfortunately, such detailed geometric data were not available for the bus stops used in the study. The “Roadway Function” variable was contemplated as a proxy for the geometric characteristics of the roadways where bus stops are located. Nonetheless, the regression tree model developed for SSI can be improved with the addition of aforementioned detailed roadway geometrics data. Investigating the effect of bus stop and roadway characteristics such as the availability of pedestrian crosswalks, medians, traffic lights, number of lanes is a promising future direction.

It is also worth mentioning that the preference of regression tree over spatial regression or heterogeneity models is based on the practical benefits of the adopted approach rather than its capacity to account for the statistical issues such as the spatial autocorrelation and the unobserved heterogeneity. That is, a spatial regression model depends on specific spatial context and it is needed to be re-estimated for every different location (or set of bus stops) as spatial context (i.e., adjacency matrix) also changes with the location. Heterogeneity models, on the other hand, have complex estimation processes, pose dimensionality challenges, and difficult to utilize for prediction (Manning et al., 2020). Although these complex models can address those issues associated with statistical estimations, they are incompatible with the purpose of providing a practical tool for estimating the SSI when the necessary crash data or specialized tools are not available. The regression tree is a practical tool as such, because once it is estimated, the tabulated tree values can be used easily to estimate SSI for a specific bus stop. Nevertheless, it is important to note that the non-parametric nature of the adopted regression tree approach limits the capacity to account for the unobserved heterogeneity and spatial autocorrelation issues, which may have a significant effect on the validity and robustness of the model outputs. It is worth reiterating that the proposed SSI metric is a stand-alone approach that does not need any statistical regression method to estimate or predict. Regression tree approach was proposed as an auxiliary tool for the cases when such data and tools are lacking.

Another caveat of the study is that the number of pedestrian trips or the percentage of workers commuting by walking were not considered in this study because such modes corresponds to 1% of the total trips made in Palm Beach, Florida (ACS, 2010). Therefore, the impact of individuals commuting by walking is very small compared to the other modes of transportation. Nonetheless, SSI can be extended to be implemented at a location such as New York City where pedestrian trips are a very important factor in safety.

Overall, the proposed SSI methodology is a step towards establishing standardized assessment metrics for pedestrian safety, yet there is still need for further development and validation. As a future direction, we are planning to validate and test the applicability of SSI metrics by utilizing data from other transit systems and increasing the geographical coverage of the crash and bus stop location data. Furthermore, SSI metrics for different time windows (i.e., morning peak, off-peak, and evening peak hours) can be studied.

Acknowledgments

This project was supported by US Department of Transportation grant DTRT13-G-UTC42 and administered by the Center for Accessibility and Safety for an Aging Population (ASAP) at Florida State University (FSU), Florida A&M University (FAMU), and University of North Florida (UNF). The opinions, results, and findings expressed in this manuscript are those of the authors and do not necessarily represent the views of aforementioned institutions.
Appendix A

See Figs. 10–13

Fig. 10. The first level of the regression tree.

Fig. 11. The second level of the regression tree starting from branch 1.
Fig. 12. The third level of the regression tree continuing from branch 1.
References