WILEY | Hindawi

Research Article

Identifying High-Risk Intersections for Walking and Bicycling Using Multiple Data Sources in the City of San Diego

Mahdie Hasani,[1] Arash Jahangiri ⓘ ,[1] Ipek Nese Sener,[2] Sirajum Munira,[2] Justin M. Owens,[3] Bruce Appleyard,[4] Sherry Ryan,[4] Shawn M. Turner,[5] and Sahar Ghanipoor Machiani[1]

[1]*Civil, Construction, and Environmental Engineering, San Diego State University, 5500 Campanile Dr, San Diego, CA 92182, USA*
[2]*Texas A&M Transportation Institute, 505 E. Huntland Dr., Austin, TX 78752, USA*
[3]*Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, USA*
[4]*School of Public Affairs, San Diego State University, 5500 Campanile Dr, San Diego, CA 92182, USA*
[5]*Texas A&M Transportation Institute, 3135 TAMU, College Station, TX 77843, USA*

Correspondence should be addressed to Arash Jahangiri; ajahangiri@sdsu.edu

Over the last decade, demand for active transportation modes such as walking and bicycling has increased. While it is desirable to provide high levels of safety for these eco-friendly modes of travel, unfortunately, the overall percentage of pedestrian and bicycle fatalities increased from 13% to 18% of total road-related fatalities in the last decade. In San Diego County, although the total number of pedestrian and bicyclist fatalities decreased over the same period of time, a similar trend with a more drastic change is observed; the overall percentage of pedestrian and bicycle fatalities increased from 19.5% to 31.8%. This study aims to estimate pedestrian and bicyclist exposure and identify signalized intersections with highest risk for walking and bicycling within the city of San Diego, California, USA. Multiple data sources such as automated pedestrian and bicycle counters, video cameras, and crash data were utilized. Data mining techniques, a new sampling strategy, and automated video processing methods were adopted to demonstrate a holistic approach that can be applied to identify facilities with highest need of improvement. Cluster analysis coupled with stratification was employed to select a representative sample of intersections for data collection. Automated pedestrian and bicycle counting models utilized in this study reached a high accuracy, provided certain conditions exist in video data. Results from exposure modeling showed that pedestrian and bicyclist volume was characterized by transportation network, population, traffic generators, and land use variables. There were both similarities and differences between pedestrian and bicycle models, including different spatial scales of influence by mode. Additionally, the study quantified risk incorporating injury severity levels, frequency of victims, distance crossed, and exposure into a single equation. It was found that not all intersections with the highest number of pedestrian and bicyclist victims were identified as high-risk after exposure and other factors such as crash severity were taken into account.

1. Introduction

According to the fatality analysis reporting system (FARS) encyclopedia, in 2016, 818 cyclists and 5987 pedestrians were killed in traffic accidents, making up 18.2 percent of all crash fatalities. After some reduction between 2005 and 2009, pedestrian and cyclist fatalities have seen an increasing trend since 2009. The overall percentage of pedestrian and bicycle fatalities, as a percentage of total road-related fatalities, increased from 13% to 18% in the last decade [1]. In San

Diego County, although the total number of pedestrian and cyclist fatalities decreased over the same period of time, a similar trend with a more drastic change is observed; the overall percentage of pedestrian and bicycle fatalities increased from 19.5% to 31.8%. Statistics also shows more Americans are walking and bicycling for commuting and recreation over that last decade [2]. Bicycle and pedestrian volume, known as exposure data, is an essential part of safety assessment. However, most existing bicycle and pedestrian networks are not equipped to routinely collect count data,

such as typical count data collection undertaken for vehicular networks (e.g., via loop detectors). Given this lack of bicycle and pedestrian data, local agencies are not able to accurately assess which facilities are in highest need of improvement. Technological advancements in transportation are creating new opportunities to investigate new sources of data to more accurately measure pedestrian and bicycle activity and risk exposure, thereby improving our safety modeling and planning.

Average annual daily pedestrian volume (AADP) and average annual daily bicyclist volume (AADB) are common metrics used in nonmotorized transportation studies. These metrics can be used as a measure of exposure in crash risk analyses. Due to limited resources and time, it is not possible to collect data for a whole year at many locations. Thus, a common practice is to collect short-term counts at several locations and apply an extrapolation method to convert short-term counts into yearly counts [3–9], which can be expressed in AADP or AADB forms. Exposure modeling can then be applied to estimate AADP or AADB for all other locations that were not originally selected. Direct-demand models have been the most popular approach for exposure estimation. Several studies have used the direct-demand model approach to develop demand models based on observed pedestrian or bicycle traffic volume and associated socioeconomic and land use variables in specific areas [10–13]. The estimated exposure can be used in risk quantification for normalization; risk is generally calculated by dividing the number of safety events by the estimated exposure derived from pedestrian and bicycle counts. Depending on how safety events (number of crashes, victims, etc.) and exposure (populations, traffic volume, AADP, etc.) are defined, different crash risk values can result for the same location. Some studies have focused on pedestrian and bicyclist crash risk modeling to investigate the factors that increase or decrease crash risk as well as selecting or developing a metric to quantify risks at specific locations [14–19].

While exposure—or the amount of pedestrian, bicyclist, or motorist activity occurring within a certain timeframe within a certain area—is highly correlated with pedestrian and bicycle crashes [20], there is a well-acknowledged dearth of data on walking and bicycling behaviors and activity levels, and researchers globally are working to address this gap [21]. This study contributed to the field by utilizing data from multiple sources, including automated pedestrian and bicycle counters, video cameras, crash databases, and other sources (e.g., GIS), to identify high-risk signalized intersections within the city of San Diego, California, USA. Data mining techniques, a modern sampling strategy, and automated video processing methods were adopted to demonstrate a holistic approach that can be applied to identify facilities with highest need of improvement. The study was conducted in four major steps: (1) Identifying the intersections for short-term video data collection. (2) Developing a vision-based intersection monitoring system to automatically detect, track, and count pedestrians and bicyclists. (3) Converting short-term counts to long-term counts collected at the selected intersections. (4) Conducting exposure modeling and risk quantification for walking and bicycling at signalized

intersections. The remainder of the paper is organized as follows. Background on exposure modeling, risk quantification, and data needs are presented next, followed by a data description section. Subsequently, the methodology and results sections discuss the four major steps mentioned above. Finally, conclusions and future direction are provided.

2. Background

Exposure and roadway crash risk studies have been conducted for different levels of the geographic area. These studies can be divided into two categories [22]: (1) area-wide approaches in which wide areas such as traffic analysis zones and census tracts are considered as units of analysis [23, 24], and (2) facility-specific approaches that take roadway segments or intersections as units of analysis [25–30]. The present study focuses on identifying high-risk signalized intersections in the city of San Diego, and therefore signalized intersections are the units of analysis for this study. Out of all these intersections, a sampling strategy is required to identify a subset of intersections for collecting short-term counts that are utilized to develop an exposure model. The exposure model can then be applied to estimate the volumes at other locations that were not part of the sample. Two general groups of sampling methods for site selection include probabilistic and nonprobabilistic sampling [22, 31]. Nonprobabilistic sampling techniques are mostly based on nonrandom factors such as engineering judgment [31]. In contrast, probabilistic sampling techniques involve some random selection in the process and thus result in better generalization. Other probabilistic sampling techniques include cluster analysis, stratified sampling, and multistage random which is basically the combination of clustering and stratification [22, 31].

2.1. Short-Term and Long-Term Data Collection. Most studies and agencies have used a manual approach [14, 32] that includes having people collect count data in the field or having them extract count data by watching video data collected at the selected intersections. To facilitate manual work, advanced video processing algorithms [33–35] can be applied to automatically count pedestrians and bicyclists. Different methods and technologies for collecting vulnerable road users are extensively discussed in [36]. Three steps can be identified when conducting video data analysis: object detection, object tracking, and behavior analysis. Object detection is the process of identifying different objects in images or video frames. Once the objects are detected, they need to be tracked frame by frame to monitor the spatial and temporal characterization of the objects. Finally, behavior analysis of the object trajectories could be carried out to obtain a desired outcome, such as speed of moving objects [37, 38], object counts [35, 38–40], waiting time [35, 40], lane keeping data [41], traffic violations [37, 42], and overtaking information [43].

As it is not feasible to collect count data for a whole year at the selected sites, even automatically, a common practice is to collect data for shorter periods of time and apply an extrapolation method to convert short-term count data to yearly data. Extrapolation has been used in several studies

for estimating Annual Average Daily Pedestrian (AADP) and Annual Average Daily Bicyclist (AADB) [4, 6, 25, 44, 45]. When collecting short-term counts, several factors such as counting period length, time of day, month, and year could potentially impact the accuracy of yearly count estimation [3, 4, 6, 45–47]. In addition to collecting short-term data, long-term count data at several locations are also required to perform the extrapolation. The long-term counts could provide daily, weekly, and monthly count patterns that are used for calculating adjustment factors to perform extrapolation.

Several extrapolation methods have been used in previous studies such as traditional or standard method, Day-by-month, day-of-year, and weather model [4, 6, 9, 45, 47]. The selection of the method depends on the geographic location and weather variation of the study area as well as the availability and duration of short-term and long-term counts. For example, the variation in temperature has a great impact on the number of pedestrian and bicyclists in certain areas. Therefore, some methods were proposed to incorporate the influence of weather on AADP and AADP [7, 9]. Another consideration in the extrapolation process is to apply a suitable method for matching short-term counters to appropriate long-term counters [6, 7, 9, 30, 47–50].

2.2. Exposure Modeling and Risk Quantification. Crash risk is generally estimated by dividing the number of safety events such as crashes by a total number of people who were likely to be involved in the safety events (i.e., exposure). Hence, crash risk is the probability of crash occurrence per unit of exposure. Focusing on pedestrians and bicyclists, the exposure has been defined in different ways such as pedestrian or bicyclist volume and estimated number of streets or travel lanes crossed [51]. At a theoretical level, it has been suggested that exposure may be defined as a measure of the number of potential opportunities for a crash to occur [14]. In most studies, the number of pedestrian/bicyclist crashes has been divided by a single exposure variable in order to calculate the crash risk. However, exposure can also be defined using multiple variables simultaneously. For example, the number of pedestrians and vehicles were used in [14]; average daily pedestrians, average daily traffic, and distance crossed were used in [52]; and hundred million pedestrian/bicycle miles of roadway traveled were used in a study conducted in Washington, D.C. [51].

When investigating intersection-related crash risk, it is important to consider the roles that different transportation modalities including pedestrians, bicyclists, motorcyclists, and low-speed mobility vehicles play in traffic and the differential risks that are faced by people traveling in these modalities. In a large-scale study, crash risks were investigated by travel mode, gender, and age groups [53]. In a study in Washington, D.C., it was found that although the number of crashes involved with pedestrians was more than two times the number of bicycle-involved crashes, the crash risks were similar after taking exposure (per 100M miles traveled) into account [32]. Also, in a tendency known as the "safety in numbers" phenomenon, areas with higher bicycle traffic flow have been found to be less risky for individual bicyclists in some cases [54].

3. Data Description

The unit of analysis in the present study is signalized intersections in the city of San Diego. A total of 1522 signalized intersection was identified using an ArcGIS shapefile. Short-term video data were collected by National Data and Surveying Services (NDS) at a sample of 45 intersections (selection of 45 intersections will be discussed in the site selection section below). These intersections were equipped with cameras, and the videos were recorded for 12 hours (7:00 am to 7:00 pm) in a workday (Tuesday, Wednesday, or Thursday) in May, June, or July 2018. Short-term data collection can be conducted for different lengths from a few hours to a few weeks, and generally, more data leads to smaller extrapolation errors. In this study, data collection length of 12 hours was used for two reasons. First, data collection for longer periods was not feasible due to budget constraints. Second, according to Nordback et al.'s study, it was found that extrapolation error rates do not decrease significantly from 12 to 24 hours [6]. Subsequently, pedestrian and bicycle short-term counts were automatically obtained through machine-vision modeling. In addition to the short-term counters (of selected intersections), 43 automated counters were also utilized among all automated counters in the county of San Diego. These counters are not located at intersections, but they continuously collect pedestrian and bicyclists counts since 2012. Data have been collected for several years from these counters. However, due to the equipment being vandalized and some issues with battery counters, data from 2015 were used, which is believed to contain the most reliable data. For every intersection, demographic characteristics, socioeconomic, and built environment variables were obtained by buffer analysis in ArcGIS. In addition, crash data involving pedestrians and bicyclists were obtained from the Statewide Integrated Traffic Records System (SWITRS) through the Transportation Injury Mapping System (TIMS). Crash data per victim for each intersection was extracted from 2006 to 2016. For all data sets, missing values and outliers were identified and dealt with by comparing saturation flow rate of bicyclists and pedestrians with actual observations, manually reviewing the data, and imputing data (e.g., using interpolation).

4. Methodology

4.1. Site Selection. A sampling strategy entailing cluster analysis and stratified sampling was utilized to identify a representative subset of intersections based on several variables. Cluster analysis is a classification technique that can be used to classify observations (intersections in our case) into distinct categories with similar characteristics. The algorithms used for data clustering can be categorized into several groups, such as partitional, hierarchical, density-based, grid-based, and model-based algorithms [55, 56]. In the present study, partitional algorithms were considered as they are the most widely used approaches. To select the best number of clusters, the Silhouette metric, Elbow, and gap statistics method can be employed [57, 58].

In stratified sampling, a few variables with different levels or thresholds are typically used to create strata, and each

analysis unit can then be associated with a stratum [26]. Stratified sampling is effective in that it ensures the sample contains observations with different levels for the variables used. However, as the number of variables increases the number of strata grows rapidly, and thus selecting one intersection per stratum could make the sample too large. Therefore, stratified sampling could reduce flexibility in selecting many variables in the site selection process. Many factors could potentially impact pedestrian and bicyclist volume at intersections, which are typically used in site selection. A number of variables have been used in similar studies for selecting count locations such as population density, median income, and proximity to commercial properties [4, 26, 32, 45, 59–61].

In this study, a multistage random approach was adopted to benefit from both stratification and clustering. The data in this study contains both numerical and categorical variables and thus Gower coefficient [62] was used to calculate distances (i.e., similarities) between two observations. After determination of pairwise distance among observations, a partitional clustering algorithm, Partitioning Around Medoids (PAM) [57], was employed to identify the clusters. The most popular partitional algorithm is K-means due to its efficiency and simplicity [63, 64]. However, K-means is more sensitive to outliers compared to PAM, since K-means employs centroids as the center of clusters and PAM employs the actual observations to define clusters. In the stratification step, depending on the number of clusters used, a stratified sampling method can be applied using one or more intersection characteristics to ensure that intersections with different levels for these characteristics are included in the sample. If the number of clusters turns out to be very high or very low, then the number of variables and/or levels of variables used for stratification could be adjusted to obtain the desired sample size.

4.2. Vision-Based Pedestrian and Bicyclist Monitoring. A vision-based monitoring system was used to count the number of pedestrians and bicyclists crossing the intersections from short-term video data. The system used consists of three steps: object detection, object tracking, and object counting. Object detection was performed by utilizing Faster R-CNN [33] to detect pedestrians and bicyclists in video frames. Faster R-CNN is a state-of-the-art real-time object detection method that has reached a high performance when applied to the PASCAL VOC 2007 test data. Subsequently, detection results were used to perform object tracking implemented by using Intersection-Over-Union (IOU) [65] tracker. The main idea of this tracker is to associate each detection result with the highest IOU to the last detection result in the previous frame. The tracker starts a new trajectory and ends the old trajectories if all detections are not associated with any trajectories. Finally, pedestrian and bicyclist counts were obtained from regions of interest which were defined as areas typically used by pedestrians and bicyclists to cross intersections. To obtain the counts, any trajectory that entered the regions of interest was counted as a crossing pedestrian or bicyclist.

4.3. Extrapolation. To estimate AADP and AADB from short-term counts, similar pedestrian and bicyclist volume patterns for each short-term data collection site need to be identified. These volume patterns are utilized for extrapolating long-term counts from short-term counts. Permanent counters, even at a different location from the short-term counters, are typically used to identify similar demand patterns. In this study, the matching process was performed in two steps as proposed in [30]: in the first step, the PAM clustering method was applied to classify long-term counters into different clusters based on traffic distribution indexes including AMI, WWI, and PPI for bicycle counters and AMI and WWI for pedestrian counters. In the second step, the classified long-term counters were used as the training data for developing a K-nearest neighbor (KNN) [57] model to match short-term counters to appropriate clusters.

Several variables have been calculated and utilized in the extrapolation process. These variables include population density, employment density, and land use density (commercial, residential, government, industrial, park and recreational) within a given counter buffer (0.25-mile buffer), as well as traffic distribution indices such as AMI, WWI, and PPI introduced in [7, 9]. These indices reflect bicyclist volume in morning peak hour over midday peak hour, weekend over weekday, and monthly variations, respectively.

Several extrapolation methods have been used in previous studies and a few -such as day-of-year [4], day-by-month [9], and weather model- have been shown to produce lower AADP and AADB estimation errors. The day-of-year method was not applicable in this study as it required the short-term and long-term data to be collected in the same year. However, short-term data collection was conducted in 2018 while long-term data used were collected in 2015. In addition, the weather model was not deemed to be beneficial due to San Diego's year-round mild weather. Thus, the day-by-month method was applied with minor modification as described below.

First, 12-hour counts were converted to 24-hour counts using (1). Equation (2) shows how day-by-month factor was calculated for every day d of the week and month m. Subsequently, AADP and AADB counts were estimated by applying day-by-month adjustment factors to the 24-hour counts using (3). The long-term counter data in the following equations refer to a counter that has been matched with the short-term counter of interest. However, it should be noted that if two or more permanent counters are matched to a short-term counter, the mean of adjustment factors across all matched counters was used.

$$P_{dm}^s = \sum_{h=7}^{18} P_{hdm}^s \times \frac{\sum_{h=1}^{24} P_{hdm}^l}{\sum_{h=7}^{18} P_{hdm}^l} \qquad (1)$$

$$PF_{dm} = \frac{AADP^l}{ADP_{dm}^l} \qquad (2)$$

$$AADP^s = P_{dm}^s \times PF_{dm} \qquad (3)$$

where

P_{dm}^s (*or B_{dm}^s for bicyclists*): pedestrian (bicyclists) count on day d of a week, in month m estimated for short-term counter s;

P_{hdm}^s (or B_{hdm}^s for bicyclists): pedestrian (bicyclist) count in hour h of day d of a week, in month m from short-term counter s;

P_{hdm}^l (or B_{hdm}^l for bicyclists): pedestrian (bicyclist) count in hour h of day d of a week, in month m from matched long-term counters;

PF_{dm} (or BF_{dm} for bicyclists): pedestrian (bicyclists) day-by-month factor for day d of a week, in month m;

$AADP^l$ (or $AADB^l$ for bicyclists): average annual daily pedestrian (bicyclist) count obtained from matched long-term counters;

ADP_{dm}^l (or ADB_{dm}^l for bicyclists): average daily pedestrian (bicyclist) volume on day d, in month m from matched long-term counters;

$AADP^s$ (or $AADB^s$ for bicyclists): average annual daily pedestrian (bicyclist) count estimated for short-term counter s.

4.4. Exposure Modeling and Risk Quantification. A wide variety of approaches and methods have been used in predicting nonmotorized activity using direct-demand models. Given the nature of the dependent variable, which is discrete in nature, and the variance was greater than the mean, the negative binomial model was selected as the exposure model of this study. As discussed earlier, several variables were considered in the analysis. Univariate and bivariate correlation analyses were first conducted to explore variables' distribution or pattern and to investigate the relationship between the dependent and independent variables. Several variable forms and functions were examined to get the best data fit.

After several model trials with different combinations of the key variables, the best models were evaluated based on the predictive accuracy of the models in terms of mean absolute error (MAE) and root mean squared error (RMSE). Cross-validation technique was employed for performance evaluation. Cross-validation is a resampling technique that helps identify a parameter value, ensuring a proper balance between bias and variance [66]. For cross-validation, a subset of the data, known as the training set, is used to train the model, and the remaining data points serve as a test set or validation set. While fitting a model on a training set, it is desirable to have minimum MSE, which minimizes the difference between the prediction and the actual observation. This research used a 10-fold cross-validation method to evaluate and compare the performance of the developed models. This method split the feature vector sets into ten approximately equally sized distinct partitions. While one set was used for testing, the other sets were used for training. Then, the procedure was repeated ten times, and all accuracy rates over these ten runs were averaged to improve the estimate. The performance evaluation criterion was the average accuracy. The final models were identified based on statistical, predictive, and intuitive considerations as well as insights from the literature.

Utilizing the estimated pedestrian and bicycle activity (i.e., exposure), risks associated with walking and bicycling at signalized intersections can be calculated. A general equation to quantify risk is shown in (4).

$$Risk = \frac{Number\ of\ safety\ events}{Exposure} \quad (4)$$

Several studies have used the number of crashes in the risk equation as the number of safety events. As a result, two different locations with the same number of crashes and exposure would lead to the same level of risk, while different number of victims might be involved at one location from the other. Taking number of victims into consideration (instead of number of crashes), a crash with multiple victims should be associated with a higher risk comparing to a crash with only one victim. In addition, crashes with higher levels of severity should be attributed to higher risks. Also, the number of fatalities (instead of crashes) could be used to provide an estimate of the relative lethality of intersections. Therefore, a combination of fatalities and injuries was utilized to provide a more holistic estimate of the risk. Crash severity was incorporated in risk quantification by utilizing crash costs associated with severity levels. Other factors, AADP (AADB) as the exposure and distance crossed, were also included in the risk equation as presented in (5). In addition to the crash cost, the equation numerator includes a term, N^k, to produce more weight (i.e., importance) on the locations with higher frequency of victims. Since crashes are rare events, it is important to magnify the number of occasions that led to fatalities and injuries. The tuning parameter, k, can also be used to provide the extent of the weight. For example, if zero is selected for this parameter, N^k becomes one, which means zero weight is given to the victim frequency. As k increases, it provides more weight on the victim frequency and consequently higher risks are resulted.

Crash cost has been used for different purposes, such as analyzing the effectiveness of a specific roadway enhancement and measuring the effect of seatbelt, and it has been estimated based on injury severity in several studies [67–71]. In the very beginning, Miller estimated motor-vehicle crash comprehensive costs by injury severity and body region [67]. Another study estimated crash costs of medium and heavy trucks by seven injury severity levels [68]. Miller et al. broke down pedestrian and pedalcyclist crash costs by age, injury severity, and body region in the United States [69]. Federal Highway Administration (FHWA 2005) also presented an estimation of crash cost based on maximum police-reported injury severity.

In crash cost studies, maximum abbreviated injury severity (MAIS) is defined as the maximum threat of a crash to a victim's life [72]. Crash cost generally results from a combination of different cost categories, including medical, emergency service, lost productivity, the monetized value of the pain and suffering, and lost quality of life costs. Collectively, these costs have been called comprehensive costs. Monetary or economic cost value of a crash can be obtained by subtracting lost quality of life from the comprehensive cost

Optimal number of clusters
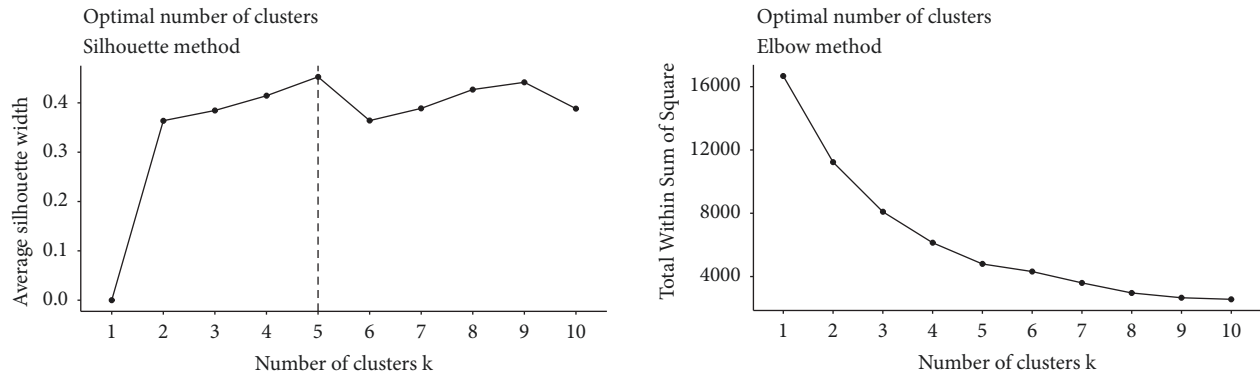Silhouette method

Optimal number of clusters
Elbow method

Figure 1: Identifying the best number of clusters: Silhouette method on the left and Elbow method on the right.

[67]. In this study, lifetime costs were used which includes medical, work loss, and quality of life costs [69].

$$Quantified\ Risk = \frac{C \times N^k}{AADP\,(AADB) \times D} \tag{5}$$

$$C = \sum_s N_s \times C_s$$

where

C: total crash cost weighted by severity;

N_s: number of pedestrians or bicycle victims with injury severity level s;

C_s: cost per victim with injury severity level s;

s: severity level = {fatal, severe injury, other visible injury, and complaint of pain};

N: total number of victims;

k: exponent of N, a tuning parameter to magnify the frequency of victims;

AADP(AADB): average annual daily pedestrian (bicyclist) count;

D: distance a pedestrian or bicyclist crossed.

5. Results and Discussion

5.1. Site Selection. Any variable that is expected to affect pedestrian and bicycle activity at intersections could potentially be included in the site selection process. A total of 18 variables were examined and after several trials with different subsets of variables and excluding variables with high correlation with other variables, a subset of 12 variables were selected to perform site selection as follows: population density, land use (parks and recreational; residential), presence of college, presence of school, transit stops density, mean traffic volume, pedestrian victims, bicyclist victims, proximity to Balboa Park, proximity to beaches, sidewalk density, and bikeway density.

Using all selected variables, signalized intersections were grouped into clusters by applying PAM clustering method. The Silhouette metric and Elbow method were applied to

identify the best number of clusters. The highest Silhouette value -that shows the highest clustering performance- was obtained when using 5 clusters, as shown in Figure 1. Based on the Elbow plot in this figure, as the number of clusters increases, the total sum of squares decreases. However, no clear elbow point is visible, and thus the best number of clusters was selected to be five based on the Silhouette method only. The geographic distribution of these 5 clusters is shown in Figure 2.

Within each cluster, stratified sampling was performed using two variables, namely the number of pedestrian victims and the number of bicyclist victims. The purpose was to ensure that the sample includes intersections with a high, moderate, and low number of victims. The number of pedestrian victims, ranging from 0 to 13, was divided into three levels (low: 0, 1, 2; moderate: 3, 4, 5; high: >=6). Similarly, the number of bicyclist victims was divided into three levels (low: 0, 1; moderate: 2, 3, 4; high: >=5). Consequently, nine strata for each cluster (3*3=9) resulted. Subsequently, a sample of 45 intersections was identified as shown in Figure 2 by selecting one intersection per stratum (5*9=45). It should be pointed out that most adjacent intersections are assigned to the same cluster unless their characteristics are very different from their adjacent intersections. In addition, the selected sample was carefully reviewed to ensure adjacent intersections are not in the sample. This was considered to avoid variable correlation of intersections sharing the same area of influence.

5.2. Vision-Based Pedestrian and Bicyclist Monitoring. Utilizing video data, machine-vision models were trained to detect and track pedestrians and bicyclists. Several pedestrians and bicyclists were labeled to perform the training task. The vision-based monitoring system was tested in several scenarios and system performance was assessed using real-world video data from stationary cameras at several signalized intersections. Figure 3 shows an example in which two pedestrians and a bicyclist were successfully detected, tracked, and counted as they crossed the intersection within the region of interest (transparent blue region). The average pedestrian and bicycle counting accuracies were 85% and 81%, respectively. Several factors impacted the model performance, including the number of pedestrians and

Cluster

- 1 (red)
- 2 (blue)
- 3 (purple)
- 4 (orange)
- 5 (green)

Intersection Selection
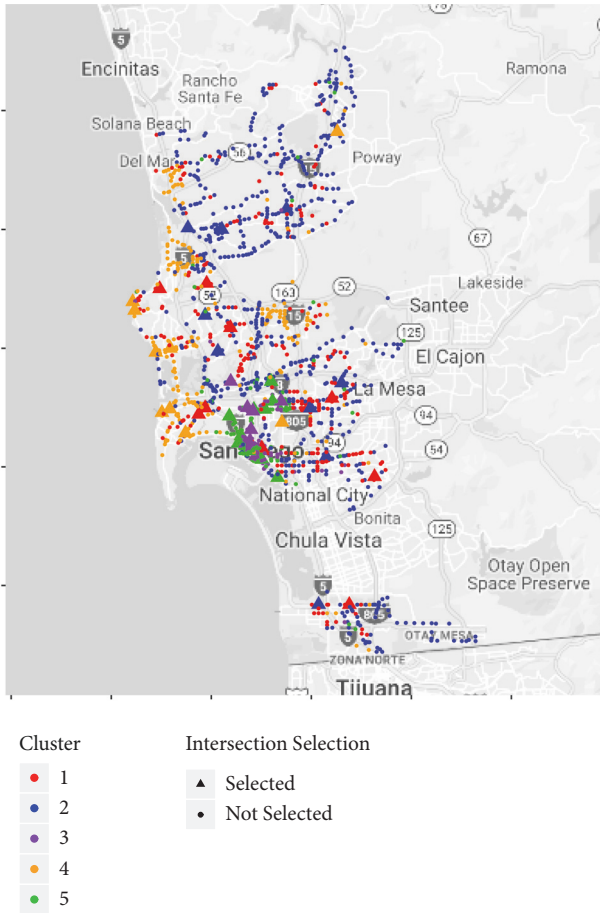
- ▲ Selected
- • Not Selected

FIGURE 2: Site selection results by cluster analysis and stratification: triangular shapes are the 45 selected intersections.

bicyclists labeled, intersection shape and size, lighting condition, occluded objects, and video quality. The best counting accuracy of 95% was achieved for both pedestrians and bicyclists in scenarios in which many objects were labeled, good lighting condition was present, video quality was decent, and pedestrians/bicyclists were not crossing in groups.

As expected, more labeled pedestrians and bicyclists led to better model performance as the models were provided with more information in terms of positioning, angles, and lighting conditions. For example, the detection accuracy increased by 68% when the number of labeled pedestrians increased from 15 to 32. Model transferability was examined by using data from one intersection to train models and testing these models on other intersections. The benefit was that the manual work of labeling was reduced, but the models performed poorly since different intersections have different shapes and sizes. Focusing on one intersection for both training and testing, the way people cross the intersection (as individual vs. in groups), lighting conditions due to time day significantly affected model generalizability. For instance, the models had difficulty detecting and tracking pedestrians and bicyclists crossing the intersection in groups since some of them were occluded by others in multiple video frames. In

addition, the quality of the video and object distances to the camera impacted the results. Cameras used in this study were set at a corner of each intersection. Detecting objects crossing the two farther intersection approaches from the camera was challenging, especially in large intersections where pedestrians and bicyclists were too small to distinguish.

5.3. Extrapolation. Pedestrian and bicycle volume patterns at permanent counters were identified by PAM clustering method. Figures 4 and 5 illustrate clusters of pedestrian counters classified into three clusters (Recreational, Mixed, and Utilitarian) based on AMI and WWI. As shown in Figures 6 and 7, bicycle counters were grouped into four clusters (Utilitarian, Recreational, Mixed Recreational, and Mixed Utilitarian) based on AMI, WWI, and PPI calculated for every counter. Pattern classification into three and four clusters have been used in past studies [46, 49, 73] and the only reason three clusters were used (instead of four) for pedestrian counters was the limited number of counters which would have led to having only one counter in a cluster, which is not recommended [6, 50].

Utilitarian counters have two distinct peak hours in mornings and evenings on weekdays. They also have a relatively uniform distribution throughout the week as shown by daily pattern. Recreational counters have higher weekend peaks than weekday peaks as expected. The daily patterns also show the highest volume on Saturdays and Sundays compared to the other days of the week. Mixed, mixed recreational, or mixed utilitarian counters represent different combination variations of utilitarian and recreational demand. These classifications for bicycle and pedestrian counters were consistent with the literature as discussed in [49]. Finally, each short-term counter was matched to one of the specified clusters, and the mean of the adjustment factors across all counters within that cluster was used to extrapolate the short-term counts (12-hour counts) to yearly counts (AADP or AADB).

5.4. Exposure Modeling and Risk Quantification. After the estimation of AADP and AADB for the sample intersections, exposure modeling was applied to calculate AADP and AADB for the remainder of intersections as discussed below. Tables 1 and 2 show the results of the negative binomial regression models of pedestrian and bicycle annual average daily volume, respectively. The tables present the explanatory variables and their estimates. The variables such as transit stop density for which buffer analysis was conducted were tested in exposure models with three different areas of influence (0.1 miles, 0.25 miles, and 0.5 miles). The number following each variable represents the buffer area of influence.

As shown in Table 1, the pedestrian model has seven variables, and all of them were statistically significant at the 90 percent confidence level. As shown in Table 2, the bicycle model also has seven variables, and all variables except one (transit stop density at 0.1 miles) were statistically significant at the 95 percent confidence level. The R-squared values for the pedestrian and bicycle models were 0.7 and 0.67, respectively, which assert that the models have decent goodness of fit. The MAE and RMSE of the models also

FIGURE 3: Pedestrians (right) and a bicyclist (left) detected while crossing an intersection within the region of interest.

TABLE 1: Negative binomial regression model for pedestrians.

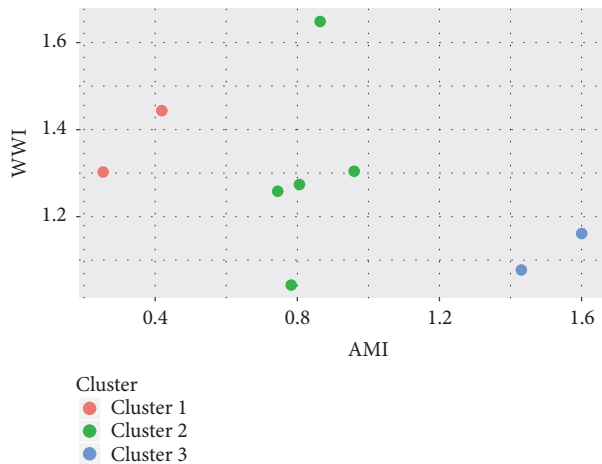| Variable	Estimates	t-stat	Sig.
(Intercept)	4.865	23.24	0.000
Transit stop density (0.5 miles)	8.782	3.81	0.000
Percentage of regular transit rider, pedestrian, or bicyclist population (0.25 miles)	3.717	1.95	0.051
Employment density (0.25 miles)	0.051	2.41	0.016
Maximum speed limit within the intersection less than 40 mph	1.135	5.37	0.000
Percentage of vacant housing units (0.5 miles)	−3.517	−2.99	0.003
Total commercial or mixed-use land area (0.1 mile)	0.190	5.01	0.000
If the area contains a higher crime count than the average crime counts among the buffers (0.25 miles)	−0.292	−1.65	0.098
N		45	
R-squared		0.70	
RMSE		1633.24	
MAE		1147.41	

FIGURE 4: Clusters identified for long-term pedestrian counters (each circle in the plot represents a long-term pedestrian counter).

indicated that the models provide a good estimate of the annual average pedestrian and bicycle volume at locations without counts.

The model results revealed important insights. The pedestrian volume was characterized by transportation network (transit stop density and speed limit), population (employment density, and regular transit rider, pedestrian, or bicyclist population), and land use (vacant housing units, commercial or mixed-use land area, and high-crime area). The best model was obtained with variables of different spatial scales. This finding is consistent with previous studies [26, 74, 75] that suggested it is unlikely to have all the variables significant at the same buffer scale. The direct-demand model for bicycle traffic included variables that represent characteristics of the transportation network (density of bicycle facility, maximum posted speed within an intersection, and transit stops density), population (total regular bicyclist population), traffic generator (presence of a school and proximity to the beach), and land use (total commercial or mixed-use area).

Interestingly, some variables, such as commercial or mixed-use land area and transit stop density, influence both pedestrian and bicycle traffic for the study area, but the spatial scale of influence varies. The commercial or mixed-use land area influences pedestrian and bicycle volume within 0.1 miles and 0.25 miles, respectively. This indicates that the commercial and mixed-use land area attracts bicyclist traffic for a larger catchment area than pedestrian traffic does. Previous studies have also indicated that commercial areas attract pedestrian [12, 75] and bicycle [11, 12, 28] activity. However, the study by Tabeshian and Kattan (2014), conducted in Canada, found a significant impact of commercial areas on pedestrian and bicycle traffic within 0.25 miles and 0.1 miles, respectively, which is contrary to this study's findings [12]. Two separate studies in Alameda
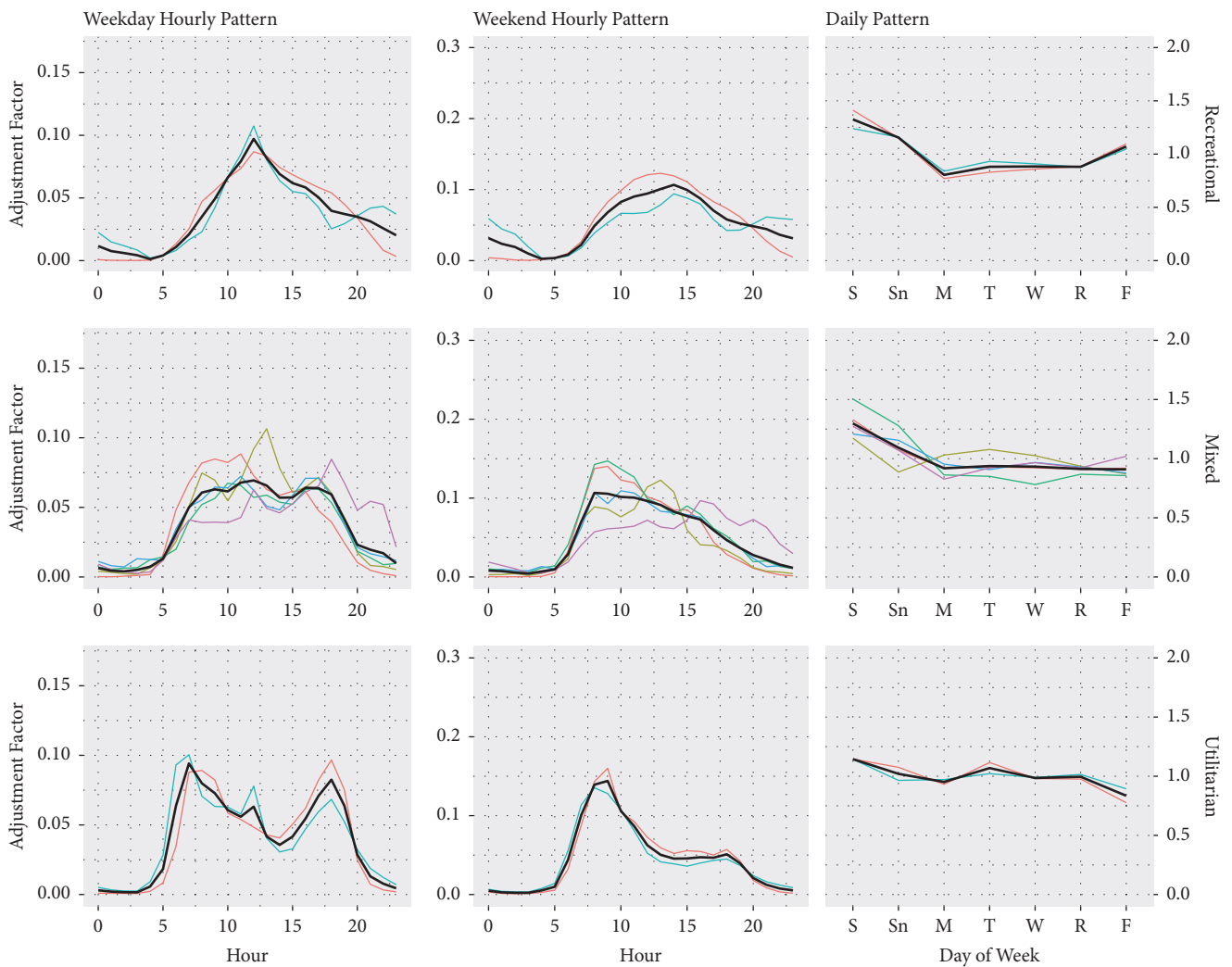
FIGURE 5: Clustering result of long-term pedestrian counters classified into three groups (Recreational, Mixed, and Utilitarian).

TABLE 2: Negative binomial regression model for bicyclists.

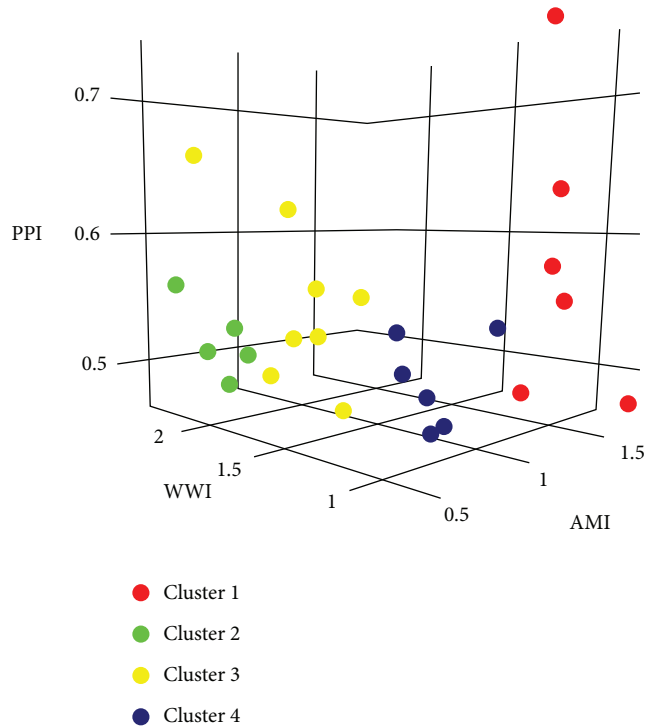| Variable	Estimates	t-stat	Sig.
(Intercept)	4.265	12.56	0.000
Regular bicyclist population (0.25 miles)	0.015	3.90	0.000
Transit stop density (0.1 miles)	0.852	1.55	0.121
Maximum speed limit within the intersection less than 40 mph	0.457	2.90	0.004
Distance between the intersection and beachfront access point less than or equal to 10 miles	0.379	2.33	0.020
Presence of a school (0.5 miles)	−0.483	−2.40	0.016
Bike facility density (0.5 miles)	1.371	2.00	0.045
Total commercial or mixed-use land area (0.25 miles)	0.021	2.18	0.030
N		45	
R-squared		0.67	
RMSE		105.93	
MAE		87.68	

FIGURE 6: Clusters identified for long-term bicycle counters.

County, California, have also found a significant influence on commercial areas on pedestrians within 0.25 miles [26] and bicyclists within 0.1 miles [28]. The comparison suggests that not only the explanatory variables but also their influence areas for nonmotorized traffic activity vary with location and community.

Similarly, transit stop density influences pedestrians within 0.5 miles, which is larger than the bicycle buffer of 0.1 miles. The results indicate that pedestrians are likely to travel more to ride a transit facility than bicyclists are. Previous studies have also observed a significant association of transit facilities with pedestrians within 0.5 miles [76] and bicyclists within 0.5 miles [11]. Given that mass transit facilities are bicycle-friendly in San Diego [77], transit riders probably make up a large proportion of pedestrians and bicyclists in the city.

Pedestrian and bicycle volumes decrease when the maximum intersection speed limit exceeds 40 mph. The finding confirms that pedestrians and bicyclists are more likely to avoid high-speed intersections and find an alternative route. Fagnant, D. J., and K. Kockelman [10] also observed a similar relationship for bicycle traffic in the Seattle, Washington, area. The finding is not surprising given the rise of traffic fatalities. A report [78] indicated that around 1,000 pedestrians and bicyclists are hit and seriously injured annually in San Diego, and in 2012, pedestrian collisions increased 20 percent compared to previous years. In 2017, there were 12 deaths and 71 serious injuries involving pedestrians and bicyclist [79]. The high crash risk could discourage pedestrians and bicyclists from using high-speed intersections.

The pedestrian model had a strong positive association between the pedestrian volume and the percentage of regular transit rider, pedestrian, or bicyclist population within 0.25 miles. As expected, the population inclined to use active modes and public transportation was more likely to contribute to the walking volume within their neighborhood. The pedestrian volume also increased with increasing employment density within 0.25 miles. Previous studies also observed similar influence in San Francisco [13] and San Diego, California [59]. The results suggest that with more people working in a neighborhood, intersections are more likely to observe higher pedestrian volume. Similarly, the negative association between pedestrian volume and total vacant housing units indicates that pedestrians are less likely to generate from neighborhoods with many vacant properties. The negative influence of crime on pedestrian volume, but not on bicyclists was also observed, which shows that people are more likely to avoid high-crime locations and conforms with previous research [80, 81].

As expected, the bicycle model indicated higher bicycle volume in areas with a larger population of regular bicyclists. The model also indicated that the intersections near beach access points (less than 10 miles) were more likely to observe high bicycle traffic. The density of bike facilities also had a positive influence on daily bicycle volume. The finding can be attributed to the recent surge of dockless bicycles in the city [82] as well as the 16 miles of separated bike paths around San Diego Bay, completed in February 2018 [83]. The Bay Shore bikeway was built with a vision to provide a scenic and convenient way for bicyclists to travel in the San Diego area. The dockless bike sharing facilities were first launched in February 2018 and added the convenience of using bicycles. The combined influence may contribute to a higher bicycle volume in locations near beach areas and with better bicycle facilities. Surprisingly, the model revealed a negative association between the presence of a school and bicyclist volume, which contradicts previous studies conducted in Canada [11, 84]. However, a study conducted in the United States suggested that the number of students (ages 5 to 18) who walk or bike to school decreased sharply in recent years due to increased traffic collisions, lack of sidewalks, and urban sprawl [85]. Perhaps the increasing collision rate in the city discouraged children from bicycling to schools, and adult bicyclists tend to avoid locations near schools.

After estimating pedestrian and bicycle volumes (i.e., AADP and AADB) as the exposure measure, the risk was quantified by applying the proposed quantified risk equation presented in (5). In this equation, number of victims and crash severity levels were obtained from the SWITRS data, distance pedestrians or bicyclists crossed at the intersection was calculated by multiplying the average number of lanes (across all approaches) by the lane width (12 ft was assumed), cost per victim was obtained based on the victim's age and injury severity as estimated by Miller et al. [69]. After experimenting several values for the tuning parameter k, a value of three was chosen for the final model. Models with smaller values of k, identified some intersections as high-risk (i.e., top 50) with only one or two victims in the past ten years. While other factors such as a small AADP (AADB) and/or a
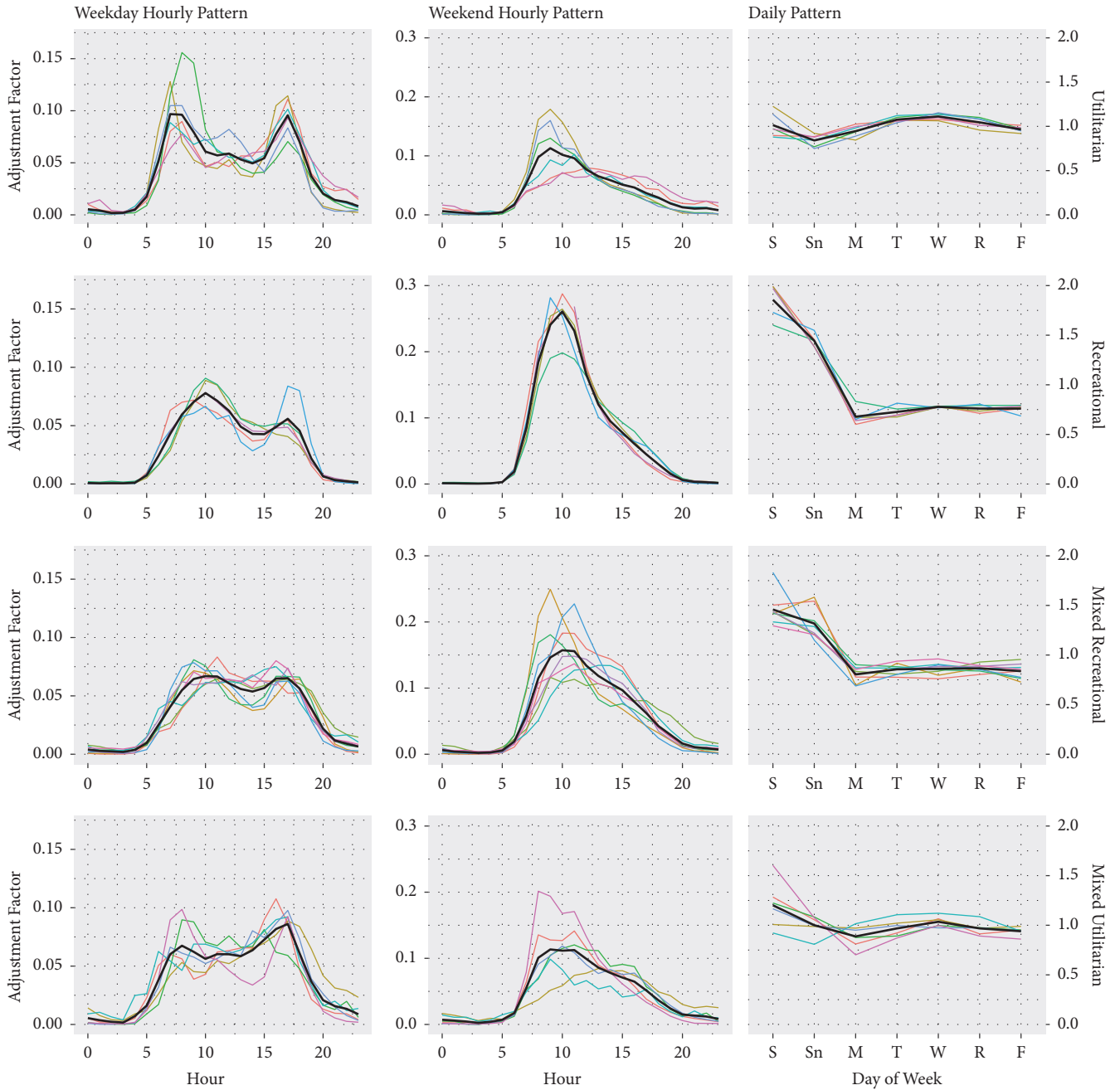
FIGURE 7: Clustering result of long-term bicycle counters classified into four groups (Utilitarian, Recreational, Mixed Recreational, and Mixed Utilitarian).

high crash cost could lead to high-risk values, it may not be practical to recognize an intersection with only one victim in the past ten years as a high-risk intersection even if the AADP is small. High values of k led to extreme values of risk especially when the number of victims was high, and thus the value of three was selected as it provided a reasonable outcome. The risk for all signalized intersections was calculated to identify high-risk intersections for walking and bicycling as mapped in Figure 8. Previously, 15 intersections, known as "fatal 15", were identified as the deadliest intersections for pedestrians in the city of San Diego [86]. As expected,

it was found that these intersections had more number of victims than other intersections. However, when exposure and other factors were taken into account using the quantified risk equation, not all intersections with the highest number of pedestrian and bicyclist victims were identified as high-risk. For example, out of 39 intersections with the highest number of pedestrian victims (Victims >= 8), only 22 were made it to the top 39 high-risk intersections based on the quantified risk. Similarly, out of 36 intersections with the highest number of bicyclist victims (Victims >= 5), 26 remained in the top 36 high-risk intersections.
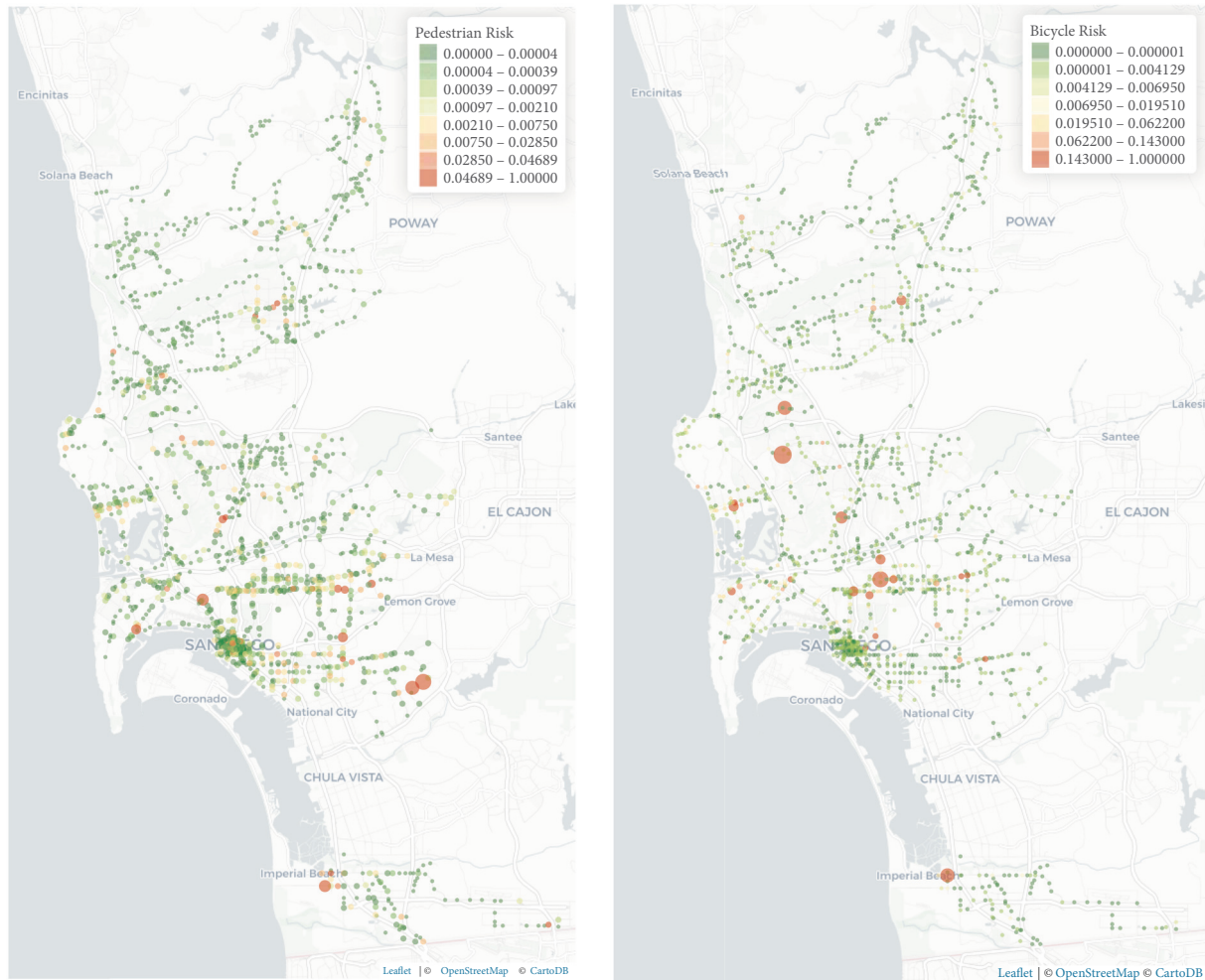
FIGURE 8: High-risk intersections for walking (left side) and bicycling (right side).

6. Conclusion

While it is accepted that pedestrian and bicyclist volume is positively correlated with the number of pedestrian and bicyclist crashes, there is a renowned lack of pedestrian and bicycle data that can negatively impact accurate risk quantification and safety evaluations. This study leveraged a combination of available data sources including automated pedestrian and bicycle counters, video cameras, crash databases, and other sources (e.g., GIS), to identify high-risk intersections for walking and bicycling. Cluster analysis and stratification were applied to identify a representative sample of locations to collect short-term video data that were used to develop a vision-based monitoring system for automatic detection, tracking, and counting of pedestrians and bicyclists. When sufficient number of pedestrians and bicyclists were annotated, pedestrians and bicyclists were not too far from the camera, they did not cross the intersection in groups, and good lighting condition was present, a high counting accuracy of 95% was obtained. Utilizing permanent counters, an extrapolation method along with a novel matching method was employed to estimate yearly counts

that were used for estimating exposure by direct-demand models. Exposure analysis identified transportation network, population, traffic generator, and land use variables as statistically significant in estimating pedestrian and bicyclist volume. Accounting exposure as a normalization factor and other factors such as frequency of victims and crash severity in quantifying risk had a significant impact on the selection of high-risk intersections; not all intersections with the highest number of pedestrian and bicyclist victims were identified as high-risk. The variables were found to be influential at multiple buffer area and showed differences across pedestrian and bicycle activity. The results underscored the importance of location and community in characterizing nonmotorized demand and targeted improvements to encourage nonmotorized activities.

The modeling framework and data sources used in this study are beneficial to conduct future analyses for other facility types such as roadway segments and also at more aggregate levels such as traffic analysis zones. The approach is also beneficial to public agencies, which can help identify high-risk facilities and prioritize them for countermeasure implementation. Since crashes are rare events, the identification

of high-risk facilities would take a long time, and thus a potential future direction is to proactively assess safety by discovering near-crash situations in video analysis. This enables researchers and practitioners to quantify risk and evaluate safety in a much shorter period of time. Another future research topic is to investigate advanced spatial modeling methodologies with direct-demand models to better understand the impact of intersections sharing the same area of influence (e.g., adjacent intersections).

Data Availability

Except the video data that may include personally identifiable information, other data may be made available. However, at this point, since the project has not been finalized yet, it is not possible to make them available within any supplementary files or any other forms. The data may be released after finalizing the project and obtaining any required permissions from the project sponsor if any.

Disclosure

The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] Fatality Analysis Reporting System (FARS), https://www-fars.nhtsa.dot.gov/Main/index.aspx, 2018.

[2] Beyond Traffic 2045, US Department of Transportation, 2015.

[3] D. Johnstone, K. Nordback, and S. Kothuri, "Annual average nonmotorized traffic estimates from manual counts: quantifying error," *Transportation Research Record: Journal of the Transportation Research Board*, 2017.

[4] S. Hankey, G. Lindsey, and J. Marshall, "Day-of-year scaling factors and design considerations for nonmotorized traffic monitoring programs," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2468, no. 1, pp. 64–73, 2014.

[5] A. R. Budowski, *Developing Expansion Factors to Estimate Cyclist Seasonal Average Daily Traffic in Winnipeg, MB*, 2015.

[6] K. Nordback, W. Marshall, B. Janson, and E. Stolz, "Estimating annual average daily bicyclists," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2339, pp. 90–97, 2013.

[7] T. Nosal, *Improving the Accuracy of Bicycle AADT Estimation: Temporal Patterns, Weather and Bicycle AADT Estimation Methods*, McGill University Libraries, 2014.

[8] T. Nosal, L. F. Miranda-Moreno, and Z. Krstulic, "Incorporating weather: comparative analysis of annual average daily bicyclist traffic estimation methods," *Transportation Research Record*, vol. 2468, no. 1, pp. 100–110, 2014.

[9] D. Beitel and L. F. Miranda-Moreno, "Methods for improving and automating the estimation of average annual daily bicyclists," in *Proceedings of the 95th Annual MeetingTransportation Research Board*, 2016.

[10] D. J. Fagnant and K. Kockelman, "A direct-demand model for bicycle counts: the impacts of level of service and other factors," *Environment and Planning B: Planning and Design*, vol. 43, no. 1, pp. 93–107, 2016.

[11] J. Strauss, L. F. Miranda-Moreno, and P. Morency, "Cyclist activity and injury risk analysis at signalized intersections: a bayesian modelling approach," *Accident Analysis & Prevention*, vol. 59, pp. 9–17, 2013.

[12] M. Tabeshian and L. Kattan, "Modeling nonmotorized travel demand at intersections in calgary, Canada: use of traffic counts and geographic information system data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2430, no. 1, pp. 38–46, 2014.

[13] R. Schneider, T. Henry, M. Mitman, L. Stonehill, and J. Koehler, "Development and application of a pedestrian volume model in San Francisco, California," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2299, pp. 65–78, 2012.

[14] M. H. Cameron, "A method of measuring exposure to pedestrian accident risk," *Accident Analysis Prevention*, vol. 14, no. 5, pp. 397–405, 1982.

[15] M.-R. Lin and J. F. Kraus, "A review of risk factors and patterns of motorcycle injuries," *Accident Analysis & Prevention*, vol. 41, no. 4, pp. 710–722, 2009.

[16] Y. Peng, Y. Chen, J. Yang, D. Otte, and R. Willinger, "A study of pedestrian and bicyclist exposure to head injury in passenger car collisions based on accident data and simulations," *Safety Science*, vol. 50, no. 9, pp. 1749–1759, 2012.

[17] K. Xie, K. Ozbay, A. Kurkcu, and H. Yang, "Analysis of traffic crashes involving pedestrians using big data: investigation of contributing factors and identification of hotspots," *Risk Analysis*, vol. 37, no. 8, pp. 1459–1476, 2017.

[18] S. Boufous, L. De Rome, T. Senserrick, and R. Ivers, "Risk factors for severe injury in cyclists involved in traffic crashes in Victoria, Australia," *Accident Analysis & Prevention*, vol. 49, pp. 404–409, 2012.

[19] D. K. Tinsworth, S. P. Cassidy, and C. Polen, "Bicycle-related injuries: injury, hazard, and risk patterns," *International Journal for Consumer and Product Safety*, vol. 1, no. 4, pp. 207–220, 1994.

[20] Federal Highway Administration, "PBCAT - pedestrian and bicycle crash analysis tool version 2.0. publication FHWA-HRT-06-090," U.S. Department of Transportation.

[21] S. Ryan, D. E. Sidelinger, S. Saitowitz, D. Browner, S. Vance, and L. McDermid, "Designing and implementing a regional active transportation monitoring program through a county-MPO-university collaboration," *American Journal of Health Promotion*, vol. 28, no. 3, pp. S104–S111, 2014.

[22] R. Greene-Roesel, M. C. Diogenes, and D. R. Ragland, *Estimating Pedestrian Accident Exposure. Publication UCB-ITS-PRR-2010-3*, Calofornia Path Program, 2010.

[23] W. Schwartz and C. Porter, *Bicycle and Pedestrian Data: Sources, Needs, and Gaps*, 2000.

[24] T. Schweizer, *Methods for Counting Pedestrians*, 2005.

[25] R. J. Schneider, L. S. Arnold, and D. R. Ragland, "Methodology for counting pedestrians at intersections," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2140, no. 1, pp. 1–12, 2009.

[26] R. J. Schneider, L. S. Arnold, and D. R. Ragland, "Pilot model for estimating pedestrian intersection crossing volumes," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2140, no. 1, pp. 13–26, 2009.

[27] M. Abdel-Aty and X. Wang, "Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1953, pp. 98–111, 2006.

[28] J. B. Griswold, A. Medury, and R. J. Schneider, "Pilot models for estimating bicycle intersection volumes," *Transportation Research Record*, vol. 2247, no. 1, pp. 1–7, 2011.

[29] I. N. Sener, K. Lee, J. G. Hudson, M. Martin, and B. Dai, *The Challenge of Safe and Active Transportation: Examination of Pedestrian and Bicycle Crashes in the Austin District*, Texas A & M Transportation Institute, 2018.

[30] M. Hasani, A. Jahangiri, and S. G. Machiani, "Developing models for matching of short-term and long-term data collection sites to improve the estimation of average annual daily bicyclists," in *Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2931–2936, IEEE, Maui, HI, USA, 2018.

[31] Y. P. Aggarwal, *Better Sampling: Concepts, Techniques, and Evaluation*, Stosius Inc/Advent Books Division, 1988.

[32] J. A. Molino, J. F. Kennedy, P. L. Johnson, P. A. Beuse, A. K. Emo, and A. Do, "Pedestrian and bicyclist exposure to risk: methodology for estimation in an urban environment," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2140, no. 1, pp. 145–156, 2009.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[34] H. Hediyeh, T. Sayed, M. H. Zaki, and K. Ismail, "Automated analysis of pedestrian crossing speed behavior at scramble-phase signalized intersections using computer vision techniques," *International Journal of Sustainable Transportation*, vol. 8, no. 5, pp. 382–397, 2014.

[35] M. S. Shirazi and B. T. Morris, "Vision-based turning movement monitoring:count, speed & waiting time estimation," *IEEE Intelligent Transportation Systems Magazine*, vol. 8, no. 1, pp. 23–34, 2016.

[36] P. Ryus, E. Ferguson, K. M. Laustsen et al., *Methods and Technologies for Pedestrian and Bicycle Volume Data Collection*, CiteSeerX, 2014.

[37] Y. Guo, T. Sayed, and M. H. Zaki, "Automated analysis of pedestrian walking behaviour at a signalised intersection in China," *IET Intelligent Transport Systems*, vol. 11, no. 1, pp. 28–36, 2017.

[38] G. Ponte, Z. L. Szpak, J. E. Woolley, and D. J. Searson, "Using specialised cyclist detection software to count cyclists and determine cyclist travel speed from video," in *Proceedings of the Australasian Road Safety Research Policing Education Conference*, Melbourne, Victoria, Australia, 2014.

[39] H. Tang, "Development of a multiple-camera tracking system for accurate traffic performance measurements at intersections," http://www.its.umn.edu/Publications/ResearchReports/reportdetail.html?id=2254.

[40] M. S. Shirazi and B. Morris, "Vision-based pedestrian monitoring at intersections including behavior & crossing count," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2016.

[41] D. Kasper, G. Weidl, T. Dang et al., "Object-oriented bayesian networks for detection of lane change maneuvers," *IEEE Intelligent Transportation Systems Magazine*, vol. 4, no. 3, pp. 19–31, 2012.

[42] M. Zaki, T. Sayed, K. Ismail, and F. Alrukaibi, "Use of computer vision to identify pedestrians' nonconforming behavior at urban intersections," *Transportation Research Record*, vol. 2279, pp. 54–64, 2012.

[43] F. Garcia, P. Cerri, A. Broggi, A. De La Escalera, and J. M. Armingol, "Data fusion for overtaking vehicle detection based on radar and optical flow," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, IEEE, 2012.

[44] J. Roll, *BicycLe Traffic Count Factoring: An Examination of National, State and Locally Derived Daily Extrapolation Factors*, Dissertations and Theses, 2013.

[45] S. Hankey, G. Lindsey, X. Wang et al., "Estimating use of nonmotorized infrastructure: models of bicycle and pedestrian traffic in minneapolis, MN," *Landscape and Urban Planning*, vol. 107, no. 3, pp. 307–316, 2012.

[46] G. Lindsey, S. X. Hankey, Wang., and J. Chen, *The Minnesota Bicycle and Pedestrian Counting Initiative: Methodologies for Non-Motorized Traffic Monitoring*, 2013.

[47] D. Beitel, S. McNee, and L. F. Miranda-Moreno, "Quality measure of short-duration bicycle counts," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2644, no. 1, pp. 64–71, 2017.

[48] S. Gadda, A. Magoon, and K. M. Kockelman, *Estimates of Aadt: Quantifying the Uncertainty*, Transportation Research Board, 2007.

[49] L. Miranda-Moreno, T. Nosal, R. Schneider, and F. Proulx, "Classification of bicycle traffic patterns in five North American cities," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2339, pp. 68–79, 2013.

[50] D. Johnstone, K. Nordback, and M. Lowry, *Collecting Network-Wide Bicycle and Pedestrian Data: A Guidebook for When and Where to Count*, Washington State Department of Transportation, 2017.

[51] S. Turner, I. N. Sener, M. E. Martin et al., "Synthesis of methods for estimating pedestrian and bicyclist exposure to risk at areawide levels and on specific transportation facilities," Tech. Rep. FHWA-SA-17-041, The Federal Highway Administration (FHWA), 2017.

[52] E. Radwan, H. Abou-Senna, A. Mohamed et al., *Assessment of Sidewalk/Bicycle-Lane Gaps with Safety and Developing Statewide Pedestrian Crash Rates*, 2016.

[53] L. F. Beck, A. M. Dellinger, and M. E. O'Neil, "Motor vehicle crash injury rates by mode of travel, United States: using exposure-based methods to quantify differences," *American Journal of Epidemiology*, vol. 166, no. 2, pp. 212–218, 2007.

[54] S. Blaizot, F. Papon, M. M. Haddak, and E. Amoros, "Injury incidence rates of cyclists compared to pedestrians, car occupants and powered two-wheeler riders, using a medical registry and mobility data, Rhône County, France," *Accident Analysis & Prevention*, vol. 58, pp. 35–45, 2013.

[55] B. Zerhari, A. Ait-Lahcen, and S. Mouline, *Big Data Clustering: Algorithms and Challenges*, Tetuan, Morocco, 2015.

[56] K. Li, H. Jiang, and A. Y. Zomaya, *Big Data Management and Processing*, Chapman and Hall/CRC, 2017.

[57] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York, NY, USA, 2009.

[58] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[59] M. G. Jones, S. Ryan, J. Donlon, L. Ledbetter, D. R. Ragland, and L. S. Arnold, "Seamless travel: measuring bicycle and pedestrian activity in san diego county and its relationship to land use, transportation, safety, and facility type," PATH Research Report, 2010.

[60] C. V. Zegeer, J. R. Stewart, H. H. Huang, P. A. Lagerwey, J. Feaganes, and B. J. Campbell, *Safety Effects of Marked versus Unmarked Crosswalks at Uncontrolled Locations: Final Report and Recommended Guidelines*, 2005.

[61] Estimating Pedestrian Accident Exposure, *Safe Transportation Education and Research Center (SafeTREC)*, California Path Program Institute Of Transportation Studies University Of California, Berkeley, Calif, USA.

[62] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, pp. 857–872, 1971.

[63] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, University of California Press, 1967.

[64] L. Kaufman and P. Rousseeuw, "Clustering by means of medoids," in *Statistical Data Analysis Based on the L_1–Norm and Related Methods*, North-Holland, Amsterdam.

[65] E. Bochinski, V. Eiselein, and T. Sikora, *High-Speed Tracking-by-Detection Without Using Image Information [Challange Winner IWOT4S]*, 2017.

[66] M. J. Chan-Lau, *Lasso Regressions and Forecasting Models in Applied Stress Testing*, International Monetary Fund, 2017.

[67] T. R. Miller, "Costs and Functional Consequences of U.S. Roadway Crashes," *Accident Analysis & Prevention*, vol. 25, no. 5, pp. 593–607, 1993.

[68] E. Zaloshnja and T. Miller, *Unit Costs of Medium and Heavy Truck Crashes*, 2007.

[69] T. R. Miller, E. Zaloshnja, B. A. Lawrence, J. Crandall, J. Ivarsson, and A. E. Finkelstein, "Pedestrian and pedalcyclist injury costs in the united states by age and injury severity," in *Proceedings of the 48th Annual Proceedings - Association for the Advancement of Automotive Medicine*, vol. 48, pp. 265–284, 2004.

[70] *Maximum Police-Reported Injury Severity Within Selected Crash Geometries, Chapter 4*, FHWA-HRT-05-051, 2005.

[71] E. Zaloshnja, T. Miller, F. Council, and B. Persaud, "Comprehensive and human capital crash costs by maximum police-reported injury severity within selected crash types," in *Proceedings of the 48th Annual Proceedings - Association for the Advancement of Automotive Medicine*, vol. 48, pp. 251–263, 2004.

[72] Abbreviated Injury Scale (AIS), *Association for the Advancement of Automotive Medicine*, 2018, https://www.aaam.org/abbreviated-injury-scale-ais/.

[73] G. Lindsey, K. Nordback, and M. A. Figliozzi, "Institutionalizing bicycle and pedestrian monitoring programs in three states," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2443, no. 1, pp. 134–142, 2014.

[74] X. Liu and J. Griswold, "Pedestrian volume modeling: a case study of San Francisco," *Yearbook of the Association of Pacific Coast Geographers*, vol. 71, no. 1, pp. 164–181, 2009.

[75] L. F. Miranda-Moreno and D. Fernandes, "Modeling of pedestrian activity at signalized intersections," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2264, no. 1, pp. 74–82, 2011.

[76] S. S. Pulugurtha and S. R. Repaka, "Assessment of models to measure pedestrian activity at signalized intersections," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2073, pp. 39–48, 2008.

[77] San Diego Metropolitan Transit System, "How to Ride: Bikes," 2018, https://www.sdmts.com/rider-info-how-ride/bikes.

[78] City of San Diego, "Vision zero saves lives," City of San Diego Comprehensive Pedestrian Collision Analysis, Transportation and Stormwater Division, 2014.

[79] City of San Diego, *Zero Traffic Related Fatalities and Severe Injuries by 2025*, 2018.

[80] A. P. McGinn, K. R. Evenson, A. H. Herring, S. L. Huston, and D. A. Rodriguez, "The association of perceived and objectively measured crime with physical activity: A cross-sectional analysis," *Journal of Physical Activity & Health*, vol. 5, no. 1, pp. 117–131, 2008.

[81] G. G. Bennett, L. H. McNeill, K. Y. Wolin, D. T. Duncan, E. Puleo, and K. M. Emmons, "Safe to walk? Neighborhood safety and physical activity among public housing residents," *PLoS Medicine*, vol. 4, no. 10, pp. 1599–1607, 2007.

[82] A. Bowen, "New bike sharing program launching in San Diego," *KPBS Public Media*, 2018, http://www.kpbs.org/news/2018/feb/15/new-bike-sharing-launches-san-diego-limebike/.

[83] San. Diego, "Association of Governments," *Bayshore Bikeway Fact Sheet*, 2018.

[84] J. Strauss and L. F. Miranda-Moreno, "Spatial modeling of bicycle activity at signalized intersections," *Journal of Transport and Land Use*, vol. 6, no. 2, pp. 47–58, 2013.

[85] TE. McMillan, "Walking and Biking to School, Physical Activity, and Health Outcomes," in *Active Living Research*, a National Program of the Robert Wood Johnson Foundation, Princeton, NJ, 2009.

[86] The Fatal Fifteen Intersections - Vision Zero, "Circulate San Diego," http://www.circulatesd.org/fatal15sd, 2018.