

# Pedestrian Safety – Fundamental to a Walkable City

Patrick McDevitt<sup>1</sup>, Preeti Swaminathan<sup>1</sup>, Joshua Herrera<sup>1</sup>, and Raghuram Srinivas<sup>2</sup>

<sup>1</sup> Masters of Science in Data Science, Southern Methodist University,  
Dallas, TX 75275 USA

<sup>2</sup> Southern Methodist University,  
Dallas, TX 75275 USA

{pmcdevitt, pswaminathan, herreraj, rsrinivas}@smu.edu

**Abstract.** In this paper, we present a method to identify urban areas with a higher likelihood of pedestrian safety related events. Pedestrian safety related events are pedestrian-vehicle interactions that result in fatalities, injuries, accidents without injury, or near-misses between pedestrians and vehicles. To develop a solution to this problem of identifying likely event locations, we assemble data, primarily from the City of Cincinnati and Hamilton County, that include safety reports from a five year period, geographic information for these events, citizen survey of pedestrian reported concerns, non-emergency requests for service for any cause in the city, property values and public transportation accessibility. We augment the data from Cincinnati with walkability scores obtained from public sources. From this assembled data set we complete both supervised learning and unsupervised learning. The supervised learning, two-part regression, identifies specific areas within the city that have the highest potential for safety improvement. It is these regions that are recommended to be prioritized for resource allocation and remedial action. The unsupervised learning, k-means cluster, is conducted to augment the overall understanding of how different neighborhoods present differing opportunities for improvement with regard to walkability and consequently pedestrian safety.

## 1 Introduction

*An early-morning walk is a blessing for the whole day -- Henry David Thoreau [1]. So, begins the choice every day for urban dwellers -- to walk or not to walk -- to have a blessing as proposed by Thoreau, or to assess the daily commute -- as summarized by Jeff Kober [2]: *My intention is to get done with this commute ... my intention will not be met until I get out of this car* -- as just a rather unpleasant means to get from point A to point B.*

A walkable neighborhood is a neighborhood with the following characteristics : a center (either a main street or public space), sufficient population density to support local businesses and public transit, affordable housing, public spaces, streets designed for bicyclists and pedestrians, and schools and workplaces within

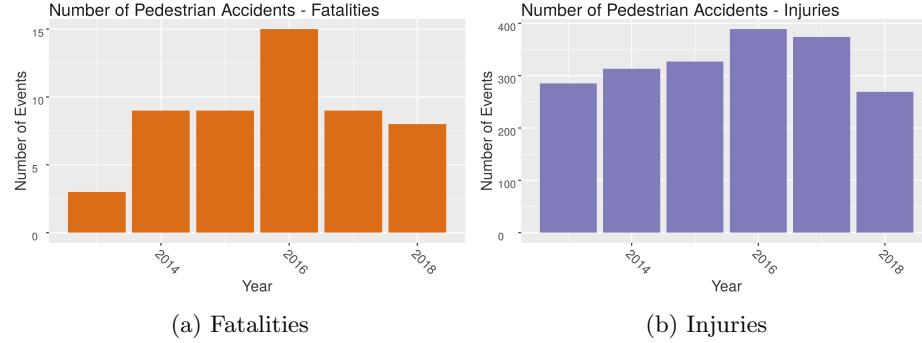
walking distance for residents [3]. As the modern urban landscape has evolved in the US over the last fifty years, pedestrianism was not often on the list of high priorities for inclusion into the development of urban environments. As a result of this trend, there have been real, and negative, consequences: economically, epidemiologically, and environmentally on the inhabitants of many cities in western developed countries [4]. Economically, we can observe that the percentage of income spent on transportation for working families has doubled, from one-tenth to one-fifth of household earnings from the 1970s to current era [4]. If we consider the health effects of urban living patterns, we observe that people living in less walkable neighborhoods are nearly twice as likely to be obese than people that live in walkable neighborhoods [4]. This statistic, coupled with the fact that Americans now walk the least of any industrialized nation in the world [5] indicate a growing health problem due in part to a lack of physical activity. When constructed on a per-household basis, carbon mapping clearly demonstrates that suburban dwellers generate nearly twice as much carbon-dioxide, the main pollutant contributing to global warming [6], than do urban dwellers due to longer commutes and larger houses [4].

There is a growing movement in the US and other western nations to promote the concept of walkable cities as healthier places to live - economically, environmentally and physiologically - than the suburban, exurban, drive-till-you-qualify model of modern western development [7] [8] [9]. As identified in the Toronto Pedestrian Charter [10] the six principles for building a vital urban pedestrian environment include: accessibility, equity, health and well-being, environmental sustainability, personal and community safety, and community cohesion and vitality. According to the city of Toronto, this is the first such pedestrian bill of rights in the world and promotes the concept that walking is valued for its social, environmental, and economic benefits.

The US is experiencing an increase in the number of pedestrian fatalities, reaching a 25-year high in 2017, with nearly 6,000 fatalities [11]. Newspaper articles in the Midwest identify fatal occurrences: [12] "An uptick in pedestrians being hit by cars in the Cincinnati and Northern Kentucky area has officials sounding the alarm. Three crashes just this week resulted in the death of three pedestrians."

For the years 2013 through 2018, Cincinnati experienced on average 12 fatalities and 350 injuries per year, as shown in Figure 1

Fig. 1: Number of Pedestrian Accidents



The supervised learning two-part model produced a linear model for predicting the cost of pedestrian-vehicle interactions and a map of the city highlighting areas with the largest potential for safety improvement. The area with the highest potential for safety improvement was located in the center of the Corryville neighborhood, with at least 8 other areas identified in neighborhoods such as the Central Business District, Westwood, Lower Price Hill, North Avondale, and Hyde Park. The unsupervised learning we implemented generated clusters of pedestrian safety issues and visualized them on the grid map of Cincinnati. The unsupervised clustering identified neighborhoods such as College Hill, Clifton, Central business district, Over-the-Rhine and many more as having prevalent safety concerns.

The remainder of this paper is organized as follows. In Section 2 we provide general information about pedestrian safety, terminology used in subsequent sections of the study, and a valuation associated to the costs incurred in pedestrian accidents. In Section 3 we describe the data sources and data sets used in this study, a description of a grid cell approach to modeling geographic data, the development of model features from the source data, and some graphical representations of the distribution of these features viewed geographically. In section 4 we describe the various models implemented, the results and analysis are presented in Section 5. Ethical considerations of this study are provided in Section 6. The Conclusion is presented in Section 7.

## 2 Pedestrian Safety

The subject pedestrian safety is supported by terminology specific to this domain. A collection of the terminology that we use in this paper is provided in this section.

Prime measurements that are used to report pedestrian safety events are fatalities, injuries, and near misses. The statistics in these categories are quoted

in number of events and are typically stated on an annualized and per capita basis.

There are a range of severities associated to the outcomes of pedestrian–vehicle accidents. A continuous real valued response variable that accounts for both the severity and the frequency of events can be established by accounting for this relative severity. We implement a response variable that is a multiple of the number of events and the cost of the event. The cost basis that we use is based on average severity costs for 5 levels of events, as established by the National Safety Council<sup>3</sup> (NSC) as shown in Table 1. As identified by the NSC, estimating methods are not the actual incurred economic costs, but also include estimations for lost quality of life. The values in the Table 1 are for a per-injured person basis. For our purposes, the actual costs are not used to establish an actual economic impact, but are used to represent a relative severity among the injury types. For instance, a fatality is assessed to be ten times more severe than a disabling injury and two hundred times more negatively impactful than an event where no injury was observed.

**Table 1.** Event Comprehensive Cost Severity, National Safety Council

Severity	Unit Cost (\$)
Fatality	10,082,000
Disabling Injury	1,103,000
Evident Injury	304,000
Possible injury	141,000
No Injury Observed	46,600

Table 2 provides definitions of pedestrian safety related terminology as used in this paper.

---

<sup>3</sup><https://www.nsc.org/Portals/0/Documents/NSCDocuments-Corporate/estimating-costs.pdf>

**Table 2.** Pedestrian safety terminology

Attribute	Description
Potential for Safety Improvement (PSI)	Measures the actual crash cost minus the expected cost of “similar” sites that can be obtained from the crash cost models. In typical usage, an explanatory model using available features is established to predict some measure of cost (e.g., fatality or injury). [13]
Hotspot	Areas with higher density or frequency of pedestrian related accidents [14].
PVI	Pedestrian-vehicle interactions

The focus of many pedestrian safety studies is the interaction between pedestrians and vehicles. Prior works have created statistical models to determine the likelihood of crashes, given information about the time of day, victim’s age, gender, and other features [15] [16] [17] [18] [19] [20]. The study done by Guo [21], et al examined the patterning and structure of road networks as a factor of pedestrian vehicle interactions (PVIs). Zhang et al [22] created a statistical model that classified different types of street crossings to determine which type of crossing was the safest, and gain insight to the relationships between the factors that contribute to a PVI.

In our study, we address the issue of pedestrian safety in a manner that is generally consistent with traffic safety analysis methods, which develop a model, typically via linear regression, and then use the variance between expected and observed to determine the areas with potential for improvement as the areas in which the observed number (or cost severity) of events most greatly exceeds the model expected values. Common in highway traffic studies is the implementation of an empirical Bayesian adjustment to the regression model estimations.

$$Expected_{Bayesian} = \eta \cdot P(E) + (1 - \eta) \cdot O(E), \quad (1)$$

where  $0 < \eta < 1$ ,  $P(E)$  = expected accidents (either in severity or number) from the model for similar locations, and  $O(E)$  is the observed accidents (either in severity or number) for a specific location.  $\eta$  is determined empirically from existing models or published standards[23]

We do not implement that additional step, as our purpose is restricted to identifying locations that have greatest positive variance (observed - expected) in comparison to all of the other grid cells in the modeled region. For this purpose, it is not additional added value to include the adjustment associated to the empirical Bayesian procedure.

### 3 Data Sets

#### 3.1 Grid cell development

To enable the development of both supervised and unsupervised learning models, we employ a grid cell approach. We superpose a grid of equi-distant points over the geographic definition of the city of Cincinnati. The grid points are spaced at approx 250 meter distance. This provides a uniform distribution of 4,196 grid cells to encompass the 80 square mile surface area of the city. Our approach then is to assemble all of the experiences from the various data sources that occurred within a grid cell as representative of that area's experience. The supervised learning model then consists of 4,196 rows (representing each individual grid cell in the map) and as many columns as data features that we complete as independent estimators of the dependent response. In the following descriptions of data sets, we display them in the aggregated grid cell to visualize the relative density of each feature within the city boundaries.

The development of the grid cells, the feature definition from each data set, and the aggregation of the response variables and independent features are all completed using customized R code.

Our rationale for using 250m grid spacing was chosen as a baseline distance for model development. It represents a distance that is intermediate in comparison to other studies using similar approaches[14]. This is a characteristic of the modeling approach that can be considered for optimization in future evaluations or developed as a hyper-parameter to improve model performance.

#### 3.2 Publicly Available Data Sets

Like many cities in the United States, Cincinnati is also a participant in the Open Data experiment. The city freely provides governmental data to the public with the goal that interested individuals and groups will use the data sets in creative ways to improve the quality of life in the city. The source of much of the data that we utilize in this study is accessed from the Cincinnati Open Data portal. Our method was to review each data set accessible from the web portal, intuit whether that data set potentially provided relevant information to include in a predictive model related to pedestrian safety, and if so, download that data set to a local hard drive for future processing. The basis for retaining a data set for incorporation in the model was based on the assumption that data sets that provide information for a five year period, with quantifiable characteristics that characterize infrastructure related to transportation (street density and locations, public transportation accessibility), services provided by the city (non-emergency service calls), and information directly related to traffic and pedestrian accidents. From this basic approach, we identified the data sets as shown in Table 3

In addition to the data available on the Open Data Cincinnati portal, we also utilize two data sets obtained directly from the city's Department of Transportation & Safety (DOTE). The first of these is the result of a survey that was

conducted from February through April 2018. The second is reports of near-miss interactions between pedestrians and vehicles that were independently reported to the city DOTE and retained in a separate database. Both of these data sets were provided to us for our use in this evaluation. Both of these data sets are further explained in a subsequent section of this paper.

Augmenting the data sets available from the City of Cincinnati, we also use data publicly available from additional sources : Walk Score®<sup>4</sup>, Zillow<sup>5</sup>, and Google Maps<sup>6</sup>.

### 3.3 Pedestrian Accidents

From the Open Data Cincinnati portal, we use the data set Traffic Crash Reports (CPD) to identify the pedestrian accident history. From this data set, there are more than 2,100 reported events involving pedestrians in the years 2013 through 2018. Of these, there are 47 reported fatalities, 2008 injuries, and 109 events reported as property damage only. The annual count for fatalities and injuries are shown in Figure 1. As a dependent response characteristic for this model, we distribute each of the events to the appropriate grid cell, apply the comprehensive cost severity factor as identified in Table 1, aggregate the total experience in each grid cell as a dollar severity amount, and then apply a kernel function to distribute the cost severity over the local geographic region. The kernel function approach is employed with the premise that local conditions, beyond just the perimeter of a 250m x 250m region in which the accident occurred, participate in providing the conditions that contribute to a pedestrian accident. Further, the kernel function acts to develop a more continuous distribution of event costs which would otherwise be localized discontinuous point functions with grid cells of high severity being surrounded by grid cells with zero cost. For our purposes, we use a kernel function radius of 0.005 degrees (latitude or longitude). This provides a radius search distance of adjacent grid cells of approximately 500 meters, resulting in the cost severity being distributed to approximately 15 neighboring grid cells. Similar to the decision to use 250m grid spacing for the model, we use the 500m radius inclusion distance for this model as it represents a distance that is intermediate in comparison to other studies using similar approaches. This is a characteristic of the modeling approach that can also be considered for optimization in future evaluations or developed as a hyper-parameter to improve model performance.

The kernel density function that we use distributes the cost severity spatially using the following quartic function[14] :

$$Cost_g = \sum_{i=1}^n \rho_i \left[ 1 - \left( \frac{d_{ig}}{r} \right)^2 \right]^2 C_i, \quad (2)$$

---

<sup>4</sup><https://www.walkscore.com/>

<sup>5</sup><https://www.zillow.com>

<sup>6</sup><https://developers.google.com/maps>

**Table 3.** Data sets supporting model

Data Set	Description
Pedestrian accidents <sup>7</sup>	Traffic Crash Reports (CPD) - Contains date, time, weather, location, road information, severity, demographic information
Traffic accidents	Same data source as for pedestrian accidents
Streets Infrastructure <sup>8</sup>	Street Centerlines (w/ PCI rating) - Street segment locations, length, width, area, material
Bus transportation <sup>9</sup>	SORTA Bus Stops - Bus stops along SORTA bus routes in the city of Cincinnati, locations, line name
Non-emergency requests <sup>10</sup>	Cincinnati 311 (Non-Emergency) Service - dates, request type, location, agency responsible, status
Cincinnati pedestrian safety survey data <sup>11</sup>	Survey input data from citizens; location, date, concern type, comments
Pedestrian near miss data <sup>12</sup>	Locations, dates of near-miss incidents reported directly to the DOTE
Property valuation <sup>13</sup>	Each property transfer, date of sale, purchase amount, property, buyer, and seller information
Walk Score®	A scoring scale rating the walkability of every address in the city
Google Maps	API used to identify latitude and longitude coordinates from local addresses
Zillow <sup>14,15</sup>	Zillow Home Value Index provides time-series of data for median market values, used to support property valuations. Zillow neighborhood provides shapefiles used to delineate neighborhood boundaries in the city.

where  $Cost_g$  is the crash cost assigned to grid cell  $g$ ,  $d_{ig}$  is the distance from the identified pedestrian accident site to the local grid cell centroids,  $r$  is the (constant) search radius,  $C_i$  is the cost severity of the accident  $i$ , and  $\rho_i$  is

<sup>7</sup><https://data.cincinnati-oh.gov/Safer-Streets/Traffic-Crash-Reports-CPD-/rvmt-pkmg>

<sup>8</sup><https://data.cincinnati-oh.gov/Fiscal-Sustainability-Strategic-Investment/Street-Centerlines-w-PCI-rating-/574p-8utc>

<sup>9</sup><https://data-cagisportal.opendata.arcgis.com/datasets/sorta-bus-stops>

<sup>10</sup><https://data.cincinnati-oh.gov/Thriving-Healthy-Neighborhoods/Cincinnati-311-Non-Emergency-Service-Requests/4cjh-bm8b>

<sup>11</sup>direct communication from Cincinnati DOTE

<sup>12</sup>direct communication from Cincinnati DOTE

<sup>13</sup>[https://www.hamiltoncountyauditor.org/transfer\\_policies.asp](https://www.hamiltoncountyauditor.org/transfer_policies.asp)

<sup>14</sup>[https://www.quandl.com/data/ZILLOW/M26\\_NFS-Zillow-Home-Value-Index-Metro-NF-Sales-Cincinnati-OH](https://www.quandl.com/data/ZILLOW/M26_NFS-Zillow-Home-Value-Index-Metro-NF-Sales-Cincinnati-OH)

<sup>15</sup><https://www.zillow.com/howto/api/neighborhood-boundaries.htm>

a normalizing factor for each accident such that the distributed costs for each accident sum to the nominal severity value of that event. For our model, the implementation of this kernel density function was implemented within the R code.

The distribution that results from the application of this data set onto the grid cell domain with the kernel density function applied are shown in figure 2. The experience is that approximately one fourth of the grid cells have no contribution from the cost severity allocation.

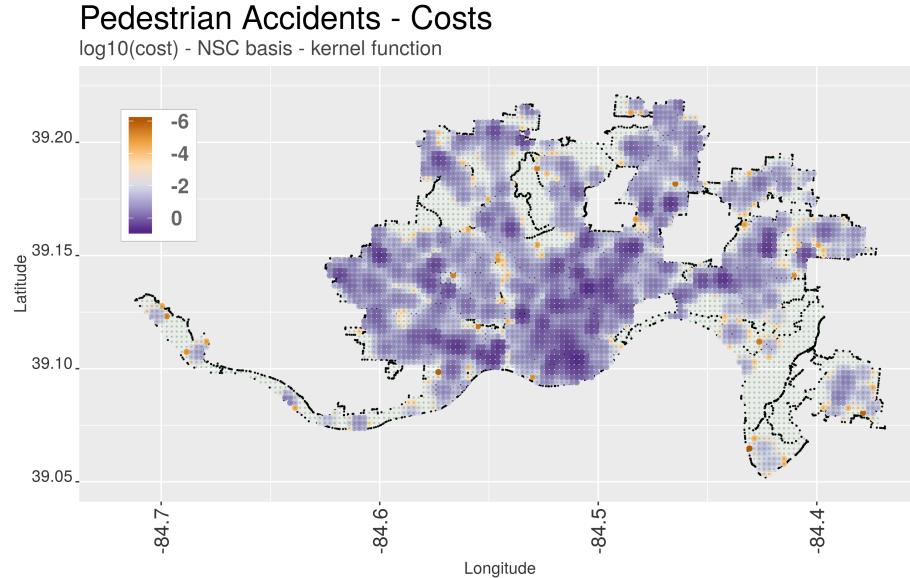


Fig. 2: Traffic Accidents - Pedestrian

This motivates us to consider the supervised learning model as a two-part model : (a) binary estimator to model the presence / absence of pedestrian accident costs, and (b) for the positive case of (a), an estimator of the magnitude of the cost severity function. This will be further outlined in the Methods and Results sections of this paper.

### 3.4 Traffic Accidents

Similar to the pedestrian accident data set in the previous section, the same data set from the the Open Data Cincinnati portal, Traffic Crash Reports (CPD), also includes a definition of every non-pedestrian involved traffic accident that occurred in the same time period. There are more than 190,000 reported traffic accidents that do not involve pedestrians in the years 2013 through 2018. As a candidate independent feature characteristic for this model, we distribute each

of the events to the appropriate grid cell, apply the same comprehensive cost severity factor as identified in Table 1, aggregate the total experience in each grid cell as a dollar severity amount. We do not apply a kernel function to distribute the cost severity over the local geographic region for the independent predictor features. In this case, the density of the actual experience is sufficiently distributed that a kernel function for density diffusion is not needed to achieve wide distribution and sufficient coverage of the accident experience among all of the grid cells. The distribution of the cost severity function for the non-pedestrian involved accidents are shown in Figure 3

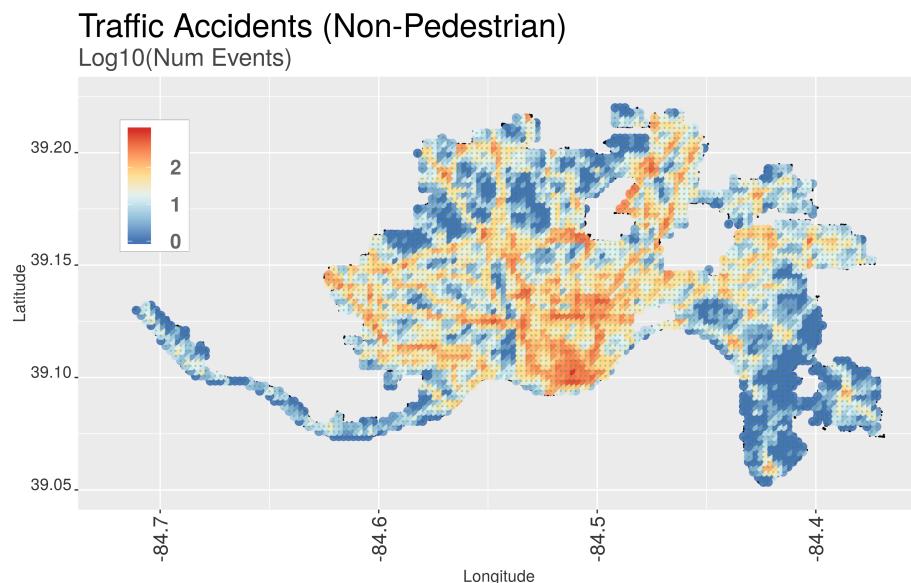


Fig. 3: Traffic Accidents - Non-Pedestrian

### 3.5 Streets Infrastructure

Typical traffic safety models include data features to characterize the density of local traffic conditions. The Open Data Cincinnati portal provides a data set called Street Centerlines (w/ PCI rating) that is used to account for these local traffic conditions. This data set includes a description of each street segment in the city and includes the characteristics : geo-locations of the street segment, surface area, surface length, surface width, number of traffic lanes, and the surface type (asphalt, concrete, etc.). We use this data set to develop a set of features as shown in Table 4. The statistics shown are summary statistics for the sample of data across all of the grid cells.

**Table 4.** Street surface features

Characteristic	1st Quartile	Median	3rd Quartile
Lane Counts	24	53	108
Sum Widths	250	563	1152
Sum Areas	131805	284640	787812
Number Streets	8	19	38

The distribution of the sum of street surface area distributed among the grid cells of this model are shown in Figure 4

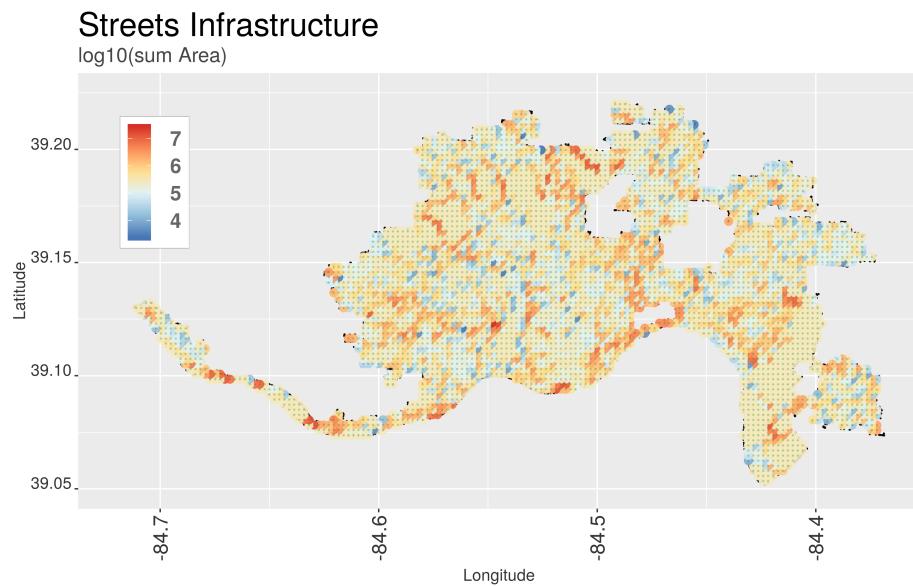


Fig. 4: Streets - Sum of Surface Area

### 3.6 Public Transportation

The data set from the Open Data Cincinnati portal includes the location of every bus stop in the Southwest Ohio Regional Transit Authority (SORTA) network. The SORTA bus network is the only public transportation network in the city. The data set includes 25,980 bus stops identified by latitude and longitude. We calculate the distance from every grid centroid to the nearest bus stop and use that value as a feature for the model.

The distribution of the distances to nearest bus stop for this model are shown in Figure 5

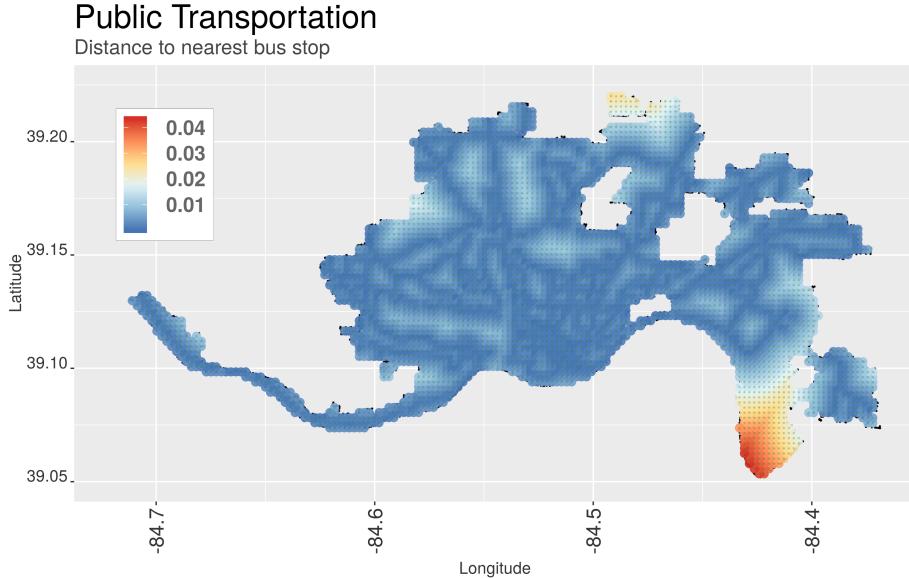


Fig. 5: Public Transportation Access

### 3.7 Non-Emergency Request for Service

The data set from the the Open Data Cincinnati portal includes a definition of every non-emergency request for service. These requests are initiated by the population and include a full range of services requested of city services. There are 613,000 requests that were made in the years 2012 through 2018. As a candidate independent feature characteristic for this model, we distribute each of the requests to the appropriate grid cell, aggregate the number of requests in each major category and also the total number of requests. We do not apply a kernel function to distribute these requests over the local geographic region for the independent predictor features. The density of the actual experience is sufficiently distributed that a kernel function for density diffusion is not needed to achieve wide distribution and sufficient coverage for service request characterization among all of the grid cells. The distribution of the total number of requests per grid cell are shown Figure 6. The number of total requests that were recorded in each category are shown in Table 5.

**Table 5.** Non-Emergency Service Requests (2012 - 2018)

Request Type	Number Requests
Animals/Insects	7765
Building Related	61616
Construction	843
Food	2153
Others	171613
Police Property	8546
Service complaint	4374
Street/Sidewalk	55555
Traffic/Signal	1167
Trash	272932
Trees/Plants	22092
Water Leak	1211
Zoning/Parking	3132
Number Total Requests	612999

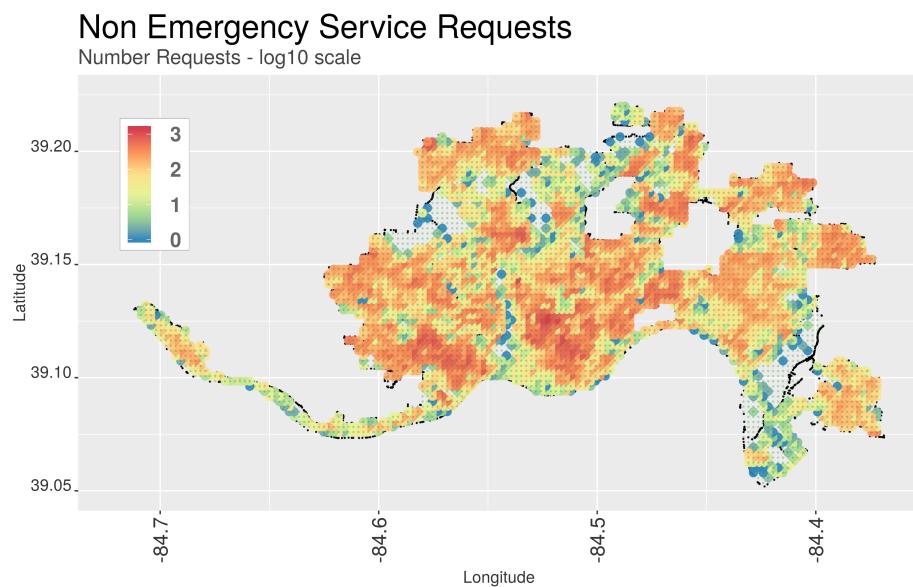


Fig. 6: Traffic Accidents - Non-Pedestrian

### **3.8 Cincinnati Pedestrian Survey**

As an avenue of data collection in response to the number of pedestrian accidents, the City of Cincinnati requested citizen input to identify specific areas in the city which are pedestrian safety concerns. The city created a web-site, which launched in Feb-2018 [24], that allows citizens to specifically identify a location on a map, within a distance of several feet of the area of concern, and report the nature of the concern in a functional user interface. The user interface was available from Feb 2018 to April 2018<sup>16</sup>[24]. The survey screen provided users with view of the city, a drop down of neighborhoods, and a list of issue categories. The user then selected a neighborhood, and selects from the pre-defined issue types to report, and also has an opportunity to write a comment. If another citizen selects the same location and issue type, comments are appended as additional comments. This provides an idea of the number of users having same issue at a particular location. Survey submissions are anonymous. The city plans to use this community input to prioritize maintenance and improvement resources.

As candidate independent feature characteristics for this model, we distribute each registered concern type to the appropriate grid cell, aggregate the number of registered concerns in each major category and also the total number of survey inputs. We do not apply a kernel function to distribute these over the local geographic region. The data features are shown in Table 6. An example of the distribution of the number of citizen inputs are shown in Figure 7

---

<sup>16</sup><http://cagisonline.hamilton-co.org/pedsafetysurvey>

**Table 6.** Pedestrian Survey - Citizen Input

Category	Description	Number
	Accessibility Issue	35
	Crosswalk Needed	311
	Double Parking	52
	Jaywalking	141
	Lack of Visibility	210
	Long Wait for Walk Signal	54
	No Bike Facilities	58
Concern type	No Sidewalks	247
	Other	448
	Parking on the Sidewalk	112
	Parking Too Close to Intersection	127
	Speeding	977
	Vehicles do not Yield at Crosswalks	515
	Vehicles Run Red Lights or Stop Signs	443
	Walk Signal is Too Short	57
	Total Number Concerns	3787
User type	Bikes	154
	Drives	1152
	Travels (Other)	186
	Uses an Assistive Device	66
	Walks	2229

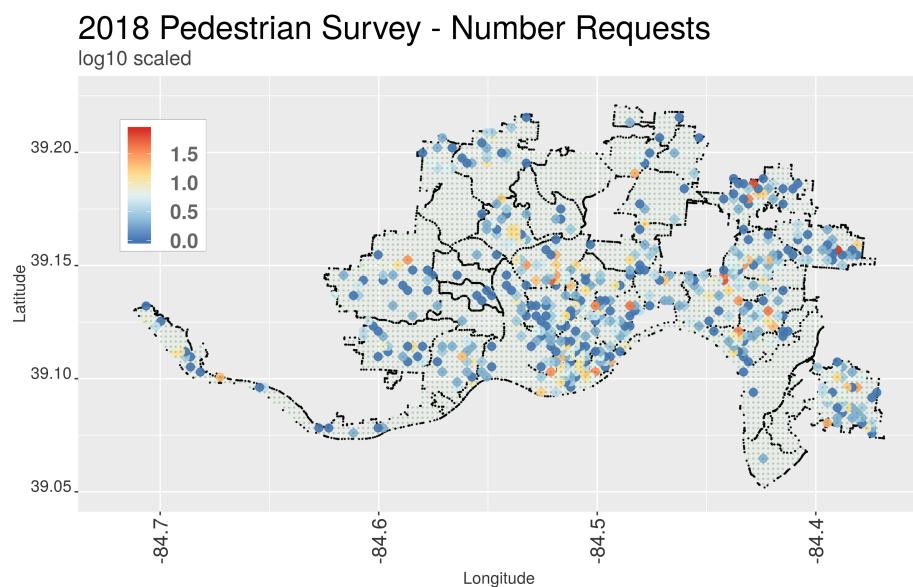


Fig. 7: Median Property Values

### 3.9 Neighborhoods

For some elements of this project, we relied on descriptions of neighborhoods, neighborhood boundaries, and real estate market valuation. To support those elements, we used data sources available from Zillow. Zillow is an online real estate based service providing information about residential properties for sale, for rent, or recently sold<sup>17</sup>. Zillow provides online access to real estate information and services to consumers. A part of the resources made available by Zillow are neighborhood boundary shapefiles. These files are shared under a Creative Commons license, allowing freedom to use the data sets provided attribution is made to Zillow.

We did observe that the Zillow neighborhood definitions include two neighborhoods, Fruit Hill and Forestville, on the eastern boundary of the city which are not, in fact, within the city incorporation limits. We dropped those neighborhoods from our model.

For this evaluation, we used the neighborhood boundaries for several purposes. First, the neighborhood boundary shapefiles for the City of Cincinnati were used to identify grid cells and experiences that are strictly within the boundaries of the city using the latitude and longitude coordinates of the shapefiles. Each grid cell in the model was attributed to one of the forty-six neighborhoods included in the Zillow defined boundaries. These neighborhood assignments were then used to support imputation for grid cells that had missing data for property value information. The grid cells without property values are imputed using the median values within that grid's respective neighborhood. The neighborhood boundaries are used in the unsupervised learning evaluation to differentiate cluster analysis results to named neighborhoods.

An additional resource that we use from Zillow is the Cincinnati Market Overview, which provides an aggregate relative change in market valuation for the Cincinnati residential market for the years 2008 through 2018. In the next section we describe the property valuation information that we use in our model that is acquired from the Hamilton County (OH) Auditor. We adjust the historical information from the Hamilton County Auditor data set using the relative market valuation change that is published on the Zillow web-site. The relative values are shown in Table 7.

---

<sup>17</sup><https://www.zillow.com/>

**Table 7.** Cincinnati Residential Market Appreciation

Year	Median Value (\$K)	Adjusted to 2018
2008	113	1.13
2009	113	1.13
2010	103	1.24
2011	100	1.28
2012	99	1.29
2013	98	1.31
2014	99	1.29
2015	100	1.28
2016	111	1.15
2017	115	1.11
2018	128	1.00

### 3.10 Property Valuation

Cincinnati is located within Hamilton County (OH). The Hamilton County Auditor provides access to data files that describe the property transfers that have occurred in the county via their web-site<sup>18</sup>. We use the last 10 years (2008 through 2018) property transfers as a means to identify local property values. Each property transfer which takes place includes descriptive information such as : property address, transfer date, type of property (residential, commercial, industrial, publicly-owned), binary flag to indicate if property transfer is considered valid (i.e., open market sale or not). For each property transfer, we use the property address information to identify the geo-coordinates for subsequent mapping to appropriate grid cell and identify the properties that are within the City of Cincinnati incorporation limits by superposition with the neighborhood shapefile as described in the preceding section. In addition, we value adjusted each transfer, based on year of sale, using the market valuation adjustments shown in Table 7. Finally, we aggregate all property transfers within a grid cell and create the data features shown in Table 8

---

<sup>18</sup>[www.hamiltoncountyauditor.org](http://www.hamiltoncountyauditor.org)

**Table 8.** Property Value Features

Data Feature
Median Sale Price – Residential – Valid Sale
Median Sale Price – Residential – Non-Valid Sale
Median Sale Price – Commercial – Valid Sale
Median Sale Price – Commercial – Non-Valid Sale
Median Sale Price – Industrial – Valid Sale
Median Sale Price – Industrial – Non-Valid Sale
Median Sale Price – Public Owned – Valid Sale
Median Sale Price – Public Owned – Non-Valid Sale

An example of the distribution of the property values for this model are shown in Figure 8

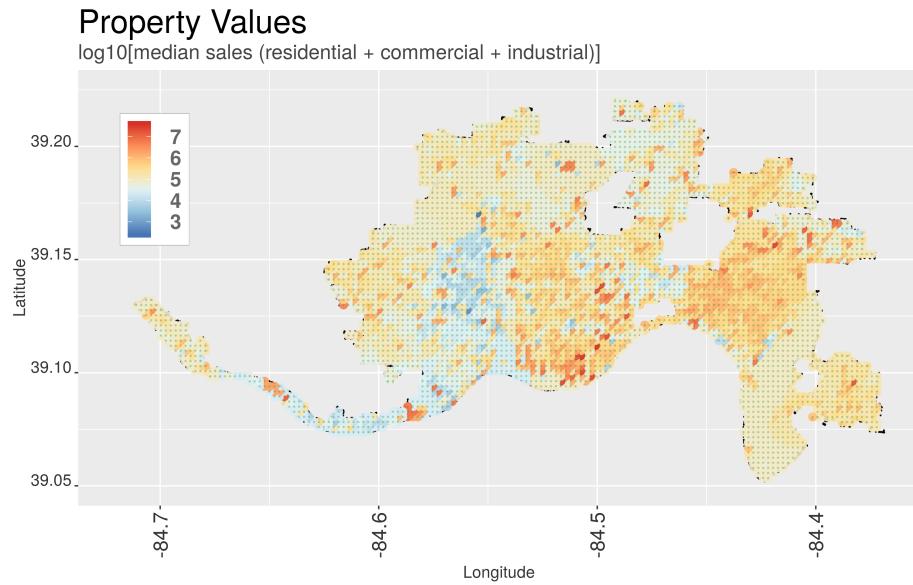


Fig. 8: Median Property Values

### 3.11 Walk Score®

Walk Score is an organization that promotes the concept of walkability and a measurement system associated to walkability. Walk Score has scored every city in the US, using a grid spacing of 500 feet. The walkability score ranges

from 0 (automobile dependent locations) to 100 (highly walkable location, not requiring an automobile for daily living). We use the walkability scores in our model as another measure that includes the interaction between pedestrians and the infrastructure of the city. Similarly to the other data sets, the individual Walk Scores available from the source data are mapped to the 250m grid for this model, and the minimum, mean, and maximum scores within each grid are calculated and implemented as candidate features in the model.

An example of the distribution of the Walk Scores as aggregated for the grid spacing for this model are shown in Figure 9

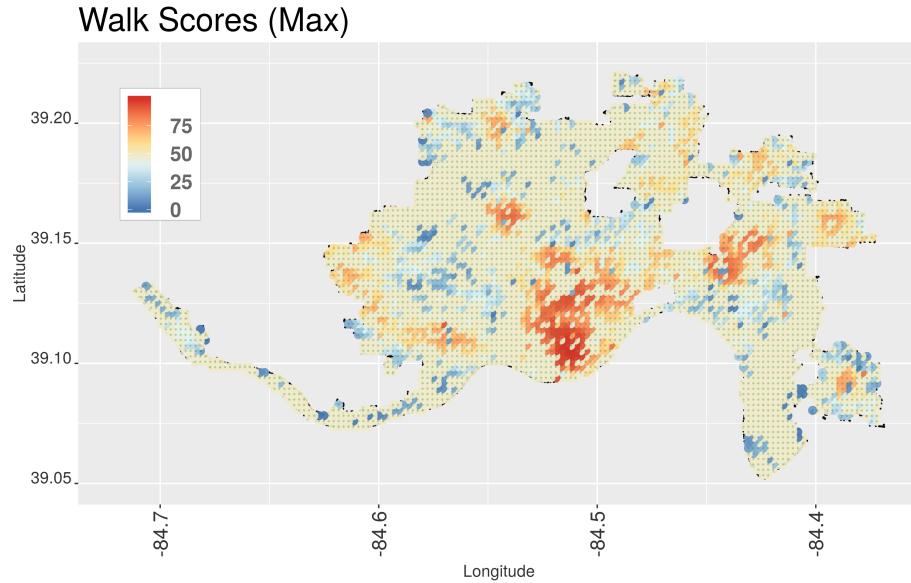


Fig. 9: Walk Score® distributed values

## 4 Methods and Experiments

### 4.1 Two-Part Model - Cost Model

As identified in Section 3.3 the bifurcated distribution of cost severity as distributed among the grid cells encourages the idea to develop a two-part model for the supervised learning element of this evaluation. We consider the two-part model as (a) a method to develop a binary estimator to model the presence / absence of pedestrian accident costs, and (b) for the positive case of (a), develop an estimator of the magnitude of the cost severity function. For the binary estimator we utilize a binary logistic regression model. The model is used to estimate the absence (no cost) and presence (cost > 0). For the second step in the two-part

model, we use a multi-variate linear regression model (on the log-transformed costs) to provide an estimation of the expected cost for the case that the cost is greater than zero. The description of this two-part model is expressed in the following relation :

$$E[Y|X] = \Pr(Y > 0|X) \times E(Y|Y > 0, X), \text{ where} \quad (3)$$

$E[Y|X]$  represents the model estimated cost severity,  $\Pr(Y > 0|X)$  is the resulted produced from the binary logistic regression (either 0 or 1) and  $E(Y|Y > 0, X)$  is the continuous valued cost severity estimation from the linear regression for the subset of grid cells that are estimated to have non-zero cost severity from the binary logistic regression.

This approach to modeling is commonly used in processes that have elements that may or may not participate in an experience, and for this samples that do participate there are a wide range of level of response, often log-normally distributed. Examples of this include health care cost studies, in which some participants in a study just do not visit medical facilities and of the participants who do participate in health care programs, some just did not incur expenses in the study period. The two-part model allows that there are two underlying populations : one of the populations is never likely to incur costs, while the second population does incur expenditures and sometimes they happen to be zero [25]. For a pedestrian safety model, we can consider similarly that two population zones exist, and the purpose of the two-part model is to identify segregation of the two populations (via the binary logistic model) and then provide estimates of cost severity among the zones that have some likelihood of event occurrence.

After the observed pedestrian events are mapped to the grid cells with the kernel function, approximately 25% of the grid cells have zero cost. It is this population in contrast to the remaining three fourths of the grid cells that are modeled with the binary logistic regression. The grid cells with expected non-zero values from the binary logistic model are then modeled using a linear regression model. In both cases, the model are built using the software R with the caret package to support cross validation and with stepwise selection for feature selection.

The 10-fold cross validation is used to reduce over-fitting and improve the generalizability of the model. The 10-fold cross validation splits the data into 10 folds. In each of the 10 iterations, one fold is selected to be the testing dataset, and the other 9 are used to train the model. This methodology is iterated throughout so that each of the 10 folds has a chance to be the testing dataset and the result of these 10 iterations is averaged. Binary logistic regression with the purpose of classifying zero and non-zero values is executed in this fashion.

Cells with a predicted non-zero value are used to create the linear model through the use of 10-fold cross validated stepwise selection process in order to determine the best complex model, ranging from 1 to 40 predictor variables. Model over-fitting can occur when a model is trained on a particular dataset to the degree that it performs extraordinarily well in predicting on the dataset, but

it fails to achieve that same level of prediction capability on other datasets. The parameters in an overly fit model do not generalize to the population at large because they are fine tuned to only reflect the dataset on which they were trained. The preliminary complex linear regression model chosen is the model with the lowest average root mean square error value, and in this study, is determined to be the model with all 40 variables.

After this preliminary model is defined, the number of features in the model is then further selectively pruned based on the variance inflation factor values and significance. For this model, variance inflation factor values larger than 5 are considered to be sufficiently collinear that they are removed from the set of mode predictors.

Variable significance is determined through the use of p-values. For this model, variables at or below the .05 p-value threshold are determined to be statistically significant to the model, while those above it are not. Variables with large p-values are removed from the model due to not significantly contributing to the predictive capability and are shown in Table 9.

With this final linear regression model, the cost of PVIs is predicted for each grid cell. The resulting prediction is then compared against the actual cost of observed incidents. The difference between the predicted and actual cost is the residual. For the evaluation of PSI, the grid cells with greatest positive residual (observed – expected) are determined to be the grid cells with the greatest potential for safety improvement. This determination is based on the fact that the cells with positive residual observed a higher severity cost than did other similar cells. Similar in this case being defined by similarity of the vector of defining independent predictor variables. To complete the evaluation, the residuals are then mapped back to the geographic map of Cincinnati to visualize the areas with greatest potential for safety improvement.

#### 4.2 Unsupervised Learning - Neighborhood Characterization

For unsupervised learning, the goal was to identify pedestrian related issues by neighborhood. T-distributed Stochastic Neighbor Embedding (t-SNE) [26] is a machine learning algorithm for visualization developed by Laurens van der Maaten and Geoffrey Hinton. It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data into low-dimensional space of two or three dimensions for visualization. The data is run through the t-SNE dimensionality reduction algorithm to visualize data in two dimensions and was run for varying perplexities of 10, 20, 50, 100, 200, and 500 to obtain a visually effective perplexity value. Perplexity 20 provides us a better result than all other perplexity values<sup>19</sup>.

To obtain optimal clustering, the original data is processed using algorithms such as k-means and hierarchical clustering. The k-means clustering with three clusters reported the highest silhouette score<sup>20</sup>. The labels created with the t-

---

<sup>19</sup><https://distill.pub/2016/misread-tsne/>

<sup>20</sup>[https://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_kmeans\\_silhouette\\_analysis.html](https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)

SNE algorithm are combined with the clustering labels of k-means, to generate a map of the city with three color-coded clusters, with each cluster representing the prevalence of different issues throughout the city.

## 5 Results and Analysis

The focus of this effort is to identify locations within the city that indicate highest potential for safety improvement. Consistent with the concept as defined and implemented in other works, the model incidents for which the observed experience exceeds the expected (model output) response are the regions with potential for safety improvement. The regions of PSI are the false negative points on the grid map. For these grid cells, the model result was no expected pedestrian events, yet the observed experience exceeded this expected response.

### 5.1 Two-Part Model - Cost Model

The results from this section provide a numerical and visual way to interpret the model. The first part of the modelling involved the binary logistic regression, whose receiver operating characteristic (ROC) curve is shown below in Figure 10. The ROC curve has an 88% area under the curve, which suggests this model has good predictive capabilities, and is backed up with a 10-fold cross-validated accuracy of 82%.

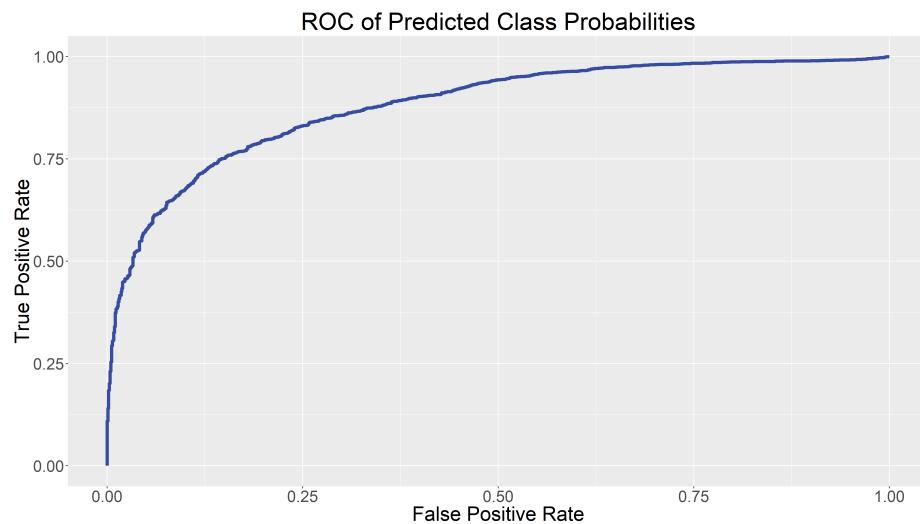


Fig. 10: ROC curve showing predicted class probabilities of the logistic regression.

The regression model is able to explain 34% of the variance in the cost of PVIs with an  $R^2$  value of .3464 and an adjusted  $R^2$  value of .3436. This model

consists of many significant variables, such as reports of faulty traffic signals and the presence of construction, and one suggestively significant variable which make sense practically, but does not meet the p-value threshold of .05. Reports of double parking (p-value = .0544), is when a vehicle parks alongside another vehicle which is already parked on the side of the road, leading to the double-parked vehicle to taking up a lane on the street. Most variables present within the model are reports to the non-emergency police telephone number, or the city's online safety survey. For each incident of report within the grid cell, the predicted total cost of incident changes by the estimated amount. Other non-report variables include the mean Walk Score value of properties within the grid cell, sum of lanes within the grid cell, distance to nearest bus stop, and the total cost of damage to people involved in non-pedestrian related (motor vehicle) accidents. All of the model coefficients can be found below in Table 9.

**Table 9.** Analysis of simplified model regression variables.

Variable	Estimate (\$)	Std. Error (\$)	t-value	p-value
Intercept	-45,600	29,300	-1.555	0.12002
Non-Pedestrian Damage Costs	294.00	14.40	20.4	< 2e-16
Misc. Non-Emergency Calls	1,330	176	7.539	5.75E-14
Mean Walk Score	3,820	604	6.33	2.71E-10
Unsanitary Food Operations	23,400	3,940	5.924	3.39E-09
Sum of Lane Count	-436.0	85.7	-5.083	3.87E-07
Distance to Nearest Bus Stop	-6,110,000	1,350,000	-4.519	6.37E-06
Faulty Traffic Signals	49,900	11,100	4.482	7.59E-06
Comm Prop, Valid, Median	0.0165	0.0051	3.248	0.00117
Animals, Roadkill, and Insects	6,430	2,050	3.132	0.00175
Comm Prop, Non-Valid, Median	0.0149	0.0048	3.1	0.00195
Speeding	-43,200	15,600	-2.769	0.00565
Vehicles not Yielding at Crosswalks	62,100	24,100	2.578	0.00998
Construction Present	31,500	12,800	2.47	0.01353
Vehicles Parking Close to Intersection	104,000	43,600	2.378	0.01747
Number of Near Misses	77,500	32,800	2.36	0.01832
Vehicles Running Red Lights	45,000	21,600	2.081	0.03753
Riders of Pedestrian Conveyances	70,100	35,700	1.964	0.04963
Double Parking	155,000	80,700	1.924	0.0544

Figure 11 illustrates the model's predictive capability as compared to the cost of incident in reality. Points which are colored in red have positive residuals, meaning the model has underestimated the cost. Green points have negative residuals, which mean the model has overestimated the cost. The further away a point is, the larger its absolute residual value is. These distances are displayed on Figure 12, where the red colored areas correspond to the distant red points. These represent areas within the city in which the model has underestimated

the cost the most have the most potential for safety improvement, because all other factors being equal, there is something about that area which is causing a higher amount of damage to pedestrians.

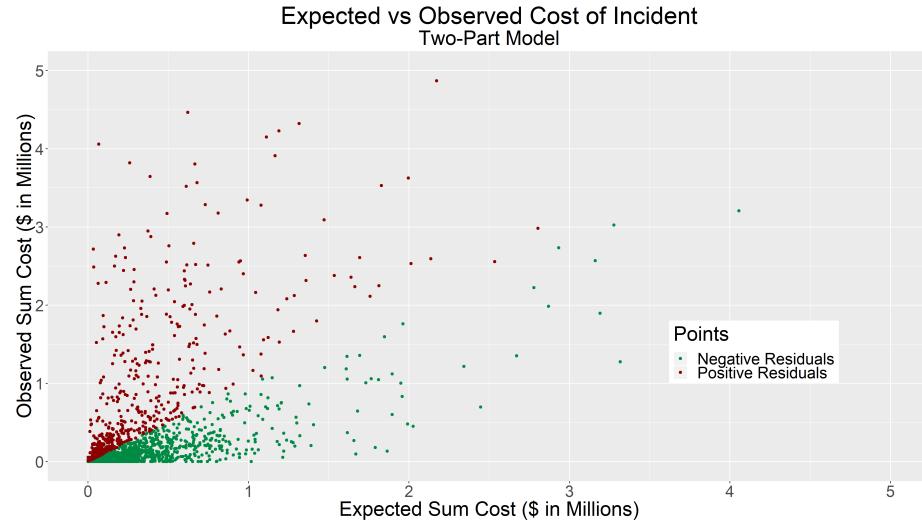


Fig. 11: Comparison of expected and observed PVI costs.

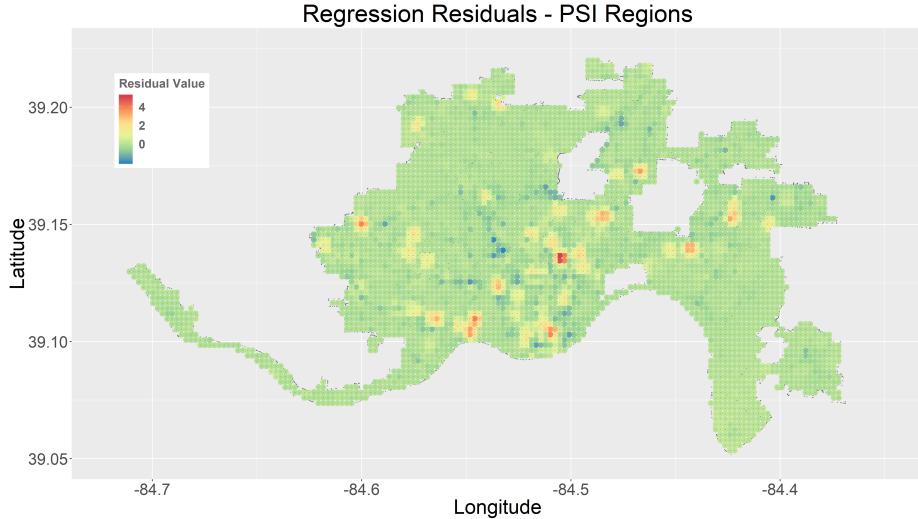


Fig. 12: A map of Cincinnati identifying areas with PSI in the darkest red.

The diagnostic plots for the simplified model are shown in Figure 13 and the most evident absurdity is the normal Q-Q plot shown in the top right. While many of the points are along the line, the Standardized residuals taper greatly at the ends of the plot. The residuals vs fitted values plot on the top left shows good evidence of linearity and equal error variance. The phenomenon of the seemingly straight diagonal sheer in the data is explained by the fact that the residual can not be negative for low fitted values because the model can not predict a negative cost. Given an fitted value of .25, the ceiling of numbers is infinity, but the floor of values is .25, which would occur when the actual value is .25 but the model estimates 0, resulting in a -.25 residual. Considering the fact that the lowest possible residual is equal to the fitted value as represented by the line  $Y = -X$ , the data are as randomly spread about the horizontal line as possible.

The parametric linear regression analysis is susceptible to outliers in such a way that highly influential outliers can reduce the predictive capabilities of the regression model. The Cook's distance metric is a representation of how the rest of the model would be affected if the observation were deleted. Deviant Cook's distance points consist of those with values of 1 or greater, and the central right plot in Figure 13 shows the highest observation having a Cook's distance value around .35. While it has the largest value of points within the dataset, it is not a cause for concern. The lower left plot is the Standardized residuals versus leverage and also aims to view the effect of potentially influential outliers that may curb the regression line towards it. Abnormalities would consist of high leverage values with high Standardized residuals that would appear at the far top right and bottom right areas on the plot. Observation 689 and 2021 are

marked as potentially influential cells in our dataset, and need to be considered when taking the model into account.

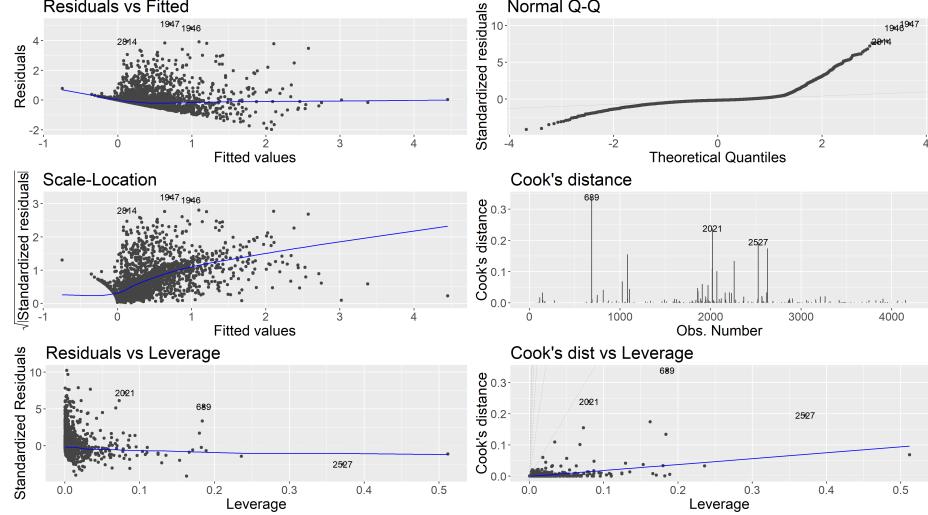


Fig. 13: Diagnostic plots for the simplified linear model.

## 5.2 Unsupervised Learning - Neighborhood Characterization

Figure 14a is the result of the k-means clustering algorithm and is split into three distinct, colored areas. The blue cluster is identified by highly populated areas which are distinguished by a prevalence of water leaks, faulty traffic signals, construction presence, and poor sidewalks compared to the other areas. The red cluster is identified by sparsely populated areas where there are many reports and complaints regarding lack of parking zones, improper or inefficient waste disposal, and reports of animals, roadkill and insects. The third cluster, yellow, is differentiated by reports of unsanitary food conditions, improper or inefficient vegetation removal, have a relatively higher presence of law enforcement facilities, and Inspection request for ADDI (American Dream Downpayment Initiative)<sup>21</sup> which include: ceiling leaks, water damaged walls, and mold,

Narrowing the scope to pedestrian safety related issues, the blue clusters have a higher incidence of reporting long waits for walk signals, low visibility, and lack of crosswalks. The yellow clusters are signified by lack of sidewalks, jaywalking, and speeding, while the red colored cluster areas have a higher rate of reports that walk signals are too short, vehicles are parking too close to intersections, and vehicles not yielding at crosswalks.

<sup>21</sup><https://www.cincinnati-oh.gov/community-development/linkservid/EDCBB922-8209-43D7-979199C1CB458EC3/showMeta/0/>

Fig. 14: Cluster results on map of Cincinnati

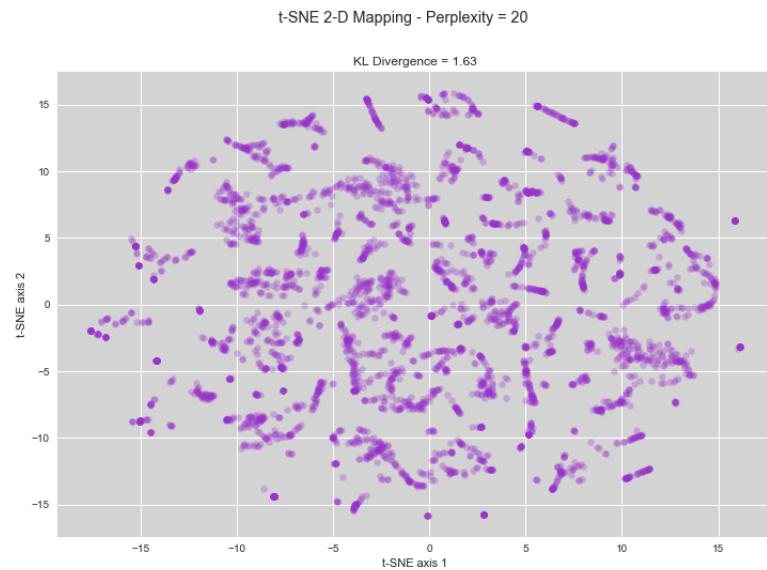
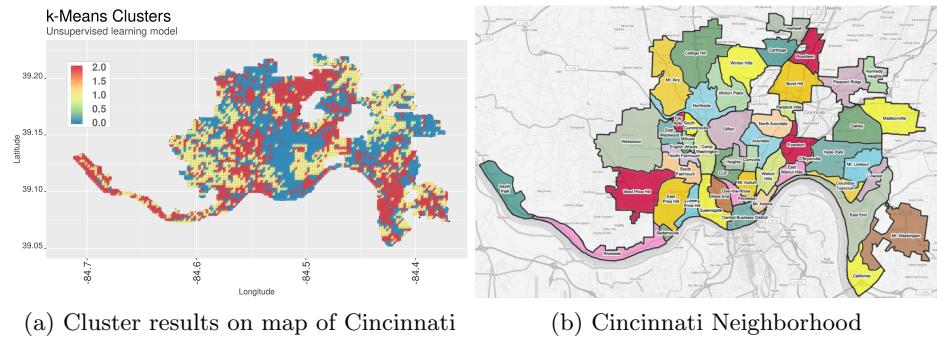


Fig. 15: Spread of data visualized using t-sne.

## 6 Ethics

As a means to evaluate compliance with ethical considerations, we use the model of the ACM Code of Ethics (the Code). Within the Code, there are four primary sections, e.g., General Ethical Principles, Professional Leadership Principles, etc., with each primary section providing additional subsections for self-assessment compliance to the Code. For each sub-section, we self-scored categorically as either Y, n/a, or D, where Y indicates that the work completed for this project rather obviously complies with the Code, n/a indicates that that section of the Code is less obviously significant for this project, and D indicates that that section of the Code identifies a potential ethical dilemma that is worthy of additional discussion to demonstrate compliance or at least point out the potential ethical challenge identified from this self-assessment. The most significant elements that we self-assess as D are : §1.2 - Avoid harm; §1.4 - Be fair and take action not to discriminate; and §3.7 - Recognize and take special care of systems that become integrated into the infrastructure of society.

From these three elements, we consider that the significant question to evaluate is: how may these research findings be interpreted and used ?

The result of this project provides a recommended prioritization for the allocation of municipal resources for the purpose of improving a pedestrian safety problem. Allocation of public resources is often as much a political challenge as it is a scientific challenge. There is no global objective function that assigns absolute social value to any decision of resource allocation. That is, in fact, the work and challenge of public officials. Within the general framework of public decision making it is recognized that facts, reports, and recommendations which are or were essentially the result of objective research are frequently interpreted in a way that suits the interpreter for their own agenda – in some cases for personal gain – financially or politically. We have to admit for this case, then, that this evaluation is potentially subject to a personally motivated interpretation. The debate about using scientific research to guide public policy is long and continuing. With that recognition, the task falls to us to identify what steps are taken to reduce the risk of unintended uses of this report.

First, the report as written has limited scope for direct application to policy. The recommendations included are applicable to the specific time period and data evaluation associated to the City of Cincinnati. The methods presented here can be widely applied (and in fact, that is the goal of this research), but in current form it would be difficult to justify using these results for direct resource allocation in any other municipality. The model developed here used very specific local experiences – accidents, reported near accidents, local conditions survey, property valuations, social media, and all other elements that contributed to this model are local and specific to the City of Cincinnati and to the current time period. As such, the specific recommendations are not generalizable. The method is generalizable; the specific results may provide indications of what local elements in other municipalities may prove indicative or at least useful for a similar exercise, but in any rational discourse, it would be difficult to extend the

specific recommendations from this study to municipalities beyond the extent of this study.

Secondly, this report is submitted to representatives of the Cincinnati City Council and the Department of Safety . By distributing the results to more than one department and to a reasonably broad audience reduces risk of information being used for narrowly scoped or interests which are not generally aligned with good public discussion, debate, and utilization. Further, the data and methods deployed here were developed based on early and continuing input from representatives of these multiple departments within the City. Thus, early and often participation of multiple stakeholders provided the opportunity to have balanced input to the research, thus improving likelihood of balanced output and utilization. And, by respecting and incorporating the input from multiple perspectives from within the City provides higher likelihood of acceptance (and perhaps adoption) of the resulting recommendations.

Thirdly, within the City of Cincinnati, there are currently several on-going initiatives dedicated to improving pedestrian safety. The other initiatives are, in some cases, significantly funded, and are the work product of several departments within the City, predominantly the Department of Safety, that are the prime stakeholders in promoting public safety in the City. These other initiatives are an order of magnitude more significant, both from resource commitment, and for intended impact, than is this study. It is not our aim to minimize the potential impact of the recommendations from this study, but we are cognizant of the relative significance of this study within the larger context of the on-going programs within the City. In any decision making forum for the City, we consider the likelihood of these results having the capability to be used for unintended or inappropriate outcomes to be sufficiently unlikely.

The ethical considerations associated to this project are adequately assessed in the spirit of the ACM Code. The identified risks are appropriately mitigated.

## 7 Conclusions and Future Work

This paper presented an analysis to quantify the relative potential for pedestrian safety improvement for each of the 4,196 grid cells defined for this evaluation. Each of the grid cells represents an approximate 250m by 250m surface area within the city of Cincinnati, Ohio. The quantified PSI for each grid cell is determined based on the observational data that is acquired from publicly available sources. The data used to develop the expected cost severity functions are varied in nature – property values, requests for non-emergency service, access to public transportation, along with observed frequencies and severities of traffic accidents not involving pedestrians. This approach to pedestrian safety modeling that does not rely on specific infrastructure definitions (e.g., intersection types) or actual traffic counts (e.g., vehicle miles traveled per street segment) is a novel approach that the authors have not seen previously published.

The data sets utilized in this study are not in the range typically characterized as big data; however, the volume of data processed to arrive at the feature

definitions is still substantial. The extraction of predictive features from the non-emergency experiences required characterizing more than 600,000 individual service calls. The regression model indicates that a wide range of experiences from moderately-sized data sets can provide model predictive performance that compares favorably with traditional models that are based more closely on traffic patterns and infrastructure characteristics. This study produced adjusted  $R^2$  values of 0.34, while studies based on traffic counts and infrastructure definitions provide adjusted  $R^2$  values in the range of 0.25. Thus, the use of publicly available data that is recorded for reasons other than its use in this study provides a path to complete statistical studies of sufficient performance while at the same time reducing reliance on more costly data gathering methods (traffic counts, for instance).

The grid cells that are identified to have the greatest PSI are in the neighborhoods of: Corryville, Central Business District, Westwood, Lower Price Hill, Avondale, and Hyde Park among others. These specific regions are recommended to be prioritized for remediation.

The unsupervised learning application completed here also identifies useful results to gain insights into the nature of unique advantages and disadvantages from the perspective of pedestrians interacting with their local environment. The results of the clustering analysis conducted from these data sets identify three unique clusters, each identifying different opportunities for improvement and each in a set of different neighborhoods. The results identify that crosswalks and visibility are problematic in the neighborhoods of College Hill, Clifton, the Central Business District, and Over-the-Rhine; that speeding and lack of sidewalks are prevalent in West Wood, West Price hill, Pleasant Ridge, and Kennedy Heights; and that vehicles violating traffic signals and failing to yield at crosswalks are dominant characteristics in Sayler Park, East End, Mt. Washington, Winton Hill, and Carthage. It is worth noting that this classification of differing neighborhood characteristics was arrived at with the data sets which were not assembled for the purpose of understanding urban walkability needs, yet as modeled and evaluated from the unsupervised learning, useful insights are developed.

For future consideration, this work developed a solution using logistic and linear regression methods. The opportunity also exists to explore additional modeling methods from the machine learning toolkit to evaluate potential model improvements from that approach. Random forest modeling is a likely candidate next step, as random forest models can produce good accuracy in data sets of non-normal distributions.

An obvious next step to evaluate improvements in this approach is to include the traditional vehicle miles traveled and infrastructure definitions as predictor variables. The combination of the features derived for this model along with the additional granularity of traffic patterns can likely provide additional improvements in model performance. This study demonstrates that data sets that characterize the interaction of citizens to their environment support reasonable model results; there is every reason to consider additional types of similar data

(e.g., locations of schools, churches, hospitals, various types of establishments) as potential improvement opportunities. From a different aspect, the location tracking devices that are ubiquitous in cell phones and smart watches likely provide a level of detailed characterization between the movement of people and their environment, including development of time series interaction with local geography.

## References

1. Thoreau, H.D., Torrey, B., Sanborn, F.B.: The Writings of Henry David Thoreau.... Volume 20. Houghton, Mifflin (1906)
2. Bowen, A.: Zen commute can take you to a better place. Chicago Tribune
3. Walk Score: Walkable neighborhoods. (2018)
4. Speck, J.: Walkable city: How downtown can save America, one step at a time. Macmillan (2013)
5. Lee, S.M.: Suburban living linked to bigger carbon footprint. SFGate
6. National Geographic: Climate 101 : Air pollution. National Geographic
7. Leyden, K.M.: Social capital and the built environment: the importance of walkable neighborhoods. *American journal of public health* **93**(9) (2003) 1546–1551
8. Steffen, A., Gore, A.: Worldchanging. Das Handbuch der Ideen für eine bessere Zukunft. München: Knesebeck (2008)
9. Doyle, S., Kelly-Schwartz, A., Schlossberg, M., Stockard, J.: Active community environments and health: The relationship of walkable and safe communities to individual health. *Journal of the American Planning Association* **72**(1) (2006) 19–31
10. Toronto City Council: Toronto pedestrian charter unveiled at council. Toronto (2002)
11. Domonoske, C.: Pedestrian fatalities remain at 25-year high for second year in a row. National Public Radio (2018)
12. Kelley, A.: Police stress pedestrian safety after three fatal crashes. WLWT Cincinnati (2018)
13. Ohio Department of Transportation: Safety study guidelines. Office of Systems Planning and Program Management (2017)
14. Xie, K., Ozbay, K., Kurkcu, A., Yang, H.: Analysis of traffic crashes involving pedestrians using big data: Investigation of contributing factors and identification of hotspots. *Risk analysis* **37**(8) (2017) 1459–1476
15. Brüde, U., Larsson, J.: Models for predicting accidents at junctions where pedestrians and cyclists are involved. how well do they fit? (1993)
16. LaScala, E.A., Gerber, D., Gruenewald, P.J.: Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. *Accident Analysis & Prevention* **32**(5) (2000) 651–658
17. Lyon, C., Persaud, B.: Pedestrian collision prediction models for urban intersections. *Transportation Research Record: Journal of the Transportation Research Board* (1818) (2002) 102–107
18. Ladron de Guevara, F., Washington, S., Oh, J.: Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in tucson, arizona. *Transportation Research Record: Journal of the Transportation Research Board* (1897) (2004) 191–199

19. Pulugurtha, S.S., Sambhara, V.R.: Pedestrian crash estimation models for signalized intersections. *Accident Analysis & Prevention* **43**(1) (2011) 439–446
20. Ukkusuri, S., Hasan, S., Aziz, H.: Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. *Transportation Research Record: Journal of the Transportation Research Board* (2237) (2011) 98–106
21. Guo, Q., Xu, P., Pei, X., Wong, S.C., Yao, D.: The effect of road network patterns on pedestrian safety: a zone-based bayesian spatial modeling approach. *Accident analysis and prevention* **99** (2017) 114–124
22. Zhang, C., Zhou, B., Chen, G., Chen, F.: Quantitative analysis of pedestrian safety at uncontrolled multi-lane mid-block crosswalks in china. *Accident analysis and prevention* **108** (2017) 19–26
23. Kolody, K., Perez-Bravo, D., Zhao, J., Neuman, T.: Highway safety manual user guide. *Transportation Research Board Report Number FHWA/AASHTO* (2014)
24. Department of Transportation and Engineering: City launching pedestrian safety survey for all 52 neighborhoods. City of Cincinnati (2018)
25. Buntina, M.B., Zaslavskyb, A.M.: Too much ado about two-part models and transformation?: Comparing methods of modeling medicare expenditures. *Journal of Health Economics* **23** (2004) 525–542
26. wiki: t distributed stochastic neighbor embedding