Pedestrian Safety – Fundamental to a Walkable City

Patrick McDevitt[1], Preeti Swaminathan[1], Joshua Herrera[1], and Raghuram Srinivas[2]

[1] Masters of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
[2] 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 a near–miss between pedestrian and vehicle. The occurrence of these events detracts from the safety of the citizens of the city. To develop a solution to this problem of identifying likely locations of events, we assembled data sets, primarily from the City of Cincinnati, that include safety reports from a 5 year period, geographic information for these events, a citizen survey that identifies pedestrian reported concerns, and a database of all requests for service for any cause in the city. We augmented that data with sports-tracking geolocation movement data of pedestrians, runners, and cyclists. From this assembled data set we completed unsupervised learning using a self–organizing map, excluding the event data. The event data was then mapped into the self-organizing map clusters to identify the statistical likelihood of events in each cluster. The results indicate a statistically significant association between clusters and events. The results identify locations in urban areas for prioritized remediation enabling a proactive approach to improved pedestrian safety.

1 Introduction

An early-morning walk is a blessing for the whole day - Henry David Thoreau [**?**]. 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 [**?**]: *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 [?]. 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 [?]. 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 [?]. So much so that working families are currently spending more of their budget on transportation than housing. 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 [?]. This statistic, coupled with the fact that Americans now walk the least of any industrialized nation in the world [?] 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 [?], than do urban dwellers due to longer commutes and larger houses [?].

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 [?] [?] [?]. As identified in the Toronto Pedestrian Charter [?] 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 [?]. Newspaper articles in the Midwest identify fatal occurrences: [?] "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."

As one avenue of response, the City of Cincinnati has 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 [?], 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 city plans to use this community input to prioritize maintenance and improvement resources.

2 Pedestrian Safety

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

Prime measurements 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 the both the severity and the frequency of events can be established by accounting for this relative severity. We have implemented a response variable that is a multiple of the number of events and the cost of the event. The cost basis that we used is based on average severity costs for 5 levels of events, as established by the National Safety Council [**?**] as shown in Table 1.

Table 1. Event Cost Severity, 2012 NSC

Severity	Unit Cost ($)
Fatality	4,538,000
Incapacitating injury	230,000
Non-Incapacitating injury	58,700
Possible injury	28,000
No injury	2,500

Table 2 below is presented as a primer on pedestrian safety related terminology along with an explanation of the significance of each term in relation to the evaluation presented in this paper.

Table 2. Pedestrian safety terminology

Attribute	Description
U.S. Census Bureau units of measure	The U.S. Census Bureau reports data within geographic units. The census block is the smallest geographic unit used by the Census Bureau. Census blocks are typically bounded by streets or natural features. There is no standard size, either by surface area or population, that is used by the Census Bureau. The data reported in a census block is 100 per cent of households reported data. There is no sampling or estimations used in census block reported data. Census blocks are assembled into block groups, and block groups then constitute census tracts [**?**].
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). [**?**]
Vehicle miles of travel (VMT)	A method to account for volume of vehicle traffic. The value is the total annual miles of vehicle travel within a specified zone. Values available from the U. S. Dept of Transportation can be aligned with U.S. Census Bureau urbanized areas. VMT is sometimes also characterized on a per capita basis.[**?**]
Hotspot	In this context, hotspots are areas with higher density or frequency of pedestrian related accidents [**?**].
Regression–to–the mean	Regression to the mean; since traffic fatality events are low volume incidents (on the order of magnitude of 10s for most cities). RTM is a consideration in studies of pedestrian safety because an area in which fatalities occur in one year may not be repeated in the next, even in the absence of implemented changes. [**?**]
Conditional script questionnaire (CSQ)	A survey tool for assessing expected human behavior under alternative situations. In the context of pedestrian safety, a CSQ requests respondents to self–assess their likelihood to ignore the driving code under different scenarios. The CSQ responses are used in establishing sub–populations that pose increased risk for pedestrian safety. The sub–populations can be based on demographic features (e.g., gender, age) or situational features (e.g., late for work) [**?**].

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, gen-

der, and other features [?] [?] [?] [?] [?] [?]. The study done by Guo [?], et al examined the patterning and structure of road networks as a factor of pedestrian vehicle interactions (PVIs). Zhang et al [?] 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 addressed the issue of pedestrian safety in regard to PVIs by proactively identifying intersections which are of high risk, as opposed to prior studies which have only identified the contributing factors.

3 Data Sets

CAGIS, Cincinnati Area Geographic Information System launched an online survey from Feb 2018 to April 2018 for pedestrian safety [?]. Users login to http://cagisonline.hamilton-co.org/pedsafetysurvey/ . The survey screen provides users with view of the city, with a dropdown of neighborhoods, and a list of issue categories. The user then selects a neighborhood, and selects a predefined issue type to report, and writes their comment. If another user selects the same location and issue type, comments are appended as additional comments. This gives an idea of number of users having same issue at a particular location. Survey submissions are anonymous.

Additionally, the date of the issue created, typical mode of user (walk, drive, bicycle), and intersection of the location selected are captured. There are 3788 records in our survey dataset with 8 useable columns. There are missing data and categorical data in incorrect format. These require EDA to clean-up and update missing information. Other than the survey data, we have various sources of data which would be combined to form the dataset for the project. All acquired from the City of Cincinnati are listed in Table 4. Supplementary data sets from external sources are listed in Table 5.

Table 3. Survey data attributes

Attribute	Description
REQUESTID	Survey key
REQUESTTYP	Provides categories for issue
REQUESTDAT	Date on which the issue was raised
COMMENTS	Issue description
USERTYPE	Mode of transport of the user
NEAR_INTER	Intersection nearest to location of issues
NEAR_STR	Street of issue
STRSEGID	ID specific to city of Cincinnati
Additional_COMMENTS	Additional descriptions for the same issue by different users
SNA_NAME	Neighborhood name

Table 4. Cincinnati based data sets

Data Set	Source	Evaluation summary
Census and demographic	https://www.cincinnati-oh.gov/planning/reports-data/census-demographics	Contains census data for each neighborhood, split into cencus tracts Data is organized at the neighborhood level, not street level as our final dataset is
Cincinnati open data	https://data.cincinnati-oh.gov	Contains economic, neighborhood safety, and health related data for city of Cincinnati Data is not granular
Cincinnati pedestrian safety survey data		Contains survey data from citizens who have reported problems by using the web-page Data is organized at the street intersection level Data was collected from Feb - Apr 2018
Income and house price	http://www.city-data.com/nbmaps/neigh-Cincinnati-Ohio.html	Contains statistics on age, house prices, income and more Data gathered as a collection of public and private sources Data organized at both the neightborhood, and census block group level
Traffic crash reports	https://data.cincinnati-oh.gov/Safer-Streets/Traffic-Crash-Reports-CPD-/rvmt-pkmq	Crash details at intersection is available, with details on person ages and injury level Provided by the Cincinnati Police Department

Table 5. Supplementary data sets

Data Set	Source	Evaluation summary
Google Maps	https://developers.google.com/maps	This API gives us granular data on walking and biking - distance, direction and other information Latitude and Longitude information for grid areas can be derived using this API
OpenStreetMap	https://www.openstreetmap.org	Similar to Google maps; we are researching ways to implement this into our study
Strava		Strava is a website and mobile application used by runners and bikers to upload their GPS-tracked activity to the Strava website Data can be used to create a heatmap of high activity areas
Walk Score®	https://www.walkscore.com	A scoring scale across US which gives an idea of the current walkability of the city
Zillow	https://www.quandl.com/data/ZILLOW/M26_-NFS-Zillow- Home-Value-Index-Metro-NF-Sales-Cincinnati-OHh	The Zillow Home Value Index contains monthly time-series of data which represent Zillow's estimation of the median market value of home sales.

4 Methods and Experiments

Analysis approach

For this project, our approach consisted of two parts. Part 1 involved extensive EDA, using unsupervised techniques to identify clusters and trends within the data. Part 2 consisted of adding cost of injury as in Table 1 to our dataset to perform regression. The dataset was simplified using correlation metrics. Feature importance of each regression method was tabulated. This approach aiedd in quantifying the intensity of accidents and predicting how expensive accidents might be. Models with high R^2 and low RMSE are considered good models. Additional algorithms were used in conjunction with above 2 approaches. T-SNe was used as a visualization tool to view clustering in 2-dimensional space, and PCA was used to reduce the dimensionality of the data set for regression. Furthermore, natural language processing was used to extract sentiment from the survey data comment field. The dataset was shuffle split into 80:20, 20% holdout for cross-validation. 80% will be split into 80:20 for training and testing. Machine learning and deep learning algorithms were used as listed in Table 6. Models were compared with each other on different performance metrics.

Table 6. Algorithms and evaluation methods

Algorithm	Models
Regression	1. Multiple Regression
	2. Support Vector Machine
	3. Decision Tree
Unsupervised Learning	1. Clustering
	a. K-means
	b. Hierarchical
	2. Gaussian Mixture Models
	3. Neural Networks
	a. Self Organizing Maps
Additional	1. T-Distributed Stochastic Neighbor Embedding
	2. Principal Component Analysis
	3. Linear Discriminate Analysis
	4. Natural Language Processing

5 Results

6 Analysis

7 Ethics

Table 7. ACM Code of Ethics

	section	YnNA	Comment
1,0	GENERAL ETHICAL PRINCIPLES.		
1,1	Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.		
1,2	Avoid harm.		
1,3	Be honest and trustworthy.		
1,4	Be fair and take action not to discriminate.		
1,5	Respect the work required to produce new ideas, inventions, creative works, and computing artifacts.		
1,6	Respect privacy.		
1,7	Honor confidentiality.		
2,0	PROFESSIONAL RESPONSIBILITIES.		
2,1	Strive to achieve high quality in both the processes and products of professional work.		
2,2	Maintain high standards of professional competence, conduct, and ethical practice.		
2,3	Know and respect existing rules pertaining to professional work.		
2,4	Accept and provide appropriate professional review.		
2,5	Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks.		
2,6	Perform work only in areas of competence.		
2,7	Foster public awareness and understanding of computing, related technologies, and their consequences.		
2,8	Access computing and communication resources only when authorized or when compelled by the public good.		
2,9	Design and implement systems that are robustly and usably secure.		
3,0	PROFESSIONAL LEADERSHIP PRINCIPLES.		
3,1	Ensure that the public good is the central concern during all professional computing work.		
3,2	Articulate, encourage acceptance of, and evaluate fulfillment of social responsibilities by members of the organization or group.		
3,3	Manage personnel and resources to enhance the quality of working life.		
3,4	Articulate, apply, and support policies and processes that reflect the principles of the Code.		
3,5	Create opportunities for members of the organization or group to grow as professionals.		
3,6	Use care when modifying or retiring systems.		
3,7	Recognize and take special care of systems that become integrated into the infrastructure of society.		
4,0	COMPLIANCE WITH THE CODE.		
4,1	Uphold, promote, and respect the principles of the Code.		
4,2	Treat violations of the Code as inconsistent with membership in the ACM.		

8 Conclusions (and Future Work)