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 [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]. 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 [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."

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 [13], 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 [14] 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 [15].
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). [16]
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.[17]
Hotspot	In this context, hotspots are areas with higher density or frequency of pedestrian related accidents [18].
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. [18]
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) [19].

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 [20] [21] [22] [23] [24] [25]. The study done by Guo [26], et al examined the patterning and structure of road networks as a factor of pedestrian vehicle interactions (PVIs). Zhang et al [27] 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 [13]. Users login to http://cagisonline.hamilton-co.org/pedsafetysurvey/ . The survey screen provides users with view of the city, with a drop down 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 usable 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.

One of the primary supplementary datasets used in analysis was acquired from Strava, a social media platform for athletes which allows them to upload their GPS-recorded workouts for friends and followers to see. Strava provided data of pedestrian workouts in the city of Cincinnati, which include latitude and longitudinal data of members' public workouts. This dataset was gathered through the Strava API, and Strava Metro service.

City's WalkScore and TransitScore dataset was created using the WalkScore API, a website that provides a scoring system. WalkScore ia an index ranging from 0 to 100, higher the score for a location, efficient the location is for walking. Transit Score is an index ranging from 0 to 100 which gives the accessibility of a public transport at a given location. Additionnally, We are using google maps API to provide us with accurate latitude and longitude for locations.

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 census 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 neighborhood, 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.openstreet–map.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. Geolocation based visual representation are used to derive location based patterns. Data like near miss, public complains, accidents plotted on map of Cincinnati provides compelling evidence of issues in city.We have utilized google API to extract latitude and longitude at street level for city of Cincinnati. Part 2 consisted of adding cost of injury as in Table 1 to our dataset to perform regression. Our dataset consist of accidents with injury, accidents with no injury and near misses. The dataset was simplified using correlation metrics. Feature importance of each regression method was tabulated. This approach aided 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% hold-out for cross-validation. We have done 10-fold cross-validation using stratified shuffle split from the python scikit library. 80% was further 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. For this project, our approach consisted of two parts. Part 1 involved extensive exploratory data analysis, 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 so that only significant variables would be used in modeling. This approach aided in quantifying the intensity of accidents and predicting how expensive accidents might be. Models with high R^2, a measurement of how well data fits to the modelled regression line, and low root mean square error, a measure of the difference between the modelled regression line and the data, were considered good models. Additional algorithms were used in conjunction with above 2 approaches. T-distributed stochastic neighbor embedding (t-SNE) was used as a visualization tool to view clustering in 2-dimensional space, and principal component analysis (PCA) was used to reduce the dimensionality of the data set for regression, resulting in fewer variables within the dataset. 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. The 80% was then split into a further 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, such as the area under curve and brier score.

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

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.

8 Conclusions (and Future Work)

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. walkscore.com: 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. Department of Transportation and Engineering: City launching pedestrian safety survey for all 52 neighborhoods. City of Cincinnati (2018)
14. National Safety Council: Estimating the costs of unintentional injuries, 2012. United States Census Bureau, http://www.nsc.org/NSCDocuments_Corporate/Estimating-the-Costs-of-Unintentional-Injuries-2014.pdf (2012)
15. U.S. Census Bureau: Geographic terms and concepts – census tract. United States Census Bureau (nd)
16. Ohio Department of Transportation: Safety study guidelines. Office of Systems Planning and Program Management (2017)
17. U.S. Department of Transportation: Vmt per capita. Transportation.gov (2016)
18. 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
19. Gaymard, S., Tiplica, T.: Conditionality and risk for the pedestrian: modelling with the bayesian networks. International journal of injury control and safety promotion **22**(4) (2015) 340–351

20. Brüde, U., Larsson, J.: Models for predicting accidents at junctions where pedestrians and cyclists are involved. how well do they fit? (1993)
21. 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
22. Lyon, C., Persaud, B.: Pedestrian collision prediction models for urban intersections. Transportation Research Record: Journal of the Transportation Research Board (1818) (2002) 102–107
23. 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
24. Pulugurtha, S.S., Sambhara, V.R.: Pedestrian crash estimation models for signalized intersections. Accident Analysis & Prevention **43**(1) (2011) 439–446
25. 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
26. 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
27. 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