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. To develop a solution to this problem of identifying likely locations of events, we assemble 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 augment that data with sports-tracking geolocation movement data of pedestrians, runners, and cyclists. From this assembled data set we complete unsupervised learning using a self–organizing map, excluding the event data. The event data is subsequently 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 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 the 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 [14] as shown in Table 1.

Table 1. Event Cost Severity, 201X NSC NEEDS CITATION!!!: https://www.nsc.org/Portals/0/Documents/NSCDocuments_Corporate/estimating-costs.pdf

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

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 [17] [18] [19] [20] [21] [22]. The study done by Guo [23], et al examined the patterning and structure of road networks as a factor of pedestrian vehicle interactions (PVIs). Zhang et al [24] 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

The Cincinnati Area Geographic Information System is an online survey which was available from Feb 2018 to April 2018 for pedestrian safety [13]. Users logged in to http://cagisonline.hamilton-co.org/pedsafetysurvey/ . 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 a pre-defined 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 which require clean-up and updating of 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 3. Datasets that originated from the Open Data Cincinnati are listed in Table 4. Supplementary data sets from external sources are listed in Table 5.

Table 3. Cincinnati based data sets

Data Set	Source	Evaluation summary
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
Pedestrian near miss data		A dataset provided by the city of Cincinnati containing latitude and longitudinal data of near-miss incidents within the city
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

Table 4. Cincinnati open data datasets

Data Set	Evaluation Summary
Cincinnati 311 non-emergency service requests	This dataset contains records of all non-emergency service requests to the city of Cincinnati from 2012 to 2018. Contains reports from graffiti to relevant traffic information such as roadkill and potholes.
2017 NFIRS Cincinnati Fire Department Incident Data	National Fire Incident Reporting System dataset which reports on the full range of fire deparment activities with reports of number of employees and vehicles deployed to every incident.
Street Centerline (w/ PCI rating)	City street centerline data with pavement condition index ranking; used to determine road condition and composition.
Traffic Crash Reports (CPD)	The Cincinnati traffic crash reports dataset compiled by the Cincinnati Police Department. It contains records of responses to traffic crashes, containing a wide-range of crash data such as persons involved, injury level, day, time, and weather.

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
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

4.1 Data Ingestion and Grid Aggregation

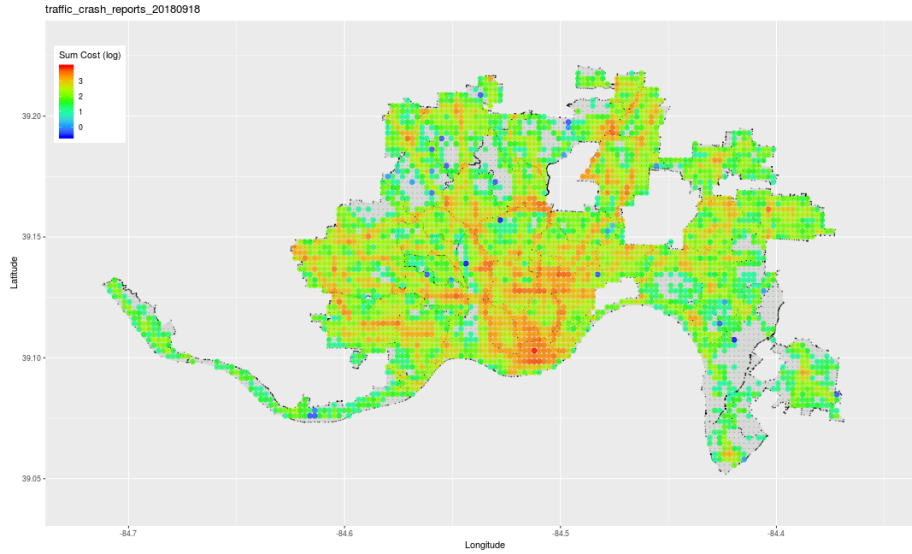Initial results from part 1 of our methodology have produced graphs as shown in Figure 1.

Fig. 1: Sum cost of traffic crashes reported from 2012 to October 2018 involving pedestrians by grid cells.

4.2 Random Forest - Binary Model

4.3 Multi-Variate Linear Regression - Cost Model

In order to create a linear cost model to predict the cost of a PVI, the data must lend itself to the assumptions of a parametric model. These assumptions include normality, linearity, constant variance, and independence. The response variable within the data has a very abnormal distribution, with nearly 80% of the reported incidents of PVI resulting in a 0 cost event. This positively skewed data does not meet the assumption of normality even after multiple types of transformations, notably including the logarithmic transformation. In order to meet the assumption of normally distributed data, the data is subset to exclude 0 cost events, and the few events resulting in extreme damage of $5 million or more. This data subset dramatically reduces the skewness present in the dataset, and allows for linear regression modelling to proceed.

The reduced scope dataset is first run through the 10-fold cross validated stepwise selection of the caret package in R, in order to determine the best full model, ranging from 10 to 35 predictor variables. The function iterates through each number of predictor variables from 10 to 35, performing stepwise 10-fold cross validation for each number of maximum predictor variables. The 10-fold cross validation is a model validation technique used to reduce overfitting and promote the generalizability of the model. Model overfitting 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 10-fold cross validation splits the data into 10 pieces. In each of the 10 iterations, one is selected to be the testing dataset, and the other 9 are used to train the model. This methodology is repeated so that each of the 10 pieces has a chance to be the testing dataset. The result of these 10 trials are averaged and compared against the cross validation results at each other maximum variable value from 10 to 35. The full model chosen is the model with the lowest average root mean square error value, and in our study, was determined to be the model with 30 variables.

Once the full model is selected, it is then manually pruned based on the variance inflation factor values and significance. Variance inflation factor values larger than 5 signify collinearity between variables, and a violation of one of the parametric assumptions in linear regression. Variable significance is determined by the p-value, and the typical value used is .05. Variables at or below the .05 p-value threshold are determined to be statistically significant to the model, while those above it are not. In spite of this black and white view on p-values, some variables are left in the model which are not significant at the .05 level, but are suggestive of being significant influences to the model, as shown in PLACEHOLDER.

With the reduced model, the cost of injury to pedestrians is predicted for each event, and the resulting prediction is compared against the actual cost of incident. The difference between the predicted and actual cost of injury is the residual, and the residuals are then mapped back to the Cincinnati map to visualize areas with potential for safety improvement.

4.4 Non-Supervised Learning - Neighborhood Characterization

5 Results

5.1 Data Ingestion and Grid Aggregation

5.2 Random Forest - Binary Model

5.3 Multi-Variate Linear Regression - Cost Model

5.4 Non-Supervised Learning - Neighborhood Characterization

6 Analysis

6.1 Data Ingestion and Grid Aggregation

6.2 Random Forest - Binary Model

6.3 Multi-Variate Linear Regression - Cost Model

6.4 Non-Supervised Learning - Neighborhood Characterization

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. Ohio Department of Transportation: Safety study guidelines. Office of Systems Planning and Program Management (2017)
16. 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
17. Brüde, U., Larsson, J.: Models for predicting accidents at junctions where pedestrians and cyclists are involved. how well do they fit? (1993)
18. 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
19. Lyon, C., Persaud, B.: Pedestrian collision prediction models for urban intersections. Transportation Research Record: Journal of the Transportation Research Board (1818) (2002) 102–107

20. 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

21. Pulugurtha, S.S., Sambhara, V.R.: Pedestrian crash estimation models for signalized intersections. Accident Analysis & Prevention **43**(1) (2011) 439–446

22. 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

23. 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

24. 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