A Framework to Assess Pedestrian Exposure Using Personal Device Data

Grace Douglas[1], Linda Ng Boyle[1], Haena Kim[1], Anne Moudon[1], Steve Mooney[1], Brian Saelens[2], Beth Ebel[2]
[1] University of Washington, Seattle, WA
[2] Seattle Children's Hospital, Seattle, WA

Capturing pedestrian exposure is important to assess the likelihood of a pedestrian-vehicle crash. In this study, we show how data collected on pedestrians using personal electronic devices can provide insights on exposure. This paper presents a framework for capturing exposure using spatial pedestrian movements based on GPS coordinates collected from accelerometers, defined as walking bouts. The process includes extracting and cleaning the walking bouts and then merging other environmental factors. A zero-inflated negative binomial model is used to show how the data can be used to predict the likelihood of walking bouts at the intersection level. This information can be used by engineers, designers and planners in roadway designs to enhance pedestrian safety.

INTRODUCTION

Pedestrian exposure is an important metric that can accurately capture the likelihood of pedestrian-related crashes. Previous pedestrian exposure measurements can be grouped into three categories: area-based measurements, point/segment-based measurements, and trip-based measurements.

Area-based pedestrian exposure measures is represented in terms of aggregated densities and/or self-reported walking activity. Aggregated densities commonly used are population-, employment-, and resident-based metrics (Quistberg et al., 2017). However, area-based metrics cannot account for variability of pedestrian activities over time or distance (Mooney et al., 2016). Self-reported walking activity is also an area-based source that asks participants to record their walking distance, walking duration, and number of trips. Because they are self-reported, issues with sampling biases and reporting reliability exists (Ralph et al., 2020). *Point/Segment-based* measures or pedestrian counts are usually collected at specific points (i.e. intersection) or segments (i.e. one side of a city block)(Zegeer et al., 2001). This data is used to assess exposure at specific locations but does not provide insights on walking patterns before or after the sampled location (Greene-Roesel et al., 2007).

The limitations of area-based and point/segment based measures are their inability to capture changes in pedestrian walking patterns over time and within a trip. With advances in portable lightweight electronic devices, it has become easier to collect data on subtle differences in walking patterns within a trip. Such *trip-based* pedestrian exposure measures include space-time walking path estimation, crossing behavior, and other physical activity. Lam et al. (2014) used this space-time approach and a shortest path algorithm to define a pedestrian path and uses the product of crash frequency and walked distance to estimate pedestrian exposure. A disadvantage of data from these electronic devices is that the recorded information is often not usable as collected. The sensors often include noise, missing values, and displaced locations, which need to be cleaned.

Pedestrian exposure measures should also be scalable for the research question of interest. For example, when comparing pedestrian-involved crash rates from state to state, each states' population may be sufficient as an exposure measure. However, within each state, the population metric will need to consider road, traffic, and environmental factors for each crash.

The goal of this paper is to provide a framework for capturing a more robust pedestrian exposure measure that can be scalable to address various pedestrian-related research questions. Pedestrian walk patterns were assessed using walking bouts, which captures the path counts and walking times at various location, points on the roadway, and time. The framework is presented at the intersection-level; which have a large number of vehicle-pedestrian interaction when compared to other road segments (Quistberg et al., 2015). That said, the walking bouts can also be aggregated to other levels (i.e. by individual, by group of people, by day, by spatial unit).

METHOD

Figure 1 presents the framework used to capture pedestrian exposure at the intersection level. The data processing was largely conducted using the statistical software program R (version 3.2.2), with the simple features ('sf') package. This ensured the file was accessible in R for data analysis. However, because the filtering of walking bouts and environmental predictors can be time consuming in R, the open source geographic information system, QGIS, was also used to validate the outcomes. Each dataset (pedestrian, intersection, environment) was also scaled appropriately before merging together.

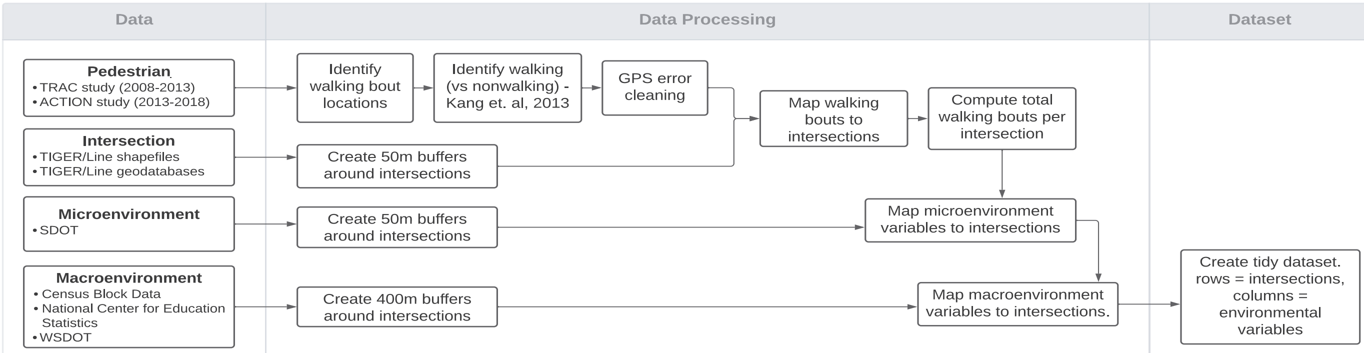

Figure 1: Data Processing of GPS-derived walking bouts and mapping of environmental factors

Data

Pedestrian Data

Pedestrian data were collected as part of two previous National Institute of Health (NIH) projects designed to examine the impact of transit and transit changes on walking and physical activity (B. Kang et al., 2013). Both projects used adult participants living in Seattle and the surrounding areas. Participants were asked to wear electronic devices that track their movements and to complete activity diaries that detailed their travel behavior each week. Data from the electronic devices captured walking bouts, defined as continuous walking with breaks less than or equal to two minutes within a seven minute rolling window (B. Kang et al., 2013).

- *Travel Assessment and Community (TRAC) Study*: Walking bouts were collected in three waves (2008-2009, n=707; 2010-2011, n=599, 2012-2013, n=525). The original study goal was designed to assess walking activities given the implementation of light rail in King County, WA (B. Kang et al., 2017).

- *Assessing Choices in Transportation in our Neighborhood (ACTION) Study*: Walking bouts were collected in three waves (2013-2014, n=590; 2015-2016, n=398; 2017-2018, n=382). The original study goal was designed to assess walking activities given the implementation of two bus rapid transit lines in King County, WA (Saelens et al., 2003) .

Intersection Data

The intersection-level spatial data was obtained from the Census Bureau's TIGER/LINE shapefiles and geodatabases (US Census Bureau, 2021).

Micro- and Macro Environment Data

The environmental data was separated into the micro and macro level. The micro-environment includes details associated with each intersection's physical characteristics. This data describes the local characteristics of the intersection and can be provided by the city Department of Transportation (DOT). For this study, data was obtained by the Seattle DOT.

The macro-environment describes the population and environment where the intersections are located. This included information such as population density, job density, and median household income. There were three GIS-linked data sources used to describe the macro-environment. They included census block data, data from the Washington State DOT and the National Center of Education Statistics.

Data Processing

Correction of GPS errors

Prior to identifying the number of walking bouts per intersection, the GPS data need to be cleaned of any tracking errors. The GPS device used the participants' location (or points) at any given time to provide information on walking speed, direction, and duration. When GPS devices cannot reliably connect with satellite transmissions, the system will incorrectly identify the location. Some reasons for unreliable satellite contact include the presence of dense tree canopy, proximity to tall buildings, and a recent restart of the device.

The lost GPS points were filtered by comparing the GPS-measured mean speed values with distance/duration mean speed. The following steps were used to identify and correct periods where GPS location data were unreliable.

1. Compute a *derived* mean speed, ($ms_{derived}$), by dividing the lengths between walking bout points by the GPS-measured duration per walking bout, in minutes: ($ms_{derived} = \frac{distance}{duration}$).
2. Compare the $ms_{derived}$ with the GPS-measured mean speed (ms_{GPS}). ms_{GPS} is the walking speed per walking bout.
3. Calculate the difference between ms_{GPS} and $ms_{derived}$.
4. Set a filter threshold to exclude values (GPS pts) for which the mean speed differences exceeded the 99^{th} percentile.

Of the 54,347 walking bouts available, 8,379 (or 15.42%) were removed due to this process.

Number of Walking Bouts per Intersection

After the GPS errors were corrected, the walking bout count within a 50m buffered region of each intersection was computed. Figure 2 helps visualize how walking bout counts were computed within each intersection buffer.

Spatial data for each intersection was extracted from the Census Bureau's TIGER/LINE shapefiles and geodatabases. The buffers created around each intersection allowed GIS-linked environmental variables to be merged to the walking bouts at the intersection level. A 50 meter buffer around intersections was used to identify the number of walking bouts and specific micro-environmental characteristics (e.g, crosswalks: yes, no). Given the differences in scale of the macro-environmental (e.g., presence of park: yes,no) and spatial-temporal variables (e.g, weather), a 400 meter buffer around intersections was used to define these variables. These two buffer sizes are consistent with those defined by M. Kang et al. (2019).

Intersection buffers. As shown in the pop-out in Figure 2, there can be multiple walking bouts within an intersection's 50m buffer. To ensure that the same walking bouts are not counted multiple times, they are counted in one traveling direction only. In the example shown Figure 2, there are 7, not 14 walking bouts traversing through the intersection in the pop-out. One walking bout can also be captured in more than one buffer and therefore counted in more than one intersection. This is reasonable because pedestrians' walking paths often cross more than one intersection in a given bout. By counting a walking bout for each intersection buffer it crosses, our analysis can understand how pedestrian exposure changes by intersection and throughout the course of a walking bout. Lastly, Figure 2 highlights the importance of intersection buffering. Not all the walking bouts directly cross a point in the intersection, but these pedestrians are still at risk when it comes to potential crash exposure due to intersection proximity. By using a 50 m buffer, the possible walking (exposure) is now captured and can be addressed.

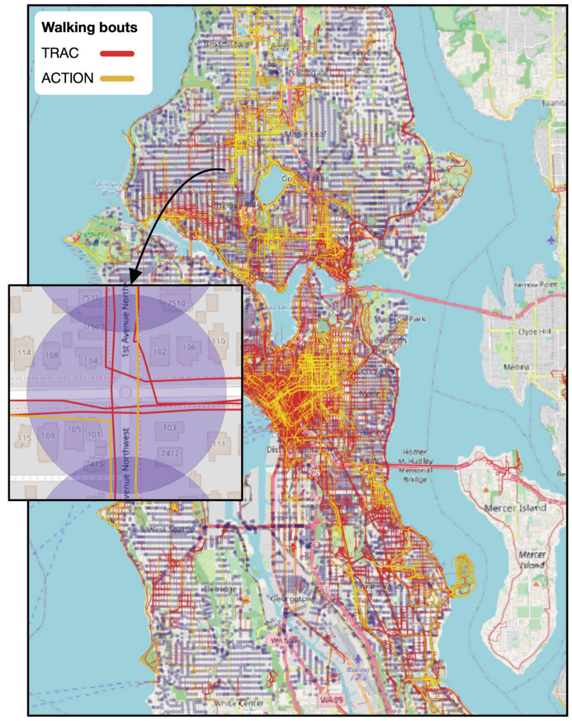

Figure 2: Intersections in Seattle, WA overlaid with 50m buffers. A subsetted visual shows an individual intersection and its corresponding buffer and walking bouts.

Micro-environment buffer. Because the micro-environmental data provides information on the intersection (e.g, signage, sidewalk length), these variables can be aggregated to the 50m intersection buffer level.

Macro-environment buffer. These variables provide more general characteristics of the area (e.g., population, environment) and were not isolated to any intersection. Hence, they were characterized within a 400m buffer of the intersection. Spatial and temporal variables that provide insights on the geographical environment were also filtered at the 400m buffer level.

DATA ANALYSIS AND RESULTS

The data was processed at the intersection level and as expected, not all intersections include pedestrians actually walking. Hence, there are many intersections with zero counts (see Table 1).

Table 1: No. of intersections with walking bouts (wb)

Data	Intersections	Int w/wb	%int w/wb
TRAC	112,341	12,635	11.24%
ACTION	110,988	15,704	14.15%

Depending on the research question, the analyst may consider segmenting the data further. For example, if the focus is only on intersections that have high pedestrian density, a model examining only intersections with pedestrian walking would be best, such as the Negative Binomial model. If the question relates to factors that impact the likelihood a pedestrians will walk at an intersection, then a zero inflated negative binomial model (ZINB) would be reasonable.

Table 2 shows the results from a ZINB model, which examines the likelihood of walking pedestrians in a given space at a given time. The ZINB model produces two outputs, which correspond to the count model results (follows a negative binomial (NB) distribution) and a zero-inflated (ZI) portion (follows a binomial distribution). The results are reported in terms of *Incidence Rate Ratio* (IRR), defined as the likelihood an incident occurs between the expressed and referenced groups.

The negative binomial or count portion (NB) shows there were on average 4.15 walking bouts per intersection for those intersections that had at least one walking bout. All environmental predictors considered were significant at a 99% confidence level. For example, the number of walking bouts is expected to be 2.22 times higher when there is a vehicle sign (includes stop and yield signs) present compared to when there is not. Because vehicle signs are often implemented to protect pedestrians, the presence of this sign could indicate the area has higher pedestrian density and is therefore indicative of walking bout count. Pedestrians may also feel safer walking where there is more vehicle regulation, which is important for city planners to know when implementing new roadway signage or pedestrian facilities.

Interestingly, the zero-inflated (ZI) portion did not show a significant impact for presence of vehicle sign or public school enrollment count. This means that while the presence of these variables is indicative of walking bout count per intersection, they do not significantly explain the likelihood of seeing any amount of walking bouts at an intersection. For example, public schools are often located in residential neighborhoods and may offer parks or other pedestrian facilities that increase the likelihood of pedestrians walking in the area. In this situation, there will be pedestrians around the school regardless of how the public school enrollment count. Hence, this variable is insignificant in the zero-inflated portion.

Other noteworthy findings include the impact of slope. Table 2 shows that a 1% increase in maximum slope percentage would decrease the expected walking bouts by a factor of 0.96. This result is consistent with the findings from Meeder et al. (2017), where a 1% increase in slope makes walking as much as 10% less attractive. The zero inflated (ZI) portion also shows that the *presence of park & ride*(ZI) increased the likelihood of walking bouts at intersections by 5.37 times.

Table 2: Zero-Inflated Negative Binomial model to predict number of walking bouts.

Predictors	Incidence Rate Ratios	CI	p
Count Model			
(Intercept)	4.15	3.79 – 4.53	< 0.0001
Maximum slope percentage(%)	0.96	0.95 – 0.97	< 0.0001
Presence of curve sign	0.85	0.76 – 0.96	0.007
Presence of vehicle sign	2.22	2.00 – 2.46	< 0.0001
Total sidewalk length(ft)	1.00	1.00 – 1.00	< 0.0001
Presence of Park & Ride	0.44	0.33 – 0.57	< 0.0001
Public school enrollment count	1.00	1.00 – 1.00	< 0.0001
Zero-Inflated Model			
(Intercept)	0.44	0.32 – 0.60	< 0.0001
Maximum slope percentage(%)	1.09	1.05 – 1.13	< 0.0001
Presence of curve sign	2.26	1.28 – 4.00	0.005
Presence of vehicle sign	0.76	0.32 – 1.79	0.530
Total sidewalk length(ft)	0.99	0.99 – 0.99	< 0.0001
Presence of Park & Ride	5.37	2.09 – 13.80	< 0.0001
Public school enrollment count	1.00	1.00 – 1.00	0.097
Observations	*14073*		
R^2 / R^2 *adjusted*	*0.964 / 0.964*		

DISCUSSION

This paper presented a framework for assessing pedestrian exposure using walking bouts. There have been previous studies using walking bouts (Quistberg et al., 2017) but the study was targeted toward pedestrian health but not necessarily safety. The framework presented in this study defined a process for examining walking bouts at the intersection level and provides the associated micro and macro environmental factors. That is, walking bouts have never been used to estimate pedestrian exposure to pedestrian-vehicle interactions, and hence pedestrian-vehicle crashes.

The data processing framework presented in this paper was aggregated to the intersection level, but given the level of data collection, it can also be scaled to a larger area-based estimate and discretized to lower levels to examine differences in sidewalk.

The feasibility of this framework was assessed using GPS-derived physical activity data from previously NIH-funded TRAC and ACTION projects. As an example, a zero-inflated negative binomial model was applied, which used the count of walking bouts within 50m buffered intersections as an outcome variable dependent on micro-, macro-, and spatial-temporal environment predictors. Some prominent predictors discussed included *presence of park & ride*, *presence of vehicle sign*, and *maximum slope percentage*.

Limitations. Data obtained through the TRAC and ACTION studies are biased towards participants who were willing to volunteer for these studies. This bias is considered in the pedestrian exposure model but adjusting for varying individual characteristics (i.e. age, gender, income, etc).

Future work. To date, there are few cities with information on pedestrian exposure at the intersection level. The proposed framework provides a strong foundation for identifying features that can impact pedestrian exposure in future studies. Our study provides insights on predictors at the micro-, macro-, and spatial-temporal environment predictors specific to Seattle, WA. It can be used to examine as well as validate the data from other cities in the US. In doing so, the model can continue to be updated and refined to increase model robustness.

ACKNOWLEDGEMENT

This project was sponsored by the US DOT - National Highway Traffic Administration (NHTSA), in collaboration with the Center for Child Health, Behavior and Development (Brian Saelens).

REFERENCES

Greene-Roesel, R., Diogenes, M. C., & Ragland, D. R. (2007, March). *Estimating pedestrian accident exposure: Protocol report* (Tech. Rep.).

Kang, B., Moudon, A. V., Hurvitz, P. M., Reichley, L., & Saelens, B. E. (2013, July). Walking objectively measured: classifying accelerometer data with GPS and travel diaries. *Med. Sci. Sports Exerc.*, *45*(7), 1419–1428.

Kang, B., Moudon, A. V., Hurvitz, P. M., & Saelens, B. E. (2017, December). Differences in behavior, time, location, and built environment between objectively measured utilitarian and recreational walking. *Transp. Res. Part D: Trans. Environ.*, *57*, 185–194.

Kang, M., Moudon, A. V., Kim, H., & Boyle, L. N. (2019, September). Intersections and Non-Intersections: A protocol for identifying pedestrian crash risk locations in GIS. *Int. J. Environ. Res. Public Health*, *16*(19).

Lam, W. W. Y., Yao, S., & Loo, B. P. Y. (2014, January). Pedestrian exposure measures: A time-space framework. *Travel Behaviour and Society*, *1*(1), 22–30.

Meeder, M., Aebi, T., & Weidmann, U. (2017, January). The influence of slope on walking activity and the pedestrian modal share. *Transportation Research Procedia*, *27*, 141–147.

Mooney, S. J., Sheehan, D. M., Zulaika, G., Rundle, A. G., McGill, K., Behrooz, M. R., & Lovasi, G. S. (2016, April). Quantifying distance overestimation from global positioning system in urban spaces. *Am. J. Public Health*, *106*(4), 651–653.

Quistberg, D. A., Howard, E. J., Ebel, B. E., Moudon, A. V., Saelens, B. E., Hurvitz, P. M., ... Rivara, F. P. (2015, November). Multilevel models for evaluating the risk of pedestrian–motor vehicle collisions at intersections and mid-blocks. *Accid. Anal. Prev.*, *84*, 99–111.

Quistberg, D. A., Howard, E. J., Hurvitz, P. M., Moudon, A. V., Ebel, B. E., Rivara, F. P., & Saelens, B. E. (2017, May). The relationship between objectively measured walking and risk of Pedestrian-Motor vehicle collision. *Am. J. Epidemiol.*, *185*(9), 810–821.

Ralph, K. M., Smart, M. J., Noland, R. B., Wang, S., & Cintron, L. (2020, October). Is it really too far? overestimating walk time and distance reduces walking. *Transp. Res. Part F Traffic Psychol. Behav.*, *74*, 522–535.

Saelens, B. E., Sallis, J. F., Black, J. B., & Chen, D. (2003, September). Neighborhood-based differences in physical activity: an environment scale evaluation. *Am. J. Public Health*, *93*(9), 1552–1558.

US Census Bureau. (2021). *Technical documentation: 2021 tiger/line shapefiles technical documentation* (Report). US Census Bureau.

Zegeer, C. V., Richard Stewart, J., Huang, H., & Lagerwey, P. (2001, January). Safety effects of marked versus unmarked crosswalks at uncontrolled locations: Analysis of pedestrian crashes in 30 cities. *Transp. Res. Rec.*, *1773*(1), 56–68.