The Built Environment, Travel Behaviors, and Active Transportation Safety

Min Gyu Kang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Anne Vernez Moudon, Chair

Brian Ernest Saelens

Qing Shen

Program Authorized to Offer Degree:

Urban Design and Planning

University of Washington

Abstract

The Built Environment, Travel Behaviors, and Active Transportation Safety

Min Gyu Kang

Chair of the Supervisory Committee:
Professor Anne Vernez Moudon
Urban Design and Planning

The goals of active and safe transportation can be achieved by creating a pedestrian-friendly environment. Since walking trips are more likely to be observed in dense urban areas where motorized travel is congested, a safe environment from motorized vehicles is crucial to protecting pedestrians and promoting walking. Thus, identifying locations where pedestrians are most vulnerable is important to further promoting this environmentally friendly and healthy mode of travel. The characteristics of the built environment at these locations help capture attributes that can affect the risk of crashes: for example, development densities and some related land uses attract pedestrian travel, while sidewalks and traffic signals can protect them from colliding with vehicles. As a result, quantifying built environment attributes around a crash-risk location is an important component in modeling pedestrian crashes. A variety of data and methods have been used to identify crash-risk locations in different studies, which have limited comparability across studies and caused complications in interpreting the results.

To date, most studies measured the overall characteristics of environments around potential crash locations for pedestrian crash modeling. However, measuring the built environment along an actual pedestrian route can more precisely capture the characteristics related to the risk of a crash than those derived from location-based approaches. Objectively measured mobility data coming from such devices as global positioning system (GPS) and

accelerometry have the potential to overcome limitations in location-based analyses. Processing the massive GPS and accelerometer datasets to reconstruct mobility patterns in terms of trips and travel modes requires robust computational power and sophisticated algorithms. Few studies have focused on understanding the details on how these algorithms process data for the purposes of quantifying travel behaviors.

In this dissertation, analyses first used a location-based pedestrian safety approach that combined built environment and crash data to identify crash-risk locations and to model pedestrian-motor-vehicle collisions. More specifically, a new protocol was developed to provide a useful tool for identifying unique pedestrian crash-risk locations at intersection and non-intersection areas. Second, the factors affecting pedestrian crashes were evaluated using fine-grained built environment data. Lastly, the automated travel behavior detection algorithm PALMS (Personal Activity Location Measurement System) was assessed: PALMS approach to translate objective measures of individual mobility patterns (e.g., GPS and accelerometry) into trips and travel modes was compared to trips recorded in travel diaries.

Studies in this dissertation contribute to the creation of consistent spatial analysis units for location-based pedestrian crash models, which makes it possible for empirical results to be comparable. A cost-effective method is offered to identify unique crash-risk locations. The dissertation also contributes to the literature by showing that factors affecting pedestrian crashes at intersection locations differed from those of non-intersection locations. It provides advanced visualization approaches to interpret empirical model results which can be used to prioritize safety countermeasures according to the characteristics of potential crash locations. Also investigated is the potential of device-collected mobility data from GPS and accelerometers to help identify individual travel modes, and to detect travel routes that can be used for quantifying the built environment attributes. The PALMS algorithm was found to better classify vehicular than pedestrian travel. Lastly, the methods developed to assess PALMS can be generalized and can serve to evaluate different approaches to travel mode classification.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the help of many people. My deepest thanks go to Professor Anne Vernez Moudon who provided the wonderful mentorship throughout my PhD life. Her academic passion, insight, and advice greatly inspired my research interests. To the rest of my dissertation committee – Professors Brian E. Saelens, and Qing Shen – I am grateful for their generous sharing of ideas, expertise, and time. I also wish to thank Professor Linda Ng Boyle for her expertise in transportation. Professor Philip M. Hurvitz deserves a special mention for his support and technical expertise. His help was essential in completing my dissertation. For both technical and personal support, I wish to thank my colleagues and friends in the UW Urban Form Lab. Special thanks go to Alon Bassok, Jason Scully, Orion Stewart, Eric Howard, Xiao Shi, and Yefu Chen. In addition, I'd like to thank the Washington State Transportation Center (TRAC) for offering great research infrastructures. Financial support was generously provided by the University of Washington Graduate School, the UW Urban Form Lab, the UW Urban Ecology Research Lab, the Washington State Department of Transportation (No. T1461), the Pacific Northwest Transportation Consortium (No. 62-9153), and the National Institutes of Health (R01 HL091881).

DEDICATION

This dissertation is dedicated to my parents, my sister, and people who have shaped me.

Chapter 1 Introduction

Promotion of active transportation is an important goal of transportation planning and public health [1,2]. Traveling by non-motorized travel modes provides clean, economic, and healthy travel options which have increased in popularity in recent years [3]. The National Household Travel Survey (NHTS) data show that in the US 9.7% of all person-trips were made by non-motorized travel modes such as walking and bicycling in 2009, compared to 6.3% in 1995 [4]. Previous studies have shown that people who walk for transportation are more likely to meet public health recommendations by accumulating more physical activity (PA) [5,6]. Systematic reviews also have shown that walking has favorable effects on many health and fitness indicators, including the risks of obesity, hypertension, colon cancer, depression, cardiovascular disease, and overall mortality [7,8]. According to the Physical Activity Guidelines for Americans (2018), sufficient PA is defined as at least 150 minutes to 300 minutes a week of moderate-intensity, or 75 minutes to 150 minutes a week of vigorous-intensity aerobic PA, or an equivalent combination of moderate and vigorous intensity aerobic PA for adults [9]. Walking is a good first activity to meet the recommendation because it does not require special skills or equipment and can be done indoors or outside by almost all age groups with little risk of injury [9–11].

There is a wide geographical variation in the use of active transportation, but walking trips are more important in urban areas [12,13]. Major obstacles to active transportation remain in low density sprawl areas that can generate long distance trips and where sidewalks are non-existent, and walking trips are more likely to be observed in dense urban areas where motorized travel is congested [12]. While walking is an effective way to obtain physical activity, pedestrians need to feel and to be safe in order to walk in these areas. Therefore, a safe environment is crucial to protecting pedestrians for the sake of promoting physical activity [14]. Studies conducted to identify factors related to pedestrian road trauma have addressed the characteristics of the micro-environment around crash-risk locations (e.g., roadway characteristics and traffic conditions) as well as the characteristics of the macro-environments around potential crash locations (e.g., neighborhood characteristics such as development densities and land uses that generate or attract pedestrian travel) [15–20]. Roadway design and pedestrian-friendly transportation facilities (e.g., wide sidewalks, traffic signals) have been found to reduce pedestrian crashes [21,22]. However, inconsistent results (regarding, for example, the association between vehicular traffic volumes, speed limits, etc., and crashes involving pedestrians) have been reported [15,16,23,24], suggesting the need for further investigations. Other built environment factors, which support safe pedestrian travel can also enhance the desirability of walking and may provide benefits of environmental health, obesity reduction, and injury prevention [14,25].

The characteristics of the micro and macro built environments can help define factors associated with crashes. As a result, quantifying built environment attributes is an important component in modeling pedestrian-motor vehicle collisions [14]. Many studies have used locations of potential crashes as the unit of analysis. These locations have been operationalized as points (e.g., intersections), polylines (e.g., roadways), polygons (e.g., census tracts), and grid cells (i.e., rasterized maps) [18,26–28] representing the spatial unit used to measure built environment characteristics. Inconsistencies in identifying crash-risk locations has limited comparability across studies and caused complications in interpreting the results. Consistent spatial analysis units for pedestrian crash models are required to make empirical analysis results comparable. In addition, spatial autocorrelation between crash-risk locations that are close together may bias analyses: overlapping buffers around locations will generate environmental characteristics of location *i* that will be associated with the outcomes of location *j* through the overlapping area [15,29,30], also suggesting the need for further investigation.

To date, most studies have measured the characteristics of locations where crashes may occur [15,31–34]. An alternative approach is to measure the characteristics of locations along an individual's walking path [14]. The built environment attributes measured along the pedestrian route (e.g., actual length of sidewalk taken by a pedestrian) can be more precisely defined and measured than those of location-based approaches: for example, the characteristics of sidewalks actually used by pedestrians can be measured as different from the characteristics of all sidewalks with a given buffered distance of a crash-risk location. The use of such devices as global positioning system (GPS) data loggers and accelerometers now provides information on the actual route or path of one person's travel. These data have the potential to overcome limitations of aggregate measures of the built environment around location-based analyses.

Device-collected data help capture locations and built environment attributes of where people actually walk. However, the translation of raw GPS and accelerometer data into travel behavior is a challenging task. Processing the massive GPS and accelerometer datasets to reconstruct mobility patterns in terms of trips and travel modes requires robust computational power and sophisticated algorithms [35]. Though these algorithms are increasingly used, limited effort has focused on understanding the details on how the data are processed to quantify travel behavior [36,37]. To obtain reliable data on individual trips and travel modes and develop the research methods in pedestrian crash modeling, it is required to assess travel behavior detection algorithms.

This dissertation addresses three specific research objectives:

1. Develop a systematic and replicable protocol to create a consistent spatial unit of analysis for use in location-based pedestrian crash modeling
2. Evaluate the factors affecting pedestrian-motor vehicle collisions at intersection and non-intersection locations using fine-grained built environment data
3. Evaluate an existing automated travel behavior detection algorithm for processing GPS and accelerometer data by comparing with travel diaries

Chapter 2 A Protocol for Defining Spatial Units for Crash-Risk Locations

---

Chapter Abstract

Background: Intersection and non-intersection locations are commonly used as spatial units of analysis for modeling pedestrian crashes. While both location types have been previously studied, comparing results is difficult given the different data and methods used to identify crash-risk locations. In this study, a systematic and replicable protocol was developed to create a consistent spatial unit of analysis for use in pedestrian crash modeling.

Methods: Four publicly accessible datasets were used to identify unique intersection and non-intersection locations: roadway intersection points, roadway lanes, legal speed limits, and pedestrian crash records. Two algorithms were developed and tested using five search radii (ranging from 20 to100 m) to assess the protocol reliability.

Results: The algorithms which were designed to identify crash-risk locations at intersection and non-intersection areas detected 87.2% of the pedestrian crash locations (*r: 20 m*). Agreement rates between algorithm results and the crash data were 94.1% for intersection and 98.0% for non-intersection locations. The buffer size of 20 m generally showed the highest performance in the analyses.

Conclusions: The present protocol offered an efficient and reliable method to create spatial analysis units for pedestrian crash modeling. It provides researchers a cost-effective method to identify unique intersection and non-intersection locations. Additional search radii should be tested in future studies to refine the capture of crash-risk locations.

Introduction

Pedestrian crashes have been a major concern globally, and especially in cities where pedestrian travel is concentrated. Given the prevalence of motor-vehicles, unprotected pedestrians are vulnerable and prone to experience serious injuries when colliding with motor-vehicles. In 2017, nearly 6,000 pedestrians were killed in motor vehicle crashes in the US [38]. While non-pedestrian fatalities decreased by 14% from 2007 to 2016, pedestrian fatalities increased at an alarming rate of 27%.

Studies have shown that the likelihood of pedestrian crashes are impacted by road and traffic (e.g. roadway width, traffic volume), the built environment (e.g. residential density, commercial land use) and socio-demographic characteristics (e.g. age, racial composition) [14,17–20,39]. Past studies have focused on modelling two types of outcomes: the severity of pedestrian injury and the frequency of pedestrian crashes. While the unit of analysis for injury severity models has been an individual pedestrian crash [39–42], crash frequency models have adopted location-based approaches. In frequency models, crash locations have been measured as points (e.g. intersection) [18,43], polylines (e.g. roadway segment) [26,44] , polygons (e.g. jurisdictional boundary) [27,45], and grid cells (e.g. rasterized map) [28,46].

Pedestrian crash-risk locations come as two main types: intersections and non-intersections [42]. An intersection is defined as the general area where two or more roadways meet [47]. Intersections are locations where most directional changes in travel take place, and consequently where conflicts between pedestrians and vehicles are high [18,48]. While intersection-specific engineering safety measures are used to mitigate these conflicts, intersection design standards often prioritize the operation of vehicles rather than the safety of walkers [17,49]. A non-intersection is any location within a roadway segment or along a transportation facility, that is not at an intersection. Past research has shown that factors (e.g., vehicle type, roadway curves) that impact collisions at intersections do not necessarily impact crashes at non-intersections [42].

The identification of crash-risk locations at intersection and non-intersection area is an important part of modeling pedestrian crashes. However, there is limited definitions, data or methods to appropriately identify crash-risk locations [50]. Regarding intersections, the definition of what constitutes an intersection might be similar in the literature [18,29,30,48,51]. However, the data and methods used to identify intersections vary across studies. For example, while some studies extracted intersection point data from nodes on intersecting roadway polylines [39], other studies conducted field investigation to obtain intersection locations [51,52]. More seriously, non-intersections have broader definitions that range from highway sections [26,53] to mid-block, cul-de-sac, curve [43], and even toll plazas [41].

There are also issues associated with the spatial analysis tools used for measuring features around crash-risk locations [54]. Complex spatial analysis using Geographic Information System (GIS) and advanced quantitative methods are often needed to measure the outcomes (e.g. the number of pedestrian crashes) and predictors (e.g. residential density) of pedestrian crash models [39,43]. Buffering techniques are widely used given the many transportation facilities that include GIS vectors [19]. A wide range of bandwidths have been used for buffering, but most are within a 100 m radius [18,43,55].

Regardless of buffer size, overlapping areas between buffers of different crash-risk locations is a major cause of spatial autocorrelation, which impacts the interpretation of pedestrian crash models [29,30,43]. The statistical models estimated in previous studies were based on the assumption that observations were mutually independent [56–58]. However, the statistical requirement that observations be independent and identically distributed (*i.i.d*) is often violated because of the overlapping buffers. Adjacent crash-risk locations are more likely to violate this assumption because they are more likely to have overlapping buffer areas.

Very few studies have accounted for the spatial autocorrelation in pedestrian safety studies. Mixed-effects models have been adopted in some studies to reflect contextual characteristics [29,30,43]. Although these models mitigate the effects of spatial autocorrelation by adopting advanced statistical methods, the source data for spatial analysis units might still involve a problem of autocorrelation derived from overlapping buffers. A different approach is the use of sampling to identify uncorrelated crash-risk locations. For instance, a subset of intersections that were considered to be independent through field investigation can be used in statistical analyses [51,52]. However, this requires extensive time and efforts, with a small sample size.

The objective of this study was to introduce a systematic and replicable protocol to create uncorrelated spatial units of analysis for pedestrian crash modeling for intersection and non-intersection areas. Although the modeling results from previous studies provide valuable insights, measurements of pedestrian crash-risk locations are often not consistent among research projects, in part due to differences in collecting and processing the source data [52]. This has led to complications in interpreting and comparing model results. A standardized method to identify pedestrian crash-risk locations would help improve the reliability, accuracy, and validity of locational factors. With a clear and replicable unit of analysis for pedestrian crash modeling, researchers and transportation planners could better understand the factors that influence the pedestrian crashes.

Methods

Data

<u>*Pedestrian crashes*</u>

Pedestrian crash data came from the Transportation Data, GIS & Modeling Office of WSDOT and covered the years between 2013 and 2017. The data included all crashes that had been reported to and recorded by local police or State Highway Patrol. There were 2,222 pedestrian crashes on state routes during the study period, with data including individual-level information such as time, weather, road condition, and socio-demographic and behavioral characteristics of both drivers and pedestrians. In the data, crash location came as milepost on state routes and county roads; and as distance from the closest intersection on city streets. Crash latitude and longitude were identified by WSDOT using Linear Referencing System (LRS) and geocoding tools in GIS.

The data included information as to whether pedestrian crashes occurred at intersections or non-intersections. The crash data was segmented into nine location types:

- Type 1: At driveway within major intersection;
- Type 2: At intersection and not related;
- Type 3: At intersection and related;
- Type 4: Circulating roundabout;
- Type 5: Exiting roundabout;
- Type 6: At driveway;
- Type 7: Driveway related but not at driveway;
- Type 8: Intersection related but not at intersection;
- Type 9: Not at intersection and not related.

For the forthcoming analysis, this information was re-categorized into two groups; intersection (type 1 to 5), and non-intersection (type 6 to 9). Among 2,222 state route pedestrian crashes, 1,423 (64%) occurred at intersection and 799 (36%) occurred at non-intersection. This information was used as a reference to test the performance of the algorithm for detecting unique intersection and non-intersection locations.

Roadway lanes and legal speed limits

Two transportation network datasets were obtained from the Office of Information Technology of the WSDOT: 1) roadway lane polyline data, 2) legal speed limit polyline data. First, roadway lane data included state routes, county roads, and city streets. The data included roadway width, number of increasing/decreasing lanes, and milepost information for state routes. Jurisdictional information such as city names was also included for county roads and city streets. There were 18,999 state route segments and 127,652 non-state route roadway segments used in this study. Second, legal roadway speed limit information was obtained as a separate dataset which contained 2,478 records. The data included speed limit information for each state route polyline segment.

Investigating intersection point data

Roadway intersection point data came from the Office of Information Technology of the Washington State Department of Transportation (WSDOT). In this dataset, intersections related to vehicular travel, and were derived not from road or street center lines, but from vehicular traffic lanes. Intersections were defined as any location where vehicular traffic could change travel direction. Different intersections were generated based on traffic direction (e.g. a left-turn lane from the north of an intersection had a different intersection with straight travel lanes than a left turn lane from the south of the same intersection) (NOTE: this definition is generated by the researchers based on their investigation of the data, and has not been confirmed by WSDOT—there is no meta data attached to the intersection dataset).

The data contained 26,204 records of intersections and provided intersection type information. Nine types were identified:

- Type G: an intersection where roadways are crossed at a common grade;
- Type A: a lane becoming an on ramp to a limited access road;
- Type E: an off-ramp lane to a limited access road;
- Type Y: a WYE (Y) connection where lanes formed three legs in the general form of a Y and the angle between two legs is less than 60 degrees;
- Type T: an entrance or an exit lane to a limited access road;
- Type N: an entrance lane to a limited access road;
- Type X: an exit lane from a limited access road; type
- Type R: roundabouts;
- Type O: on and off ramp lanes to limited access roads.

Many of the intersection types included in the data did not correspond to locations where pedestrians would cross streets or roads (Figure 1). Two trained GIS analysts investigated over 100 intersection data points and compared them with aerial photos and Google Maps to extract intersection data points where pedestrians could actually walk and cross a street or a road. Overall, they found that only intersection types G (grade intersection) and T (entrance and exit) corresponded to intersections that pedestrian would use. These intersection types were included in the pedestrian intersection data.

Several observations emerged for intersection type G, and T. Figure 1 shows intersection points identified in the data and pedestrian crash locations. In Figure 1a, Interstate-5 is a limited access highway which pedestrians are prohibited from using. Yet, the facility intersects with locations where pedestrians are allowed to cross. In Figure 1b, lanes from State Route 99, a limited access highway where pedestrians are prohibited to use, intersect with Denny Way, a city street that pedestrians can cross. The data identified five points where vehicular lanes intersect. Yet these intersection points are so close to each other that they are representing one pedestrian crossing location.

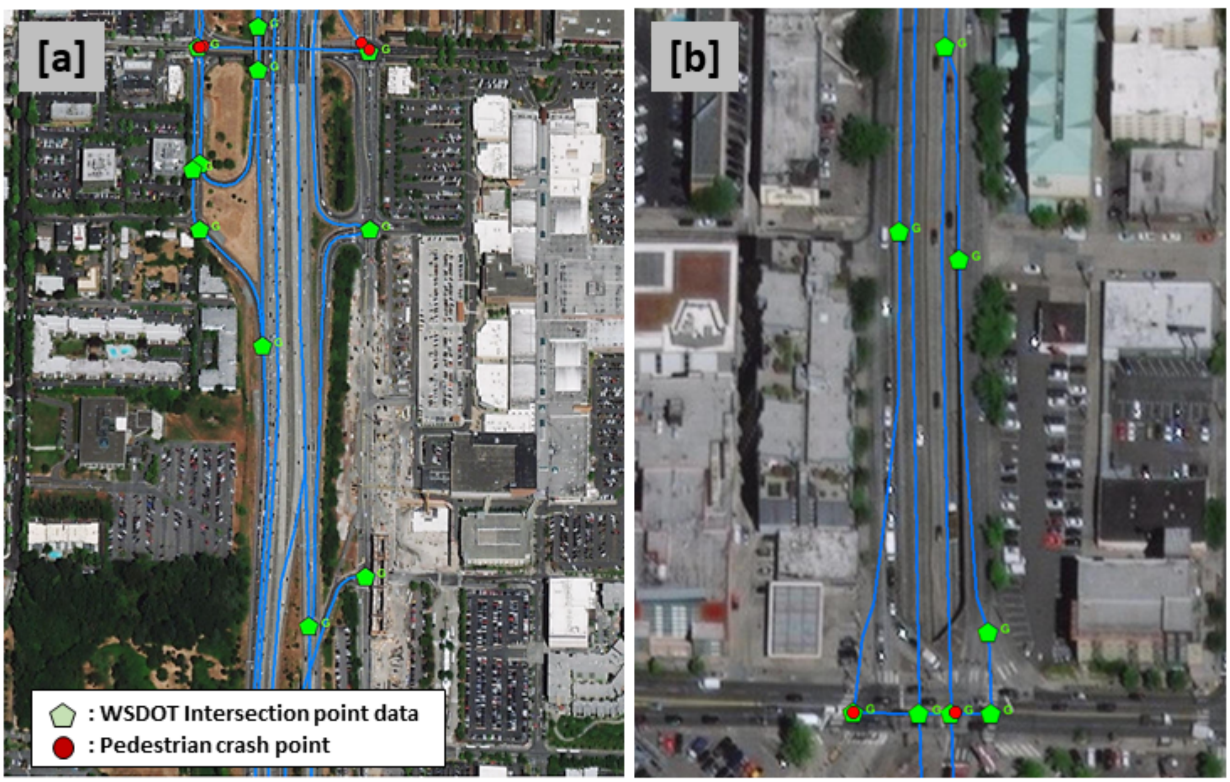


Figure 1 Examples of WSDOT data identified intersection points in Seattle, at [a] (Interstate-5 and NE Northgate Way) and [b] (state route 99 and Denny Way) in Seattle, Washington State. The blue lines represent the vehicular lanes in the respective facilities.

There were two issues identified from these aerial photos. First, in Figure 1a, the WSDOT intersection point data convey information on possible directional changes for vehicles that are not often on or near intersections used by pedestrians. In other words, the intersection point data are not restricted to streets or roads that can be crossed by pedestrians. Thus, the raw intersection GIS data might not be appropriate for modeling pedestrian crashes because some areas are not actually accessible by pedestrians. To model pedestrian crashes, intersection points on limited-access highways (e.g. interstate) needed to be removed before analysis [52]. However, some of these limited-access highways are located near local streets (e.g. city street), where pedestrians are allowed. Hence, a systematic protocol was needed to distinguish the pedestrian accessible intersection points from the inaccessible ones.

Second, as observed in Figure 1b, when multiple intersection points are in close proximity, intersection locations could be double-counted, leading to double-counting of pedestrian crashes. If crashes are allocated to only one intersection buffer, there will be cases and controls that have the similar locational attributes. Furthermore, intersections along a certain corridor will share similar roadway characteristics and land use. Also, adjacent intersections share similar traffic conditions, and therefore drivers' behaviors in those locations might also be alike. Thus, crash-risk locations in close spatial proximity are most likely correlated, leading to a biased model [30,48].

Decision tree algorithms

Two algorithms were developed and tested to detect unique crash-risk locations. Figure 2shows the steps used in the data reduction process. First, an algorithm was created to identify unique intersection locations. Figure 2a is workflow of the algorithm. Intersection point data from WSDOT were used as the input dataset for this process. Pedestrian accessible intersection points were extracted by using intersection type, road type, and legal roadway speed limit information. Figure 2b shows the decisions made for detecting unique non-intersection locations.

WSDOT pedestrian crash data were used as a baseline dataset. Buffering techniques were also used to identify non-intersection locations with crashes (cases of case-control model). Voronoi diagram techniques were applied to detect random non-intersection locations without crashes (controls of case-control model). Detailed description for each process is explained in the following sections.

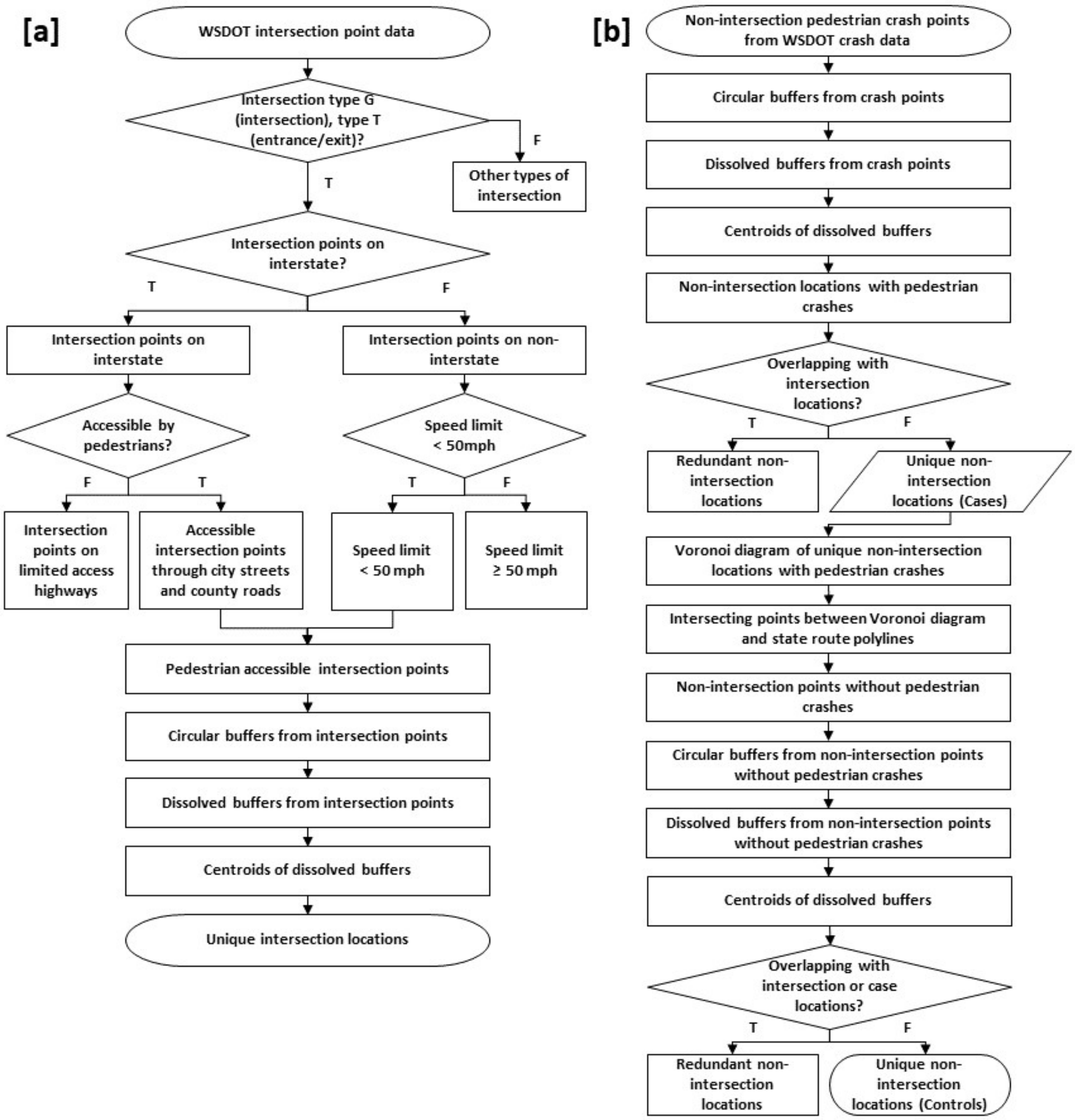


Figure 2 The decision tree algorithms show processes for detecting unique intersection locations [a] and non-intersection locations [b]

Intersection points on limited access roadways

From a legal perspective, intersection points on the main thoroughfares of limited-access roadways are not to be used by walkers for transportation purposes. However, in some cases, these intersection points were physically accessible by walkers at ramps and other locations where the highway connects to local streets such as county roads and city streets. And pedestrian crashes have occurred at some of these locations. To identify these locations, 10 m buffers from local streets were created and pedestrian accessible intersection points were extracted.

For the next step, intersection points on non-interstate routes were examined using a state route network dataset, which included legal speed limit information for each roadway segment. State route segments where the speed limits are greater than or equal to 50 mph were used to represent locations inaccessible to pedestrians. Intersections that were beyond 10 m from these segments were identified as being pedestrian-accessible. Pedestrian accessible intersection points on interstate and non-interstate route segments were then merged as a single GIS point layer for detecting unique intersection locations.

Detecting unique intersection locations

Creation of Euclidean buffers using state route intersections resulted in many overlapping buffers, leading to potential autocorrelation. Considering two intersections with overlapping buffers (i and j), environmental characteristics of intersection i will be associated with the outcomes of intersection j through the overlapping area between two intersection buffers. We applied buffering techniques described in Figure 3 to account for potential autocorrelation. Three steps were used to identify uncorrelated intersection locations without overlaps.

Pedestrian-accessible intersection points identified from previous analysis and street network data were used as an input dataset. Euclidean buffers were first developed from each intersection point to represent initial intersection locations. If there is an overlapping area between Euclidean buffers, a dissolved buffer was created to capture overlapping areas of polygons. A single centroid was then detected from each dissolved buffer and used as a unique intersection location. Lastly, Euclidean buffers were re-created from each point to represent unique intersection locations.

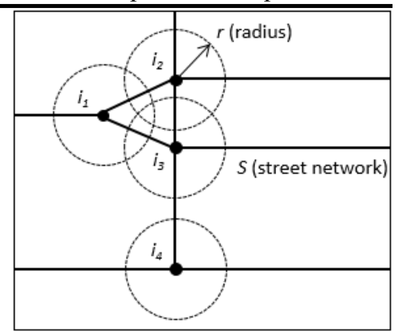
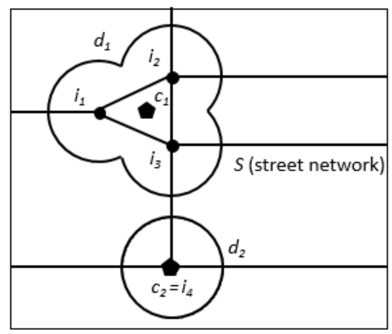
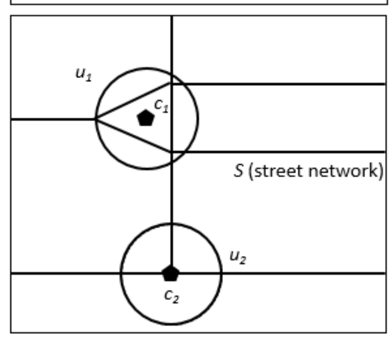
Step	*Description*	*Graphical Example*
1	• **Input**: Pedestrian-accessible intersection data points were used as an input. • **1a**: Euclidean buffers (radius: *r*) were created centered on each intersection point ($i_1, i_2, \ldots, i_n$).	
2	• **2a**: If there is an overlap between Euclidean buffers, a single dissolved buffer ($d_1, d_2, \ldots, d_n$) was created to represent overlapping area. • **2b**: A single centroid ($c_1, c_2, \ldots, c_n$) was extracted from each dissolved buffer and used as a unique intersection location.	
3	• **3a**: Euclidean buffers ($u_1, u_2, \ldots, u_n$; radius: *r*) were re-created based on each unique intersection location ($c_1, c_2, \ldots, c_n$).	

Figure 3 Processing intersection point data and street network for detecting unique intersection locations

Detecting unique non-intersection locations

According to the WSDOT pedestrian crash data, 36% of the pedestrian crashes on state routes occurred at non-intersections. To account for non-intersection pedestrian crashes, an algorithm was developed to detect non-intersection locations with or without pedestrian crashes.

Figure 4 shows the five steps used in the analysis. Non-intersection pedestrian crash points and state route network data were used as baseline datasets. Buffering techniques developed for identifying unique intersection locations were applied to detect unique non-intersection locations with pedestrian crashes (case observation). To identify non-intersection locations without pedestrian crashes (control observation), a Voronoi diagram was created based on case observation points. Since all Voronoi polygon boundaries are the farthest lines from the location of cases, the chance of overlaps between cases and controls is minimized. Lastly, to extract unique non-intersection locations removing overlapping area, same buffering techniques used in previous steps were repeated.

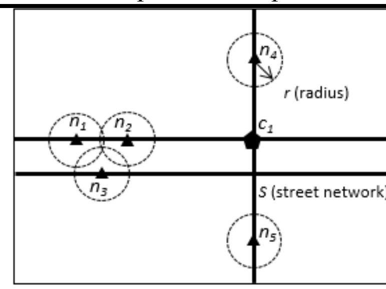
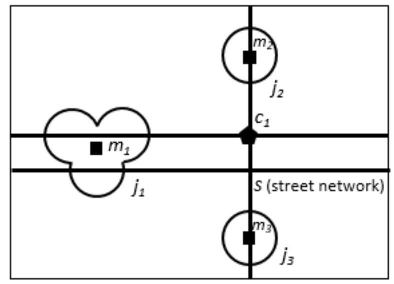
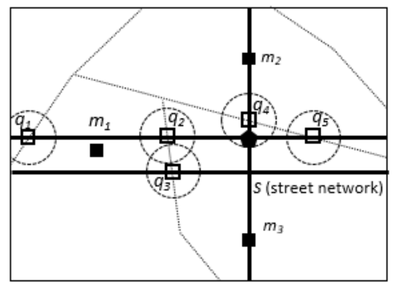
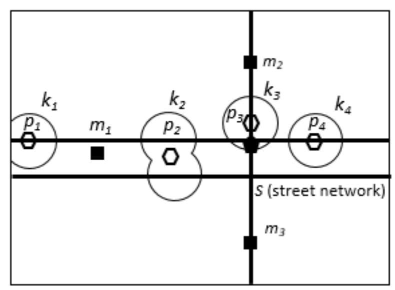
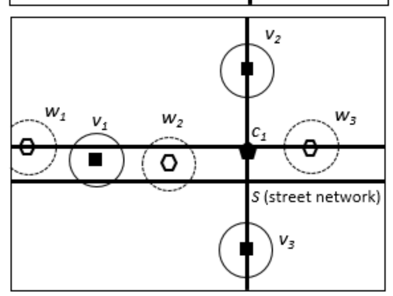
Step	*Description*	*Graphical Example*
1	• **Input**: Non-intersection pedestrian crash points ($n_1, n_2, \ldots, n_n$) and WSDOT state route network datasets were used as inputs. • **1a**: Euclidean buffers (radius: r) were created based on each non-intersection pedestrian crash point.	
2	• **2a**: If there is an overlap between Euclidean buffers, a single dissolved buffer ($j_1, j_2, \ldots, j_n$) was created to represent overlapping area. • **2b**: A single centroid ($m_1, m_2, \ldots, m_n$) was identified based on each dissolved buffer and used as a unique non-intersection location with pedestrian crashes (case observation). • **2c**: Some of unique non-intersection locations were removed if they are overlapping with intersection locations (c_n).	
3	• **3a**: A Voronoi diagram was created based on unique non-intersection location with pedestrian crashes (m_n). Any location within a Voronoi cell is closer to its associated point than to any other point input feature. • **3b**: Non-intersection points without pedestrian crashes were created on intersecting points between state route network and Voronoi polygon boundaries ($q_1, q_2, \ldots, q_n$). • **3c**: Euclidean buffers (radius: r) were created based on these non-intersection points.	
4	• **4a**: A single dissolved buffer ($k_1, k_2, \ldots, k_n$) was created to represent overlapping area. • **4b**: Centroids ($p_1, p_2, \ldots, p_n$) were detected based on dissolved buffers and used as a unique non-intersection location without pedestrian crashes (control observation).	
5	• **5a**: Euclidean buffers (radius: r) were re-created based on unique non-intersection location without crashes ($w_1, w_2, \ldots, w_n$). • **5b**: A part of non-intersection locations was eliminated if there is an overlap with intersection locations ($c_1, c_2, \ldots, c_n$) or case locations ($v_1, v_2, \ldots, v_n$).	

Figure 4 Processing pedestrian crash data and street network for detecting unique non-intersection locations

Parameter setting and comparison

The algorithm relies on Euclidean buffers with a defined search radius to identify unique intersection and non-intersection locations. Since the outcomes of the algorithm can be affected by this parameter, five performance runs were generated to test the agreement rates between the algorithm and the locational information from crash data using search radii of 20, 40, 60, 80, and 100 m.

Results

Unique intersection locations

Table 1 shows the number of unique intersection locations for case and control based on each search radius setting. The total number of unique intersection locations decreased and the relative proportion of case locations increased as a search radius increased. For example, using a 20 m radius produced 7,522 unique intersections with 10.6% having crashes, whereas a 100 m radius produced 3,019 unique intersections with 14.2% having crashes.

Table 1 Count of unique intersection locations for cases (having crashes) and controls (no crashes)

Parameter	20m	40m	60m	80m	100m
Case	794 (10.6%)	818 (12.3%)	596 (12.8%)	514 (13.8%)	428 (14.2%)
Control	6,728 (89.4%)	5,808 (87.7%)	4,056 (87.2%)	3,212 (86.2%)	2,591 (85.8%)
Total	7,522	6,626	4,652	3,726	3,019

Unique non-intersection locations

Results of the algorithm for non-intersection data were similar to results with intersection data. Table 2 shows the number of unique non-intersection locations for case and control by each parameter setting. The total number of unique non-intersection location decreased from 1,608 at

20 m to 955 at 100 m. The proportion of case (non-intersection location with state route pedestrian crashes) increased from 35.3% at 20 m to 38.7% at 100 m.

Table 2 Count of unique non-intersection locations for cases (having crashes) and controls (no crashes)

Parameter	20m	40m	60m	80m	100m
Case	567 (35.3%)	455 (38.0%)	419 (38.3%)	374 (37.9%)	370 (38.7%)
Control	1,041 (64.7%)	743 (62.0%)	674 (61.7%)	612 (62.1%)	585 (61.3%)
Total	1,608	1,198	1,093	986	955

Pedestrian crashes at crash-risk locations

Pedestrian crashes occurred at algorithm-identified intersection and non-intersection locations as well as at locations not identified as intersections or non-intersections. There were 2,222 state route pedestrian crashes in Washington State (2013-2017). The proportion of pedestrian crashes captured within intersection location buffers ranged between 35.7% (r: 100 m) and 61.4% (r: 40 m). The proportion slightly increased from 20 m to 40 m, then decreased sharply at and above 60 m. The proportion of pedestrian crashes detected by non-intersection location buffers was relatively stable compared to intersection locations. It was lowest at 40 m (22.8% of crashes) and highest at 20 m (27.6% of crashes). The proportion of pedestrian crashes occurring at locations not identified using the algorithm was lowest at 20 m (12.8% of crashes) and highest at 100 m (39.2% of crashes).

Table 3 Count of pedestrian crashes at intersection, non-intersection, and other locations

Parameter	20m	40m	60m	80m	100m
Within intersection location buffers	1,324 (59.6%)	1,364 (61.4%)	1,063 (47.8%)	928 (41.8%)	793 (35.7%)
Within non-intersection location buffers	614 (27.6%)	507 (22.8%)	530 (23.9%)	510 (23.0%)	558 (25.1%)
Others (not captured by crash-risk locations)	284 (12.8%)	351 (15.8%)	629 (28.3%)	784 (35.3%)	781 (39.2%)
Total	2,222	2,222	2,222	2,222	2,222

Agreement between crash-risk location and location type in crash data

Agreement rates between algorithm-identified locations and WSDOT-recorded crash location characteristics (intersection versus non-intersection) were computed to compare algorithm-generated and empirical data. Table 4 shows counts of pedestrian crashes within algorithm-identified intersection location buffers. A total of 1,324 pedestrian crashes occurred within 20 m buffer of intersection locations, of which 94.1% identified as intersection-related crashes from the crash data. The agreement rates decreased gradually with increasing buffer radius. The lowest agreement rate (79.2%) was found with 100 m intersection location buffers.

Table 4 Agreement rates between intersection location and crash data

Buffer Radius (r)	Count of pedestrian crashes within intersection location			
	All	Location type from crash data		Agreement
		Intersection	Non-Intersection	
20 m	1,324	1,246	78	94.1%
40 m	1,364	1,202	162	88.1%
60 m	1,063	907	156	85.3%
80 m	928	761	167	82.0%
100 m	793	628	165	79.2%

Table 5 shows concurrence between algorithm-identified non-intersection locations and WSDOT data for pedestrian crashes that were recorded as having occurred at non-intersection locations. Of the total of 614 pedestrian crashes within 20 m of non-intersection locations, 98.0% were categorized as non-intersection crashes from the crash data records. The agreement rates showed a sharp decrease after and beyond 60 m. The lowest agreement rate (77.8%) was found with the longest search radius (100 m).

Table 5 Agreement rates between non-intersection location and crash data

Buffer Radius (r)	Number of pedestrian crashes within non-intersection location			
	All	Location type from crash data		Agreement
		Intersection	Non-Intersection	
20 m	614	12	602	98.0%
40 m	507	11	496	97.8%
60 m	530	65	465	87.7%
80 m	510	86	424	83.1%
100 m	558	124	434	77.8%

Discussion

This study produced a systematic and reproducible protocol to identify unique pedestrian-motor-vehicle crash-risk locations at intersection and non-intersection areas. A unit of spatial analysis for pedestrian crash modeling was derived from two algorithms, and the reliability of the protocol was assessed by comparing the outcomes with the actual pedestrian crash data. A set of parameters was tested to check the sensitivity of the algorithm results.

Unique intersection locations were first identified using state route intersection points and street network GIS data. Since larger buffers create more overlapping areas between initial intersection locations, the total number of unique intersection locations decreased as the search radius of buffer increased. The shortest search radius (20 m) generated 7,522 unique intersection locations from 26,204 initial intersection points, which means that 3.5 initial intersection points on average were spatially correlated through overlapping areas. This problem can be mitigated by identifying unique intersection locations. Although the total number of unique intersection location was largest with 20 m buffer radius, the number of case observations was largest with 40 m search radius. Considering pedestrian crashes are rare events, a search radius should be decided based on the purpose of study not just based on the total number of observations.

Unlike intersection locations, pedestrian crash data were used to detect unique non-intersection locations. Since non-intersection locations cover much broader area than intersection locations, pedestrian crash data points occurred at non-intersections were used at the first step of algorithm to narrow down the possible candidates for unique non-intersection locations (case observation). The algorithm then adopted a Voronoi diagram to detect unique non-intersection locations without pedestrian crashes (control observation). Since any location within a Voronoi cell is closer to its associated point than to any other point input feature, newly detected control observations have the farthest distance from case observations. In this way, the possibility of overlaps between cases and controls was minimized. The number of unique non-intersection locations showed a sharp decrease between 20 m and 40 m of search radius, and the number remained almost steady after 40 m. The results imply that if a certain study requires a large number of non-intersection locations, then it would better to use a search radius shorter than 40 m, preferably 20 m.

The performance of protocol was assessed with over 2000 pedestrian crashes that occurred on state routes between 2013-2017. Intersection locations captured the largest number of pedestrian crashes with 40 m search radius (61.4% of the total of 2,222), showing a major decrease after applying 60 m search radius (47.8%). Non-intersection locations captured relatively steady number of pedestrian crashes (22.8 – 27.6%) compared to the intersection locations across five search radii. This is mainly because the algorithm for detecting non-intersection locations initially started from the location of empirical pedestrian crash points. Overall, the proportion of pedestrian crashes captured within algorithm-detected intersection and non-intersection locations was highest with 20 m search radius.

Although algorithm-detected intersection and non-intersection locations captured most of pedestrian crashes (87.2% of the total of 2,222 with 20 m search radius), there might be some misclassified crashes by the algorithm. For instance, a non-intersection pedestrian crash that occurred close to the intersection location can be classified as an intersection-related crash when using a large search radius. To assess the accuracy of the algorithm, agreement rates between algorithm-identified locations and WSDOT-recorded crash locations were computed. Although empirical pedestrian crash data are not a gold standard for evaluating the algorithm (e.g. human errors in recording and reporting pedestrian crash locations), they can still work as a useful reference to assess the protocol. The agreement rates between two measures was highest with 20 m buffer radius for intersection locations (94.1%), showing a gradual decrease as search radii increase. The agreement rates were stable for non-intersection locations between 20 m to 40 m, showing a rapid decline after 60 m.

Overall, the algorithms showed excellent performance in identifying two types of crash-risk locations. Especially, the protocol presented in this study has major benefits. First, the algorithms developed for identifying unique intersection and non-intersection locations can reduce human errors and labor hours to clean the data. There are 26,204 intersection points on state routes in Washington State. It will take a lot of time for researchers to manually classify uncorrelated intersections by comparing with satellite maps.

Second, the algorithm generates non-intersection locations with a clear definition and processes, allowing researchers to make comparisons between studies. There has been little consistency in identifying non-intersection locations in the literature, which made it difficult for empirical studies being comparable. By using the same protocol for creating non-intersection locations, empirical analysis results from pedestrian crash models on non-intersection locations can be easily compared.

Third, an individual module of the protocol can be separately used to identify crash-risk locations even though a complete dataset is not available. In this study, four publicly accessible datasets were used. Since availability of the data is different across states and countries, each individual module of algorithm can be applied and adapted circumstantially. For instance, some states do not offer intersection point data, so researchers mainly rely on roadway network GIS data to create nodes as intersection points. Still, modules for buffering techniques introduced in this study as a part of algorithm can be applied to locate unique intersection locations without overlaps. Also, modules for creating Voronoi cells also can be used to detect non-intersection locations without the input of public intersection point data.

Lastly, a larger number of crash-risk locations can be obtained if a certain research project requires them. In this study, we generated a Voronoi diagram only for one time to identify non-intersection locations without pedestrian crashes. However, this module can be repeated multiple times to create more control observations. Depending on the purpose of study or data availability, the algorithm offers a useful way to create a balanced sample of crash-risk locations.

The study has limitations. First, the algorithm eliminated overlapping intersection points and found lower number of intersection locations compared to the original intersection point data. Although, we evaluated the algorithm with multiple criteria (e.g., sample size, coverage, and accuracy), additional manual check for randomly selected data points (e.g., comparison with aerial photos) might enhance the reliability of this protocol. Second, although each module of the algorithm can be used circumstantially, non-intersection locations with crashes (case observation) are not easy to detect without empirical pedestrian crash data. The availability and quality of pedestrian crash data are different between cities and states though they are generally archived based on the police-recorded reports. Consistency of the pedestrian crash data across jurisdictions will make this algorithm more useful.

Conclusions

The protocol developed in this study provides an efficient and effective way to create spatial analysis units for pedestrian crash modeling. It can provide a substantial time saving in identifying unique intersection and non-intersection locations. The algorithm will also make it possible for researchers to compare their modeling results with other studies by using the same unit of analysis.

The algorithms showed great performance in identifying crash-risk locations at intersection and non-intersection areas. The performance of algorithms can be considered in terms of sample size, coverage, and accuracy. Depending on the objective of study, different standards can be accentuated to make a decision for search radii. For example, the buffer size of 20 m generally showed the highest performance in the analyses. However, 40 m search radius can be an alternative when the study needs larger number of cases at intersection locations. Regardless of study purpose, we suggest not to use larger than 60 m buffer size, which showed a sharp drop of performance in the analyses. Also, more divided search radii should be tested in the future studies to better capture crash-risk locations.

The protocol discussed in this study provides a tool for integrating pedestrian crash data with the transportation network for detecting unique intersection and non-intersection locations. Pedestrian crash modeling using this protocol will broaden the applicability of algorithms and enrich the discussion in the future.

Chapter 3 Environmental Risks and Benefits of Pedestrian-Motor Vehicle Collisions

---

Chapter Abstract

Background: The promotion of non-motorized travel modes is an important goal of transportation planning and public health. The steady increase in trips taken by non-motorized travel modes has raised important safety issues as walkers are the most fragile and vulnerable road users. This study explores the impact of built environment on the frequency of pedestrian-motor vehicle collisions occurring at intersection and non-intersection locations.

Methods: A retrospective cross-sectional analysis was conducted within the jurisdictional boundary of King County, Washington, USA. The outcomes were the occurrence and frequency of collisions on state routes. A number of micro and macro-environment factors were quantified for statistical modeling using geographic information system (GIS). Multivariate mixed-effects models were used to reflect correlations between crash-risk locations on state routes.

Results: Wider and major roadways were associated with higher number of pedestrian crashes. Surrounding environmental factors that encourage high pedestrian activities were also associated with higher number of pedestrian crashes on both intersection and non-intersection locations. These factors included bus ridership, employment density, park and ride, residential area, and service area of land use. Protective effects of pedestrian crashes were observed with household income in all models.

Conclusions: Environmental factors of pedestrian collision identified from this study should be prioritized to reduce the number of pedestrian crashes. Especially, the results in this study offered valuable insights about the need for developing different pedestrian safety strategies and

policies regarding two types of crash-risk location. Future studies are warranted to investigate the effects of detailed traffic condition and behavioral characteristics of drivers and pedestrians on collisions.

Introduction

There was a steady increase in the share of trips taken by non-motorized travel modes in recent years. According to the National Household Travel Survey (NHTS) data, 9.7% of all person-trips were made by non-motorized travel modes such as walking and bicycling in 2009, compared to 6.3% in 1995 [3]. The benefits of increases in non-motorized travel modes have been addressed in a broad range of literature. Increases in walking support national and state departments of transportation goals to reduce vehicle miles traveled (VMT) and traffic congestion [3,59,60]. 33% of total greenhouse gas (GHG) emissions were created from transportation sector, and CO2 accounts for 95% of the GHG emitted from motorized trips in the US [61]. Traveling by non-motorized modes can reduce a variety of GHG emissions and the use of fossil fuels. Walking is an environmentally-friendly alternative to mitigate these problems, and the promotion of active travel modes is an important goal of transportation planning [1].

Walking has been on the public health agenda since 1996, after a publication from the US surgeon general's report on physical activity showed clear benefits [62,63]. It has been identified as a common form of leisure time physical activity in Canada, the US, and Europe [1,13,64,65]. Walking has clear benefits for physical and mental health [43,66] and is associated with greater longevity and reduced chronic diseases [67–70]. Walking is also reflected in national and local directives to redress physical inactivity and obesity epidemics through active transportation [59]. With an increase emphasis on walking, there are also safety issues that need to be addressed as walkers interact with motorized road users.

Environmental correlates of pedestrian collisions

Reducing road trauma is one of the most important safety goals in concentrated urbanized areas, where the prevalence of people walking and pedestrian exposure to traffic are greatest [16]. Extensive studies have been conducted to understand the factors that are associated with pedestrian-motor vehicle collisions. These factors can be separated into micro and macro environmental factors. The micro-environmental factors include roadway characteristics and traffic conditions and are closely associated with driver and pedestrian behavior near high crash-risk locations [16–20,43]. Pedestrian collisions have been associated with higher numbers of roadway lanes [71,72], roadway functional class [23], and posted speed limits [16,23,24,43]. High density of sidewalks, transit ridership, and annual average daily traffic (AADT) have closely correlated to pedestrian and vehicle volumes [43,73], which might be associated with high number of crashes.

Macro-environments relate to the characteristics of neighborhood and activities of people. In the literature, socio-demographics such as population size and density, racial composition, employment, household income were used to estimate the number and severity of collisions [16,43,58,73–75]. Despite the pedestrian volume is an essential component of estimating pedestrian collisions, few transportation agencies collect pedestrian count data on a regular basis since collecting site-specific pedestrian volume information costs much and requires a lot of time [33]. Alternatively, various land use types were taken into account in the analysis to represent pedestrian activities [16,18,23,43].

Crash-risk locations and spatial autocorrelation

Statistical models developed for estimating pedestrian collision frequency have adopted location-based approaches to count the number of pedestrian-motor vehicle crashes within the boundary of spatial unit. A variety of spatial analysis units have been applied in previous studies, to include: regions [76], counties [56,77], districts [78], census tracts [79], census block groups [80], and traffic analysis zones (TAZs) [81]. Although various spatial units have been widely used in the literature [23], intersection and non-intersection locations can offer much finer-grained analysis results since they are generally identified as point features with higher-resolution [17,18,31,33,43]. An intersection is a place where two or more roadways cross, and a non-intersection is any location within a roadway segment or along a transportation facility, that is not at an intersection [47]. Two locations are where pedestrian-motor vehicle collisions may occur hence need more attention for pedestrian crash modeling. The modeling results from intersection and non-intersection locations might be different [42]. And, using a common model to jointly estimate impacts on the frequency of collision at both types of locations may result in biased estimates. Thus, it's necessary to investigate each location, respectively.

While intersection and non-intersection locations have clear advantages with respect to the resolution of data, they can be a challenge to analyze given the existence of spatial autocorrelation. That is, the spatial objects are correlated with adjacent objects across a spatial area, thus violating statistical assumptions of mutual independency. Generally, two approaches have been adopted in the literature to mitigate spatial autocorrelation. The first approach is to use refined spatial data that are considered to be uncorrelated. This approach requires extensive site surveys, field investigation, and data processing [82]. The other approach is applying statistical models which can control the correlations within spatial clusters. This approach is more widely used because it is relatively less time-consuming and labor-intensive compared to the first approach [27,29,56–58,83]. The structures of spatially correlated data can be clarified through mixed-effects models, which can account for the correlations within spatial clusters using random effects.

Objectives

This study intends to empirically assess the impacts of built environment attributes on the frequency of pedestrian-motor vehicle collisions at intersection and non-intersection locations. Especially, unique crash-risk location data were developed and used in the model to mitigate spatial autocorrelation. And, multivariate mixed-effects models were applied to reflect correlations within and between state routes. Specific objectives were (1) to estimate incidence rate ratio (IRR) of pedestrian-motor vehicle collisions by each environmental factor; (2) to compare significant environmental correlates of pedestrian collisions between intersection and non-intersection locations; (3) to visualize marginal effects of environmental factors and predict the number of crashes at crash risk locations, which would inform safety programs how and where to prioritize their limited resources. We hypothesized that pedestrian-motor vehicle collisions were more likely to occur on principal arterial roads with wider roadways and surrounding environmental factors that support higher pedestrian activities. Also, we expected that walkable neighborhood characteristics are related to lower number of pedestrian-motor vehicle collisions.

Methods

Study design

A retrospective cross-sectional analysis was conducted to estimate the frequency of pedestrian-motor vehicle collisions at crash-risk locations within the jurisdictional boundary of King County, Washington, USA. The county had 29.6% of the state population (2.2 million in 2017), still had a disproportionate 45.1% of all statewide pedestrian-motor vehicle collisions. The central city of King County, Seattle (population size: 0.72 million in 2017) had more than half (51.3%) of the total pedestrian crashes occurred in King County.

A protocol was developed to identify uncorrelated and unique intersection and non-intersection locations as a spatial unit of analysis. Four publicly accessible datasets were used in the process: roadway intersection points, roadway lane polyline data, legal speed limit polyline data, and pedestrian crash records. Crash-risk locations on limited access roadways (e.g. interstate) were removed from the analysis since they lacked pedestrian access. 1,455 intersection locations and 460 non-intersection locations in King County were identified and used in the analysis. A number of micro and macro-environment factors were quantified for statistical modeling using geographic information system (GIS).

Measurement and data sources

Outcome

The outcome of this study was the occurrence and number of pedestrian-motor vehicle collisions occurred on state routes between January 1, 2013 and December 31, 2017 within the boundary of King County. The data were directly obtained from the Transportation Data, GIS & Modeling Office of the Washington State Department of Transportation (WSDOT). There were

860 pedestrian-motor vehicle crashes on state routes within the study period. A Euclidean buffer with 100 m search radius from each crash-risk location was created to count the number of pedestrian-motor vehicle crashes.

All reported collisions were recorded based on the federally mandated Fatality Analysis Report Systems (FARS) protocols [16]. Each collision was counted as one incident regardless of the number of pedestrians involved. The data had assorted individual crash level information, to include: jurisdiction, time, weather, road surface condition, lighting condition, vehicle characteristics, socio-demographic and behavioral characteristics of drivers and pedestrians. Linear Referencing System (LRS) using state route milepost information was applied to locate each crash point. Latitude and longitude of each collision were derived from this method and used to map pedestrian crashes.

Micro-environments

The micro-environment factors examined included roadway design and traffic condition such as number of roadway lanes, roadway width, roadway functional class, posted maximum speed, length of sidewalk, bus ridership density, annual average daily traffic (AADT). Roadway characteristics data were collected from the Office of Information Technology of the WSDOT. The data were originally extracted from the State Highway log and the Transportation Information and Planning Support (TRIPS) mainframe database. Continuous variables (e.g. roadway width) were used in the analysis. But, categorical variables (e.g. functional class) were also created to represent mutually exclusive characteristics of variables. Sidewalk data were generated in Urban Form Lab (UFL) using various GIS data sources from King County [84]. Bus ridership density was measured by using daily average boarding and alighting information from each bus stop. Annual Average Daily Traffic (AADT) data were obtained from the WSDOT traffic geoportal. Buffering techniques were used for partitioning the space for the measurement of locational characteristics. Euclidean buffers with 100 m search radius from intersection and non-intersection locations were created to quantify micro-environment characteristics.

Macro-environments

Macro-environments are related to the characteristics of neighborhood and land use. Euclidean buffers with 400 m search radius from intersection and non-intersection locations were created to capture macro-level environmental factors, including the following: residential density, employment density, intersection density, public school, school enrollment, park and ride, trail, population size, racial composition, household income, proportion of each land use category.

The measurement of residential and employment density were conducted using SmartMaps, a novel approach to the quantification of built environment [85]. To create SmartMaps, spatially continuous surfaces of grid cells (30 m × 30 m) were produced for quantifying the unit counts within a search radius (400 m). The value of each raster cell represents the number of residential unit counts and jobs for the focal cell's neighborhood. The data for residential density were derived from the property parcel GIS data created by the Department of Assessments, King County, Washington. The source data of employment density were obtained from the tax assessor and US Bureau of Labor Statistics [86]. Intersection density was measured by counting the number of unique intersection locations within the buffers. Locations of public schools and school enrollment data were obtained from the National Center for Education Statistics (NCES). Park and ride data contained information about active park and ride lots across Washington

State, and the source data came from WSDOT. Trail data were obtained from the King County GIS Open Data.

Neighborhood sociodemographic data including population and racial composition were derived from the US census data at the census block level. Household income data were collected from the American Community Survey (ACS) at the census block group level. Parcel level land use data were obtained from the King County Department of Assessments. To quantify the sociodemographic and land use characteristics around intersection and non-intersection locations, area-weighted arithmetic values were calculated by using the proportion of census block, census block group, and parcel-level land use within the buffers.

Statistical analyses

Descriptive statistics including the distribution of data and correlation matrix were checked and analyzed before statistical modeling. A binary outcome from intersection and non-intersection location was created to examine the univariate association between environmental factors and the occurrence of pedestrian-motor vehicle collision. Odds ratios (OR) with 95% confidence intervals (CI) were estimated from this analysis.

Multivariate mixed-effects Poisson regression models were applied in the next step to estimate the incidence rate ratios (IRR) and 95% confidence interval (CI) for the relationship between environmental factors and the number of pedestrian-motor vehicle collisions. Since all intersection and non-intersection locations were clustered within state routes, the random effect component of the regression models accounted for the correlation within and between crash-risk locations and state routes. Specifically, random intercept models were applied in the analysis as follows:

$$\ln(L_{ij}) = \gamma_0 + \sum_{p=1}^{r} \gamma_p x_{pij} + U_j + R_{ij} \tag{1}$$

where L is the number of pedestrian-motor vehicle collisions, i represents a crash-risk location, j indicates a state route, γ_o is the intercept, γ_p is a regression coefficient corresponding to the p^{th} predictor variable x_{pij}, U_j is a random effect for the j^{th} cluster. γ_o+U_j is the random intercept for the j^{th} cluster. R_{ij} is a random error. This model assumes that the set of U_j, set of R_{ij}, and covariates x_{pij} are mutually independent. Since the data were not over-dispersed, the Poisson link in the regression model accounts for the distribution of number of pedestrian crashes and results in coefficients taking the form of incidence rate ratios (IRR) when exponentiated.

A full model with all variables was first estimated as a reference, then a refined model was developed based on the results from stepwise variable selection processes in R programming. The Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) were used to identify the best-fit model. When a high correlation was observed between some factors in previous analysis, those variables were excluded in the final models. Based on the refined model, two more models were estimated by crash-risk location type. Intersection and non-intersection models were separately developed to show the difference of affecting factors of pedestrian-motor vehicle collisions between two types of location.

Model visualization and marginal effects

The refined intersection and non-intersection models developed were compared visually and numerically. Coefficients and associated confidence intervals (95% CI) were plotted using a forest plot. Marginal effects of all variables (where other variables were set to mean values) were also estimated and visualized. Marginal effects of predictors with confidence intervals (95% CI) were compared by location type in the analysis.

Prediction of pedestrian collisions and PKDE maps

The expected number of pedestrian-motor vehicle collisions at each intersection and non-intersection location was estimated based on the refined models. An inverse link function was applied to calculate the number of pedestrian collisions using given environmental characteristics from each crash-risk location. The results of predicted number of pedestrian collisions were then exported as CSV format files using a unique identifier of each intersection and non-intersection location. The expected number of pedestrian crashes were merged to original intersection and non-intersection location GIS layers in ArcGIS.

A planar kernel density estimation (PKDE) map was created to calculate the density of predicted pedestrian collisions in a neighborhood around each crash-risk location. A smoothly curved surface was fitted over each crash-risk location as a distance decay function as follows:

$$PKDE_k = \sum_{i=1}^{n} \frac{1}{\pi r^2} k\left(\frac{d_{is}}{r}\right) \quad (2)$$

where $PKDE_k$ is the density of pedestrian-motor vehicle collisions at crash-risk location *s*, *r* is a search radius of the function. *k* is a weight of point *i* at distance d_{is} to location *s*. The function calculates the distance decay effect from the centroid of the raster cell to all crash-risk locations within the search radius. A 100 m search radius was used from the centroid of each grid cell. The 30 m × 30m raster cells were used to calculate and visualize the expected density of pedestrian collisions.

Results

Descriptive statistics and univariate analysis

There were 682 crash-risk locations that included any pedestrian collisions (intersection: 67.3%, non-intersection: 32.7%) and 1,233 locations with no collisions (intersection: 80.8%, non-intersection: 19.2%) in King County. Table 6 shows some of the summary statistics. Univariate analyses were conducted to describe the patterns of each continuous or categorical variable. A binary outcome from each intersection and non-intersection location was created to examine the univariate association between environmental factors and the occurrence of pedestrian-motor vehicle collision. Odds ratios (OR) with 95% confidence intervals (CI) were estimated from this analysis.

Many environmental characteristics were identified as affecting factors of the occurrence of pedestrian-motor vehicle collisions, to include: number of roadway lanes (≥ 2), roadway width, principal arterial roads, sidewalk, bus ridership, residential density, employment density, intersection density, school enrollment, park and ride, population size, and land use of service area. Whereas, some variables showed protective effects of the occurrence of collisions: posted maximum speed limits (≥ 40 MPH), trail, household income, and land use of industrial area. There were some inconclusive factors that showed statistical insignificance including annual average daily traffic (AADT), presence of public school, racial composition, park area, and residential area.

Table 6 Descriptive statistics by crash-risk locations with pedestrian collisions compared to crash-risk locations with no pedestrian collisions. Unadjusted odds ratio (OR) of pedestrian crash occurrence and 95% confidence interval (CI) for each continuous or categorical variable.

		Any crashes N: 682 Mean (SD) or Count (%)	No crashes N: 1,233 Mean (SD) or Count (%)	Unadjusted Odds Ratio [a] (95% CI)
Crash-risk location type				
	Intersection	459 (67.3%)	996 (80.8%)	0.49 (0.40-0.61)
	Non-intersection	223 (32.7%)	237 (19.2%)	Ref.
Micro-environment characteristics (100-m Euclidean area around crash-risk locations)				
Number of Roadway Lanes [b]				
	1 lane	59 (8.7%)	349 (28.3%)	Ref.
	2 lanes	475 (69.6%)	619 (50.2%)	4.54 (3.36-6.13)
	3 lanes	138 (20.2%)	214 (17.4%)	3.81 (2.69-5.41)
	4 lanes and more	10 (1.5%)	51 (4.1%)	1.16 (0.56-2.41)
Roadway width (m)		8.3 (1.9)	7.6 (2.8)	1.12 (1.08-1.16)
Roadway functional class				
	Principal arterial	446 (65.4%)	525 (42.6%)	2.55 (2.10-3.09)
	Non-principal arterial	236 (34.6%)	708 (57.4%)	Ref.
Maximum speed				
	≤ 40 MPH	415 (60.9%)	577 (46.8%)	Ref.
	> 40 MPH	267 (39.1%)	656 (53.2%)	0.57 (0.47-0.68)
Sidewalk (m)		430 (340)	250 (340)	1.15 (1.12-1.19)
Bus ridership density [c]		5,900 (13,300)	1,500 (8,100)	1.07 (1.05-1.08)
Annual average daily traffic (count)		48,000 (44,000)	52,000 (53,000)	0.98 (0.97-1.00)
Macro-environment characteristics (400-m Euclidean area around crash-risk locations)				
Residential density [d]		1,140 (1,280)	750 (1,140)	1.03 (1.02-1.04)
Employment density [e]		4,300 (13,500)	2,000 (6,600)	1.02 (1.01-1.04)
Intersection density (count per km^2)		10 (6.2)	8.7 (6.3)	1.03 (1.02-1.05)
Public school				
	Presence	107 (15.7%)	160 (13.0%)	1.25 (0.96-1.63)
	Absence	575 (84.3%)	1,073 (87.0%)	Ref.
School enrollment (count)		120 (390)	70 (240)	1.05 (1.02-1.08)
Park and ride				
	Presence	121 (17.7)	145 (11.8%)	1.62 (1.25-2.10)
	Absence	561 (82.3)	1,088 (88.2%)	Ref.
Trail (m)		240 (420)	330 (530)	0.96 (0.94-0.98)
Census (blocks, block groups)				
Total population [f]		2,200 (1,300)	1,600 (1,200)	1.47 (1.36-1.58)
Race (% of white) [f]		77.1 (10.6)	76.3 (11.9)	1.01 (1.00-1.02)
Household income ($USD) [g]		66,767 (21,526)	78,275 (26,785)	0.82 (0.79-0.86)

Parcel-level land use			
Park area (%)	1.2 (4.1)	1.6 (5.6)	0.98 (0.96-1.00)
Residential area (%)	31.5 (19.2)	30.1 (21.3)	1.00 (1.00-1.01)
Industrial area (%)	0.4 (1.4)	1.0 (2.9)	0.88 (0.83-0.93)
Service area (%)	12.7 (12.8)	8.5 (11.2)	1.03 (1.02-1.04)

[a] Odds ratio (OR) per rescaled unit (sidewalk and trail per 100 meters; bus ridership per 1,000 count; annual average daily traffic per 10,000 traffic; residential density per 100 units; employment density per 1,000 jobs; school enrollment per 100 students; total population per 1,000 people; household income per 10,000 $USD).
[b] The total number of increasing and decreasing roadway lanes was divided by two.
[c] SmartMaps (daily average boarding and alighting per square km using 100-meter buffer).
[d] SmartMaps (units per square km using 400-meter buffer).
[e] SmartMaps (jobs per square km using 400-meter buffer).
[f] Using census block information.
[g] Using census block-group information.

Multivariate mixed effects Poisson models

<u>Modeling for all types of crash-risk location</u>

Multivariate mixed effects Poisson models were applied to estimate the risk and protective factors of pedestrian-motor vehicle collisions at all types of crash-risk location (Table 7). A full model that included all factors was initially developed, followed by a refined model using a stepwise variable selection process in the statistical package, R version 3.5.0. Some environmental factors were removed from the final model since they had a high correlation with other factors (e.g. sidewalk and total population).

Regarding micro-environment characteristics, the number of lanes (≥ 2) was strongly associated with higher number of crashes compared to a single lane. Principal arterial roadways showed more than two times higher number of crashes when compared to non-principal arterial roadways. Bus ridership showed an IRR slightly higher than 1. At the macro-environment level, employment density, residential area, and service area were associated with higher number of pedestrian crashes. The presence of park and ride also showed an IRR higher than 1. Household income and industrial area showed significantly lower IRR (<1.00).

Table 7 Multivariate mixed effects Poisson models for estimating pedestrian-motor vehicle crash frequency. Incidence rate ratio (IRR) and 95% confidence interval (CI) comparison between full model and refined model.

		Full Model IRR [a] (95%CI)	Refined Model IRR [a] (95% CI)
Crash-risk location type			
	Intersection	0.70 (0.61-0.79)	0.72 (0.64-0.81)
	Non-intersection	Ref.	Ref.
Micro-environment characteristics (100-m Euclidean area around crash-risk locations)			
Number of Roadway Lanes			
	1 lane	Ref.	Ref.
	2 lanes	1.67 (1.21-2.31)	2.21 (1.67-2.94)
	3 lanes	1.47 (0.98-2.21)	2.40 (1.75-3.31)
	4 lanes and more	1.52 (0.87-2.68)	2.67 (1.67-4.27)
Roadway width (m)		1.07 (1.02-1.12)	
Roadway functional class			
	Principal arterial	2.24 (1.85-2.71)	2.49 (2.09-2.98)
	Non-principal arterial	Ref.	Ref.
Maximum speed			
	≤ 40 MPH	Ref.	
	> 40 MPH	0.80 (0.69-0.92)	
Sidewalk (m)		1.07 (1.05-1.09)	
Bus ridership density		1.02 (1.01-1.02)	1.02 (1.01-1.02)
Annual average daily traffic (count)		0.98 (0.96-1.01)	
Macro-environment characteristics (400-m Euclidean area around crash-risk locations)			
Residential density		0.99 (0.98-1.00)	1.01 (1.00-1.01)
Employment density		1.02 (1.01-1.02)	1.02 (1.01-1.02)
Intersection density (count per km^2)		1.00 (0.99-1.01)	
Public school			
	Presence	1.02 (0.83-1.25)	
	Absence	Ref.	
School enrollment (count)		1.02 (1.00-1.04)	
Park and ride			
	Presence	1.45 (1.24-1.68)	1.46 (1.26-1.70)
	Absence	Ref.	Ref.
Trail (m)		0.98 (0.97-1.00)	
Census (blocks, block groups)			
Total population		1.15 (1.07-1.24)	
Race (% of white)		0.99 (0.99-1.00)	
Household income ($USD)		0.93 (0.90-0.97)	0.91 (0.87-0.94)
Parcel-level land use			
Park area (%)		0.98 (0.96-1.00)	
Residential area (%)		1.01 (1.00-1.02)	1.01 (1.01-1.02)

Industrial area (%)	0.93 (0.88-0.98)	0.89 (0.85-0.94)
Service area (%)	1.02 (1.02-1.03)	1.03 (1.02-1.03)

[a] Incident rate ratio (IRR) per rescaled unit (sidewalk and trail per 100 meters; bus ridership per 1,000 count; annual average daily traffic per 10,000 traffic; residential density per 100 units; employment density per 1,000 jobs; school enrollment per 100 students; total population per 1,000 people; household income per 10,000 $USD).

Modeling by crash-risk location type

Based on the refined model, separate intersection and non-intersection models were developed to identify differences between the two location types (Table 8). At the micro-environment level, the number of lanes (≥ 2) was associated with higher number of crashes. However, it was not statistically significant in non-intersection model. Roadway functional class and bus ridership density showed an IRR higher than 1 in both models. In terms of macro-environment characteristics, employment density, the presence of park and ride, residential area and service area were associated with higher number of crashes in two models. Residential density was also related to higher number of crashes. However, it was statistically significant only in the intersection model. Household income was associated with lower number of crashes in both models. Industrial area also showed a lower number of crashes. However, it was significant only at intersection locations.

Table 8 Multivariate mixed effects Poisson models for estimating pedestrian-motor vehicle crash frequency by crash location type. Incidence rate ratio (IRR) and 95% confidence interval (CI) comparison between intersection and non-intersection model.

	Intersection Model IRR [a] (95%CI)	Non-intersection Model IRR [a] (95% CI)
Micro-environment characteristics (100-m Euclidean area around crash-risk locations)		
Number of Roadway Lanes		
1 lane	Ref.	Ref.
2 lanes	3.20 (2.22-4.62)	1.02 (0.67-1.54)
3 lanes	3.37 (2.25-5.06)	1.05 (0.63-1.74)
4 lanes and more	4.31 (2.15-8.61)	1.10 (0.58-2.10)
Roadway functional class		

Principal arterial	2.64 (2.11-3.31)	2.01 (1.49-2.72)
Non-principal arterial	Ref.	Ref.
Bus ridership density	1.02 (1.01-1.02)	1.03 (1.02-1.04)
Macro-environment characteristics (400-m Euclidean area around crash-risk locations)		
Residential density	1.01 (1.00-1.02)	1.00 (0.99-1.01)
Employment density	1.02 (1.01-1.02)	1.02 (1.01-1.03)
Park and ride		
Presence	1.47 (1.23-1.76)	1.62 (1.24-2.12)
Absence	Ref.	Ref.
Census (block groups)		
Household income ($USD)	0.89 (0.85-0.93)	0.92 (0.86-0.99)
Parcel-level land use		
Residential area (%)	1.01 (1.01-1.02)	1.01 (1.00-1.02)
Industrial area (%)	0.89 (0.84-0.94)	0.92 (0.84-1.01)
Service area (%)	1.03 (1.02-1.03)	1.03 (1.02-1.03)

[a] Incident rate ratio (IRR) per rescaled unit (bus ridership per 1,000 count; residential density per 100 units; employment density per 1,000 jobs; household income per 10,000 $USD).

Model visualization

Model comparison

The forest plots for the intersection and non-intersection models are shown in Figure 5. The plots include the incidence rate ratio (IRR) and 95% confidence interval (CI) of the model in each panel. The vertical black line on each panel represents the null hypothesis (IRR=1). Bars that do not touch the vertical black line indicate a significant relationship. Positive and negative relationships were visualized with two colors (blue > IRR of 1, red < IRR of 1). Many factors were not statistically significant in non-intersection model (Table 8).

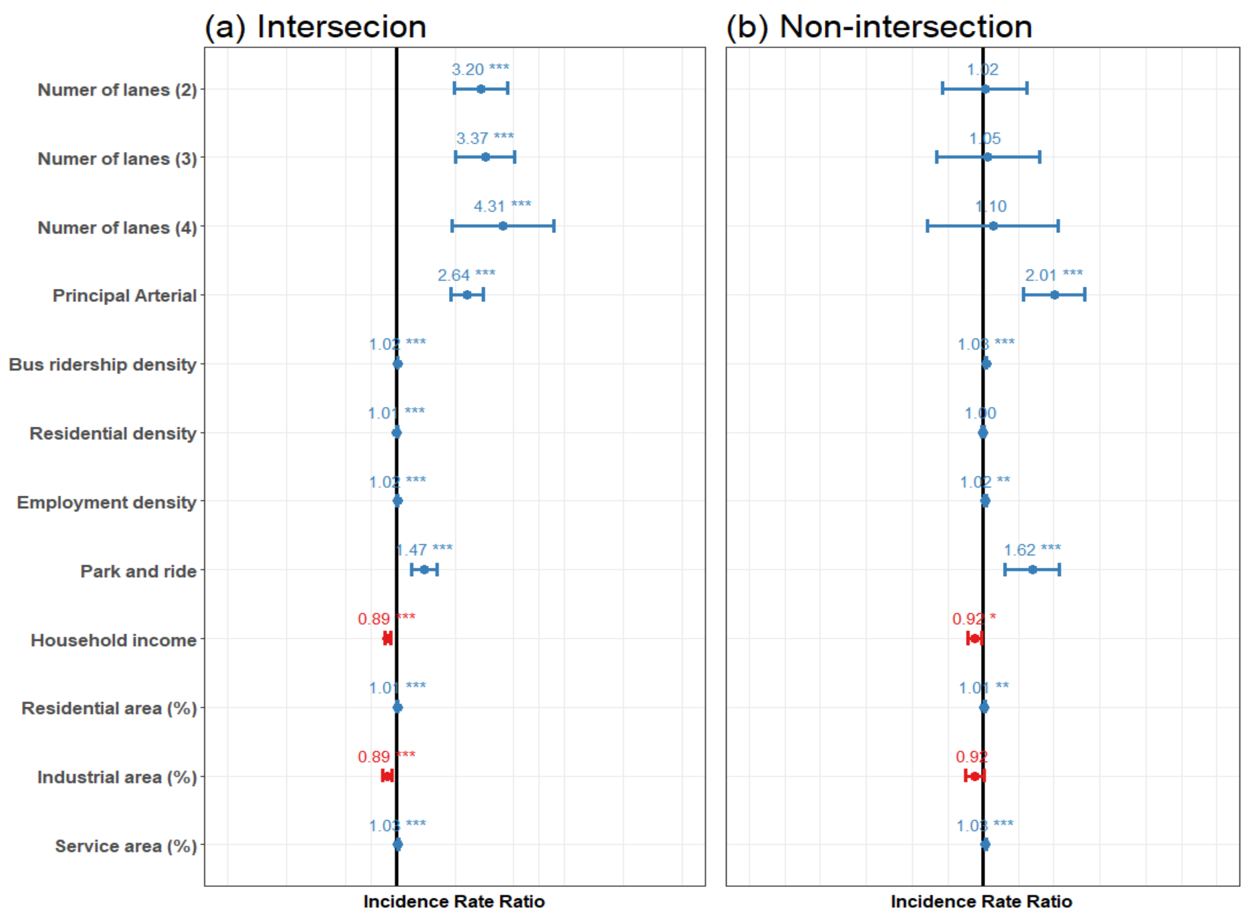


Figure 5 Forest plots of incidence rate ratios (IRRs) and 95% confidence interval (CI) for comparison between intersection and non-intersection models. A vertical black line on each panel represents null hypothesis (IRR=1).

Marginal effects

The marginal effects are shown in Figure 6. Marginal effects of predictors by location type were compared in each panel. There were several crash risk factors identified including bus ridership density, employment density, residential area, and service area. The incidence rate ratio (IRR) decreased with higher values of other variables including household income and industrial area.

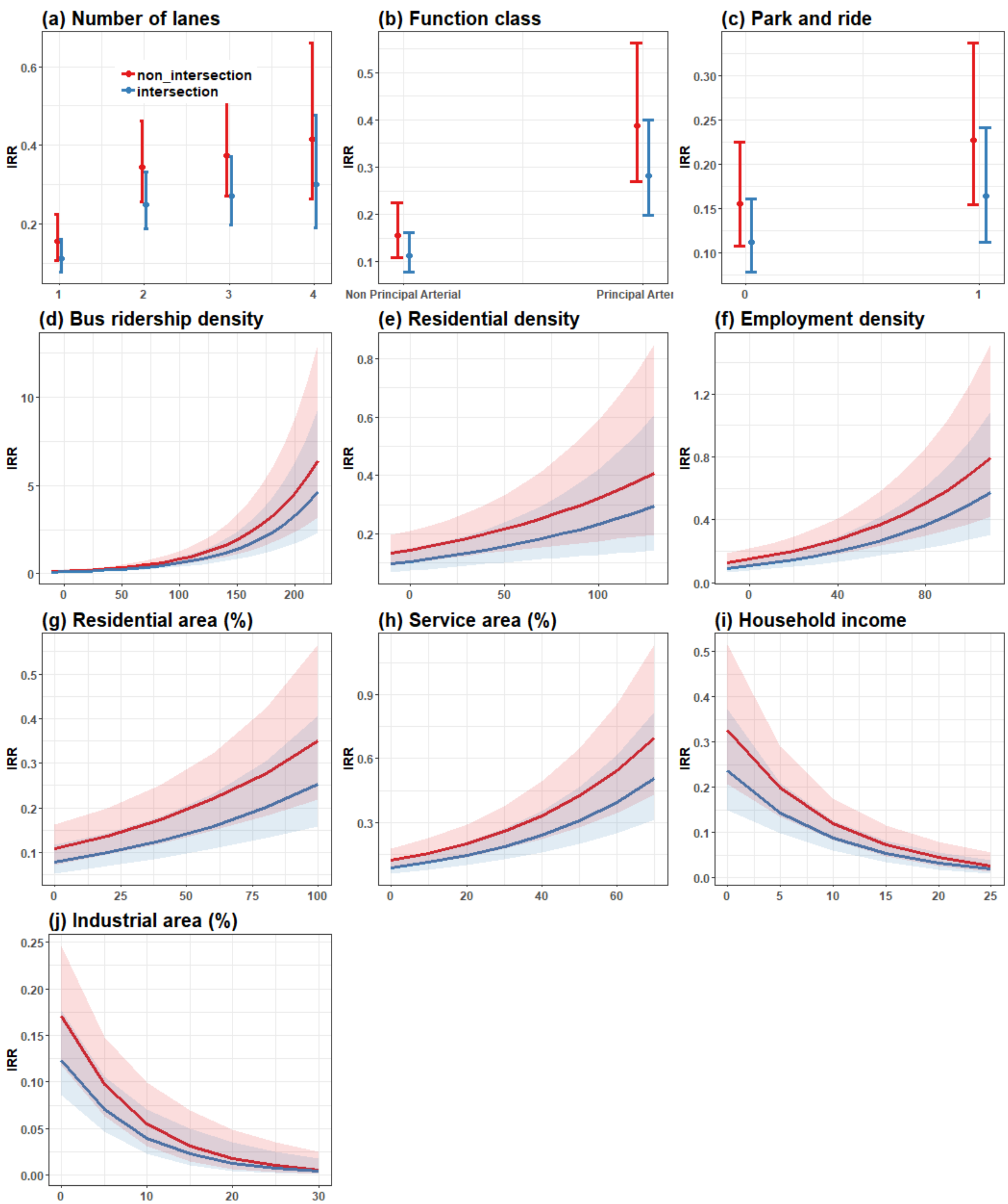


Figure 6 Marginal effects and 95% confidence interval (CI) of predictors by crash-risk location type

PKDE maps of expected pedestrian collisions

Figure 7 shows the planar kernel density estimation (PKDE) map of predicted number of pedestrian-motor vehicle collisions along the state routes based on the multivariate mixed-effects modeling results. High number of collisions was predicted on downtown areas and a part of arterial roadways. The PKDE value means the predicted number of pedestrian-motor vehicle crashes per square kilometer.

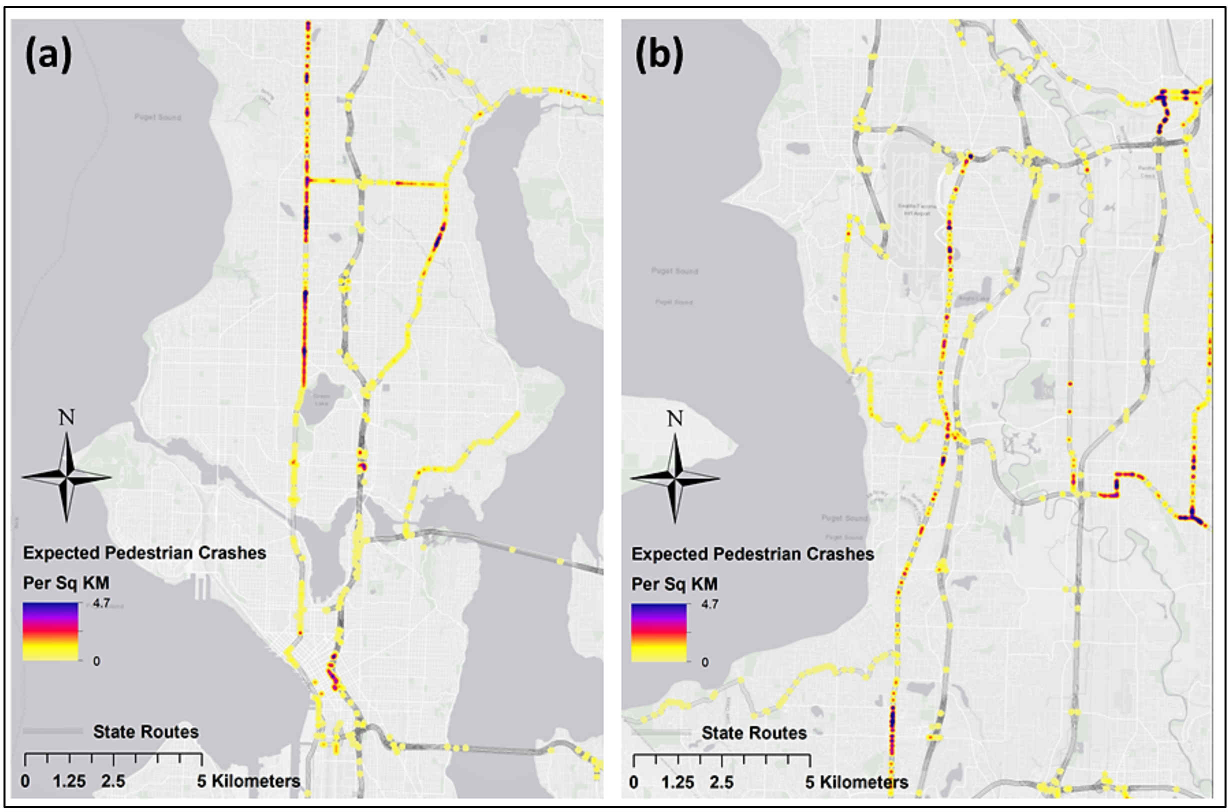


Figure 7 Planar kernel density estimation (PKDE) maps of expected number of pedestrian-motor vehicle collisions along the state routes in the city of Seattle (a), Renton and Kent (b).

Discussion

In this study, environmental risks and benefits of pedestrian-motor vehicle collisions were evaluated by applying multivariate mixed-effects models. Unique intersection and non-intersection locations were identified to reduce the risk of spatial autocorrelation. A systematic protocol was applied to eliminate overlapping areas between crash-risk locations. Retrospective cross-sectional analyses were conducted using pedestrian collision frequency as the statistical outcome. Micro and macro environmental characteristics were quantified and assessed by creating Euclidean buffers from intersection and non-intersection locations.

All location models were first developed to estimate the effects of environmental factors on the pedestrian collision frequency. A full model was presented as a reference using all environmental variables. A refined model was then estimated based on the results from descriptive analysis and variable selection processes. Since these models did not capture interaction effects of location with other variables, two location-specific models were estimated based on the refined model by crash-risk location type; one for intersection and one for non-intersection. These models were visualized and compared through forest plots and marginal effects plots. Prediction of pedestrian collisions was made and mapped based on location-specific models.

In the refined model using all crash-risk locations, several micro-environment variables were shown to have a significant association with pedestrian crashes. The number of roadway lanes (≥ 2) showed a strong association with higher number of crashes. Higher number of roadway lanes is strongly related to higher volumes of vehicles [16]. In our model, principal arterial roads showed high number of collisions compared to non-principal arterial roads. Major arterials are generally used by larger vehicles (e.g. trucks) with higher speeds when compared to vehicles on local/residential roads. Bus ridership density was also associated with higher number of crashes. Since higher density of transit ridership is related to higher pedestrian activity [87], this variable represents high volume of pedestrians around crash-risk locations.

We hypothesized that higher vehicle speed would be associated with higher number of pedestrian collisions. However, our full model showed that the higher posted speed limits (> 40 MPH) was associated with lower number of crashes when compared to those posted speeds of 40 MPH or less. Because posted maximum speed limits are different from direct measures of vehicular speed, this variable may be picking up other locational characteristics. A state route with a high posted speed limit is typically not accessible by pedestrians since they lack transportation sidewalks and other facilities for walkers. For these reasons, the posted speed limits showed negative (or non-significant) relationships with high (or severe) pedestrian-motor vehicle collisions in previous studies [16,43].

Neighborhood and land use characteristics were included as macro-environments in the model. Locations with more housing and businesses were represented as high residential density, employment density and proportion of service area. These factors capturing the level of pedestrian activity were associated with high number of crashes in the model. The presence of Park and Rides might generate walk trips to and from the transit stops [88], and it was associated with higher number of crashes. Increased proportion of residential area within crash-risk location was also related to higher number of crashes. The protective effects of pedestrian-motor vehicle collisions were found in household income and proportion of industrial area. Lower income areas were associated with higher number of crashes, similar to findings in previous studies [89–91]. Industrial areas were less attractive places for walkers, lacking pedestrian infrastructural facilities [92].

Affecting factors of pedestrian-motor vehicle collisions were different between intersection and non-intersection models. The number of roadway lanes was statistically significant only in the intersection model. When a pedestrian is exposed to a roadway for longer time (e.g. crossing four-lane roadway), the risk of collision will be high. However, this association was only effective at intersection locations where most crossings of pedestrians are occurring. Residential density and industrial land use were also statistically significant only in the intersection model though the signs of coefficients were same between two models. The results of other micro/macro-environments were consistent in two models.

The models developed in this study shows many benefits. First, spatial autocorrelation was mitigated by using unique crash-risk location data and statistical models that control the correlation within and between clusters. Specifically, a systematic protocol to create unique crash-risk location data without overlaps was used in the study. Further, the spatial autocorrelation between crash-risk locations caused by similar environmental attributes along a vehicle corridor was mitigated by applying multivariate mixed-effects models. Fixed and random effects were estimated to account for variations within and between state routes. Although these approaches required extensive GIS data processing and computational power, it led to enhanced performance of modeling results.

Second, comparison between intersection and non-intersection location models gave insight into what kind of countermeasure should be prioritized on each type of location. Since some micro-environment variables were only significant in the intersection model, different pedestrian safety strategies might be required for each type of location. Previous studies pointed out the limitation of engineering based measures to reduce the number of pedestrian collisions [16,43]. Generally, non-intersection locations have less complicated roadway design and traffic signal system compared to intersection locations. Thus, behavioral countermeasures of drivers and pedestrians might be more important than engineering-based solutions at these locations.

Third, statistical modeling results were provided with advanced visualization techniques which are comprehensible to policy makers and general audience. Forest plots are easier to grasp than tables of statistical modeling results. The association between each environmental characteristic and the frequency of pedestrian collisions was visualized in a marginal effects plot by controlling other factors. Especially, the expected number of collisions were estimated based on final models and provided as PKDE maps. These maps may serve a policy purpose by prioritizing locations with high predicted number of pedestrian collisions [43].

There were limitations observed. First, the study covered intersection and non-intersection locations within the boundary of King County, Washington. Although we collected extensive environmental datasets from various sources, some data were not available for the county. Previous studies conducted with the city level data included more detailed information on traffic condition [32,79,81]. Considering the results from non-intersection location model where some environmental factors were not significant, further models including detailed traffic condition variables should be tested even though the model covers smaller areas.

Lastly, behavioral characteristics of drivers and pedestrians were not taken into account in this study. Since the unit of analysis was an individual crash-risk location, it was difficult to obtain aggregate behavioral information of drivers and pedestrians. These data were only available through pedestrian collision data which were created based on the police records. Since the collision data only included characteristics of occurred collisions (case observation), another dataset needs to be developed to include control observations for modeling purpose.

Alternatively, individual crash level modeling where the outcome is severity of pedestrian crash can be estimated using behavioral characteristics of two parties in further studies.

Conclusions

In this study, environmental risks and benefits of pedestrian-motor vehicle collisions at intersection and non-intersection locations were examined using micro and macro environment data. Several environmental factors capturing roadway characteristics, traffic condition, neighborhood characteristics, and land use were associated with high number of pedestrian-motor vehicle crashes. Especially, differences in results between intersection and non-intersection location models provided justification for the separation of two models in the analysis. Also, the results offered valuable insights about the need for developing different pedestrian safety strategies and policies regarding two types of crash-risk location.

Since wider major roadways on intersection locations showed high number of crashes, pedestrian safety programs should be more focused on them. Other correlates of high number of pedestrian-motor vehicle collisions, which were common to both intersection and non-intersection locations were related to pedestrian volumes, to include: bus ridership, residential density, employment density, park and ride, residential area, and service area of land use. Locations with high rates of these characteristics should be prioritized to reduce the number of pedestrian collisions. Household income showed robust protective effects of pedestrian collisions in all models, suggesting that great attention should be paid to low income neighborhoods. Lastly, future studies are warranted to investigate the effects of more detailed traffic condition and behavioral characteristics of drivers and pedestrians on collisions.

Chapter 4 Capturing Fine-Scale Travel Behaviors

Chapter Abstract

Background: Device-collected data from GPS and accelerometers for identifying active travel behaviors have dramatically changed research methods in transportation planning and public health. Automated algorithms have helped researchers to process large datasets with likely fewer errors than found in other collection methods (e.g., self-report travel diary). In this study, we compared travel modes identified by a commonly used automated algorithm (PALMS) that integrates GPS and accelerometer data with those obtained from travel diary estimates.

Methods: Sixty participants, who made 2,100 trips during seven consecutive days of data collection, were selected from among the baseline sample of a project examining the travel behavior impact of a new light rail system in the greater Seattle, WA (USA) area. GPS point level analyses were first conducted to compare trip/place and travel mode detection results using contingency tables. Trip level analyses were then performed to investigate the effect of proportions of time overlap between travel logs and device-collected data on agreement rates. Global performance (with all subjects' data combined) and subject-level performance of the algorithm were compared at the trip level.

Results: At the GPS point level, the overall agreement rate of travel mode detection was 77.4% between PALMS and the travel diary. The agreement rate for vehicular trip detection (84.5%) was higher than for bicycling (53.5%) and walking (58.2%). At the trip level, the global performance and subject-level performance of the PALMS algorithm were 46.4% and 42.4%, respectively. Vehicular trip detection showed highest agreement rates in all analyses. Study

participants' primary travel mode and car ownership were significantly related to the subject-level mode agreement rates.

Conclusions: The PALMS algorithm showed moderate identification power at the GPS point level. However, trip level analyses found lower agreement rates between PALMS and travel diary data, especially for active transportation. Testing different PALMS parameter settings may serve to improve the detection of active travel and help expand PALMS's applicability in geographically different urbanized areas with a variety of travel modes.

Introduction

Identifying and understanding human location and movement is a crucial part of transportation planning, public health, and health geography research [93]. Previous studies have shown that people who walk or bike for transportation are more likely to meet public health recommendations by accumulating more physical activity [5,6]. Since the 1970s, governmental planning agencies have relied on travel survey data in order to create transportation models [94–97], including identifying origins and destinations of people's reported movements, travel modes, and related activities. More recently, a growing interest in active transportation (primarily walking and bicycling) on the part of public health [5,6,98,99], has lead researchers to obtain more objectively measured mobility data through the use of such devices as global positioning system (GPS) data loggers and accelerometers to inform transportation models. However, computational algorithms are required to process the massive quantities of GPS and accelerometer data generated in these studies [36,95,98,100–102], and to date, limited work has evaluated these algorithms to understand the details on how they quantify fine-scale travel behaviors.

Self-reported travel diaries are one of the most common instruments for obtaining data on people's locations and movements [103,104]. Traditionally, travel diaries have been considered as the comprehensive source of information for travel behaviors [105]. This self-report approach allows for the collection of data on trip purposes, departure and arrival times, travel mode, and related respondent characteristics including age, sex, race/ethnicity, education level, income, health status, etc. Unfortunately, self-reported travel surveys are a burdensome, leading to low compliance and inaccuracy. The resultant data are susceptible to human errors such as recall or social desirability bias [106]. Common problems include missed trips, incomplete entries, and misreported time stamps and travel behavior characteristics [107]. Nevertheless, important behavioral data such as purpose of travel, visited place names, addresses, and certain travel modes cannot be obtained without study participants' direct input [108]. In addition, in many instances, the primary measurement method for travel behaviors remains self-reported because of cost and feasibility [103,104].

Combined use of individual-based GPS and accelerometer devices to measure location, speed, and physical activity levels has provided new opportunities to characterize travel behaviors at fine spatial and temporal scales [108,109]. During the last few years, the use of these types of devices to obtain objectively measured travel behaviors has taken a salient role in active transportation and physical activity studies [110,111]. Although not entirely error-free, these techniques are considered to provide more objective and accurate measures of travel behavior than the traditional methods that rely on active respondent reporting [112]. As passive data collectors, GPS and accelerometer devices reduce study participant burden and likely enhance data quality [113]. Also, GPS data offer information on traveler's route choice and corresponding speed, which have been difficult to collect through travel survey methods [95].

However, processing massive GPS and accelerometer data sets to reconstruct mobility patterns in terms of trips, trip origins and destinations, and travel mode, requires robust computational power, sophisticated algorithms, and expertise in data management skills [36]. In addition, algorithms often require expert user input for specification of parameter settings for data processing.

PALMS (Personal Activity Location Measurement System) is a web-accessible system enabling the development of travel behavior and physical activity variables from device data [114,115]. Its main purpose is to merge and process time-stamped data from devices such as GPS data loggers, accelerometers, and heart-rate monitors. PALMS was developed by the Center for Wireless and Population Health Systems, University of California, San Diego. PALMS identifies trips and places, and it categorizes trips into three travel modes: walking, bicycling, or vehicular trips. In addition to processing raw time-stamped data, PALMS can aggregate the data into more manageable sets by day, participant, or event.

Besides PALMS, several algorithms have been developed to measure and quantify visited locations, travel behaviors, and physical activities by using GPS and accelerometer data. Some of these algorithms rely solely on the geographical coordinates information collected via standalone GPS data loggers or mobile phones [36,93,116]. These algorithms use frequency, density, and speed information from the spatially-referenced data to detect and classify activity locations and trips. Other algorithms including PALMS use GPS with accelerometer data to identify fine-scale travel behaviors such as walking [101,117,118].

The performance of the PALMS algorithm was evaluated in previous studies [98,114], and results showed moderate agreement rates for travel modes (agreement using SenseCam [114]: 65.3-93.4%; agreement using travel logs [98]: 74.2-89.8%). The performance of other algorithms varied considerably across studies, to include: the percentage of correctly identified trips that were recorded in a travel diary (78.9-86.0%) [95]; the proportion of correctly identified stop locations (92.3%) [36]; the proportion of locations for which a GPS-interview results match was found (50-100%) [116]; the proportion of GPS data time that correctly identified activity location and trip occurrence by comparing with recall interview results (95.8%) [35].

In the travel diary framework, respondents are instructed to record characteristics of individual trips, rather than at fixed time intervals. Therefore, many previous studies assessing specific algorithms used trips as the unit of analysis by connecting consecutive GPS data points sharing a single identified travel mode [95,119–122]. Since people perceive their travel behaviors based on the trip unit, this approach is appropriate.

However, since PALMS produces output data as GPS points with corresponding trip/place information at every time interval (e.g., minute), previous studies assessed the PALMS algorithm using individual GPS point records as the unit of analysis. PALMS was assessed in previous studies by comparing its classification of trips by mode with SenseCam images from forty adult cyclists [114], and travel logs from two research assistants [98]. In these studies, the PALMS algorithm was evaluated through a collective measure; person-level GPS data were coalesced into a single data set and analyzed in aggregate. Data were aggregated because the sample size was too small to conduct subject-level analyses, and socioeconomic and other demographics information was not obtained from study participants.

The objective of the present study was to compare PALMS travel behavior information with that of travel diaries to address the strengths and weaknesses of the PALMS automated algorithm. According to reviews of GPS data processing methods and studies of the reliability of wearable activity trackers [123,124], no other study assessing PALMS or other algorithms conducted the analyses at both GPS and trip levels using the same data sources. In the present study, we first combined data from all subjects to assess PALMS global performance, and second, we evaluated PALMS performance at the participant level to assess individual performance. Lastly, to investigate reasons for possible discrepancies across different levels of analyses, we looked into the socioeconomic characteristics of study participants.

Methods

Data Development

Participants and Sampling

Participant data came from the Travel Assessment and Community (TRAC) project, which recruited >700 baseline participants within the greater Seattle, WA (USA) area between July 2008 and July 2009. The purpose of TRAC was to investigate effects of a new light rail transit (LRT) system on people's travel behaviors in two follow-up data collection efforts [101,125].

For this study, we sampled 60 subjects from baseline TRAC participants. Because the analysis focused on travel mode identification, we conducted a stratified random sample of participants to ensure adequate coverage of the less common travel modes (e.g., transit, walking). Participants were first sorted into groups of drivers (58.1% of the full cohort), transit users (3.8%), or walkers (30.9%), based on the travel mode that each person reported most frequently in his or her travel diary (the remaining 7.2% of participants used bicycle, motorcycle, or taxi for their primary travel mode). From among each of the 3 groups, we randomly selected 20 participants.

Subjects in the final sample were 52.4 years old on average, 46.7% were male, 73.3% were white non-Hispanic, 65% completed a bachelor's or higher degree, 48.3% had full time jobs, 48.3% reported an annual household income ≤ $50k USD, 43.3% were married/partnered. The average household size was 2.1 with 0.4 children. Households had an average of 1.1 cars, and 45% lived in single-family housing.

Accelerometer, GPS, Travel Diary, and LifeLog

Participants were enrolled for one week and wore a GPS data logger (GlobalSat DG-100; New Taipei City, Taiwan) to record geospatial locations, and a hip-mounted accelerometer (ActiGraph GT1M; Pensacola, FL, USA) to measure movement. Participants also recorded place names, addresses, times of arrival and departure, activities at each place, and travel mode from place to place in a travel diary during the same days in which they wore the GPS and accelerometer.

Using the time-stamp as a common identifier, accelerometer, GPS, and travel diary data were combined into a "LifeLog", which is an individual-level master table for each participant. One record in the LifeLog represents a 30-second time-stamp with corresponding accelerometer counts, GPS latitude, longitude, and speed, and travel diary place or trip characteristics. Detailed methods and description for creating the LifeLog were documented in previous studies [101,108].

Processing GPS and accelerometer data with PALMS and Merging it with the LifeLog

PALMS allows researchers to specify analytical parameter settings for identifying trips and places. In this study, PALMS version R4 default parameters were used to process GPS and accelerometer data from the sixty subjects. PALMS calculates the distance and speed between sequential GPS points. Subsets of GPS points were flagged as being members of trips if they spanned ≥100m within an interval of 180 seconds. Trips with a 90th percentile speed of ≥25

$km \cdot h^{-1}$ were categorized as driving. Trips with a 90th percentile speed $\geq 10\ km \cdot h^{-1}$ and $<25\ km \cdot h^{-1}$ were classified as bicycling. Finally, trips with a 90th percentile speed $\geq 1\ km \cdot h^{-1}$ and $<10\ km \cdot h^{-1}$ were identified as walking. Stationary places between trips were identified as having a duration of $\geq$ 300 seconds with GPS points within a 30 m radius.

To generate complete tables (i.e., with one GPS measurement per accelerometry epoch), PALMS imputes GPS points that were missing due to signal loss by duplicating records at the mean coordinate of the 20 records collected before signal loss (see Meseck et al. 2016 [126] for detailed information on imputation of GPS data). For the 60 participants over the one-week measurement period, there were 1,712,721 measured records, and PALMS imputed 177,779 GPS records (10.4%). Output from PALMS included original variables (XY coordinate, speed, accelerometry counts), as well as calculated trip and place variables. For comparison with the LifeLog data, PALMS was configured to use 30s time-stamps.

Results from PALMS were exported as CSV format files. We linked PALMS trip and place information with the LifeLog using subject ID, date, and time stamp as common identifiers. Incomplete travel diary records (e.g., missing trip start or end time, 0.01%), trips with an unspecified travel mode (0.3%), trips taken on ferries (0.002%), and tours (trips recorded with the same start and end location, 0.3%) were removed from the data set. Consequently, 987,550 observations of GPS points with complete PALMS and LifeLog data (99.4% of merged data) remained.

GPS point level analyses

PALMS and travel diary data were first compared at the GPS point level to determine the amount of agreement between PALMS and travel diary identified trips, places, and travel mode.

For trip/place identification comparison, two variables were created for each observation: one indicating whether the PALMS algorithm classified the GPS point as part of a trip or place, and another indicating whether the travel diary classified the observation as part of a trip or place. These two variables were used in a 2×2 contingency table to calculate trip classification agreement at the GPS point level.

For travel mode identification comparison, three variables were created for each observation. PALMS classifies trips into one of three travel mode categories (pedestrian, bicycle, or vehicle), whereas the travel diary included 14 options: 1 (auto/truck/van), 2 (carpool/vanpool), 3 (bus), 4 (light rail), 5 (monorail/trolley), 6 (heavy rail), 7 (dial-a-ride/paratransit), 8 (school bus), 9 (ferry), 10 (taxi/shuttle bus/limousine), 11 (motorcycle/moped), 12 (bicycle), 13 (walk), 14 (airplane). For this analysis, we recoded the ten motor vehicle-based travel modes in the travel diary (1-8, 10, 11) as vehicle. These three categories were used in a 3×3 contingency table to calculate travel mode classification agreement.

The merged GPS point level data were examined at the subject level to compare the number of trips and places for each participant *i*. The average difference (D) between PALMS and travel diary was calculated in Equation (1), where x_{P_i}, x_{TD_i} are the number of trips (places) in PALMS *i* and travel diary *i*, and *n* is the number of study participants.

$$D = \frac{\sum_{i=1}^{n}\left(x_{P_i} - x_{TD_i}\right)}{n} \tag{1}$$

Trip level analyses

Trips recorded in the travel diary were used as the unit of analysis in a trip-level data set. PALMS and travel diary data were first tested for matching based on the overall number of trips, and more stringently, based on the temporal overlap between trips identified in these two sources. Each travel diary recorded trip had a unique trip number, together with information about starting time, ending time, duration of a trip, and travel mode. Trips from PALMS also had unique trip IDs and were matched with travel diary trips based on subject ID, date, and time.

Trips from PALMS and the travel diary were considered to match if they had any temporal overlap. We computed an agreement rate (AR_t) as the proportion of trips recorded in travel diaries that were matched to PALMS trips (Equation 2), where m_t is the total number of matching trips, and x_{TD_t} is the total number of trips from travel diaries.

$$AR_t = \frac{m_t}{x_{TD_t}} \times 100 \tag{2}$$

Next, because times reported in travel logs may not match device times precisely, we performed a sensitivity analysis by repeating the matching analysis using different proportions of overlap time. Specifically, we tested >0%, >25%, >50%, >75% and 100% of PALMS trip duration overlapping with travel diary trip duration. We computed the percentage of overlapped time (OT) between each PALMS and travel diary trip in Equation (3), where r is duration of temporal overlap between PALMS and travel diary, and t_{TD} is time duration of a trip recorded in travel diary.

$$OT = \frac{r}{t_{TD}} \times 100 \tag{3}$$

Lastly, trip-level OT data were aggregated to subject-level agreement rates (AR_i) using the five OT cut-off points of >0%, >25%, >50%, >75% and 100%, as an assessment of PALMS subject-level performance in Equation (4), where m_i is the total number of matching trips from participant *i*.

$$AR_i = \frac{m_i}{x_{TD_i}} \times 100 \tag{4}$$

To investigate factors associated with differential PALMS subject-level performance across the sixty participants, personal characteristics including primary travel mode and socioeconomic factors (e.g., age, sex, race, education, household income, children, car ownership) were taken into account. Two types of response variables were used in statistical modeling. First, subject-level agreement rates were used as a response variable. Second, since subject-level agreement rate is the proportion of correctly identified trips recorded in travel diary, we also investigated the relationship between the number of matching trips and participants' personal characteristics. Ordinary least squares and linear mixed effects models were used for the first response variable; negative binomial and mixed effects negative binomial models were applied for the second response variable.

Results

GPS point level performance

<u>*Classification of GPS points in PALMS and travel diary*</u>

PALMS classified 56.0% of trip observations in the travel diary as trips and 94.7% of place observations in travel diary as places (Table 9). The level of agreement between the two measures was also assessed by the inter-rater reliability test. Cohen's kappa statistic was 0.463 (p-value < 0.01) between PALMS and travel diary.

Table 9 Trip and place classification agreement at the GPS point level

		Travel Diary (LifeLog)			
		Trip		Place	
		GPS Point Counts	(%)	GPS Point Counts	(%)
PALMS	Trip	41,495	(56.0%)	48,016	(5.3%)
	Place	32,613	(44.0%)	860,970	(94.7%)
Total		74,108		908,986	

Table 10 shows the travel mode classification results across all participants at the GPS point level. The overall GPS point level agreement rate for travel mode match between PALMS and travel diary was 77.4%. Cohen's kappa statistic was 0.484 (p-value < 0.01). The agreement rate between PALMS and travel diary observations was higher for vehicle versus bicycling and walking.

Table 10 Travel mode classification at the GPS point level

		Travel Diary (LifeLog)					
		Vehicle		Bicycle		Walking	
		GPS Point Counts	(%)	GPS Point Counts	(%)	GPS Point Counts	(%)
PALMS	Vehicle	25,963	(84.5%)	1,024	(45.6%)	2,271	(26.6%)
	Bicycle	2,916	(9.5%)	1,202	(53.5%)	1,299	(15.2%)
	Walking	1,835	(6.0%)	20	(0.9%)	4,965	(58.2%)
Total		30,714		2,246		8,535	

<u>*Comparing trips and places identification by individual participant*</u>

Figure 8 shows the difference in the number of places and trips per day extracted from the travel diary and from PALMS by participant. Differences did not appear to be consistent across participants. PALMS found an equal or greater number of trips for 71.7% of the sample and fewer trips for 28.3% of the sample relative to the travel diary (Figure 8a). Also, PALMS identified an equal or greater number of places for 55% of the sample, and found fewer places among 45% of the sample (Figure 8b). On average, PALMS identified 3.1 more trips and 2.1 more places per day by participant during the seven consecutive days than were found in the travel diary.

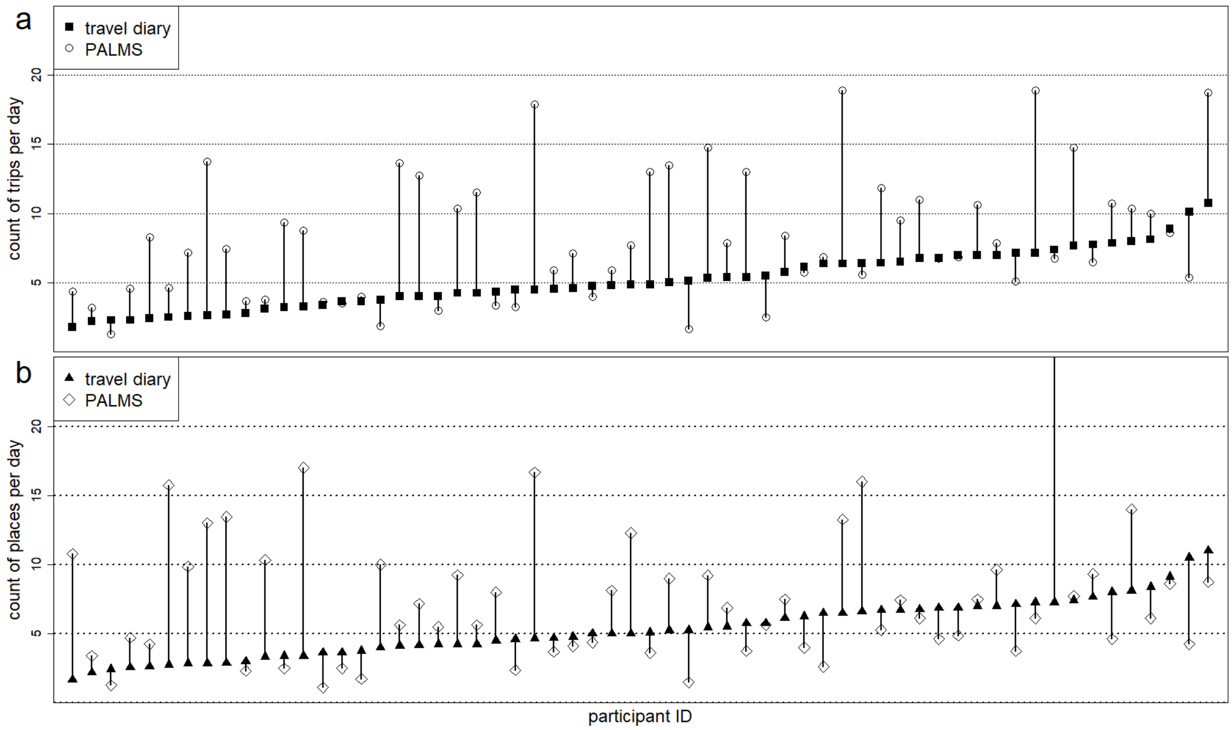


Figure 8 Count of trips (a) and places (b) per day by study participant

Comparing travel diary and PALMS data for one day of activity from two subjects

Visualization of travel diary and PALMS outcomes with corresponding maps provides some insight into the observed discrepancies. Figure 9 shows travel diary and PALMS data for one day of activity from two subjects. The upper and lower panels show an example of high and low agreement between travel diary and PALMS outcomes, respectively.

In the high agreement case (Figure 9a, Figure 9b), eight vehicular trips and two walking trips were reported in the travel diary; PALMS identified seven vehicular trips and two walking trips for this day for this subject. Based on these visuals, we could conclude that both travel diary and PALMS travel behavior identification algorithm worked well for this subject for this day.

In the low agreement case (Figure 9c, Figure 9d), the travel diary records included six vehicular trips and one walking trip. However, PALMS identified eight vehicular trips, seven bicycling trips, and nine walking trips within the same time period. Considering GPS speed and accelerometer count patterns between 09:00 and 14:00, and 19:00 to 00:00, this subject may have neglected recording some walking trips in travel diary. On the other hand, a single vehicular trip might have been identified as multiple vehicular and bicycling trips in PALMS. It is possible that vehicular trips with low speed were identified as bicycling trips in PALMS.

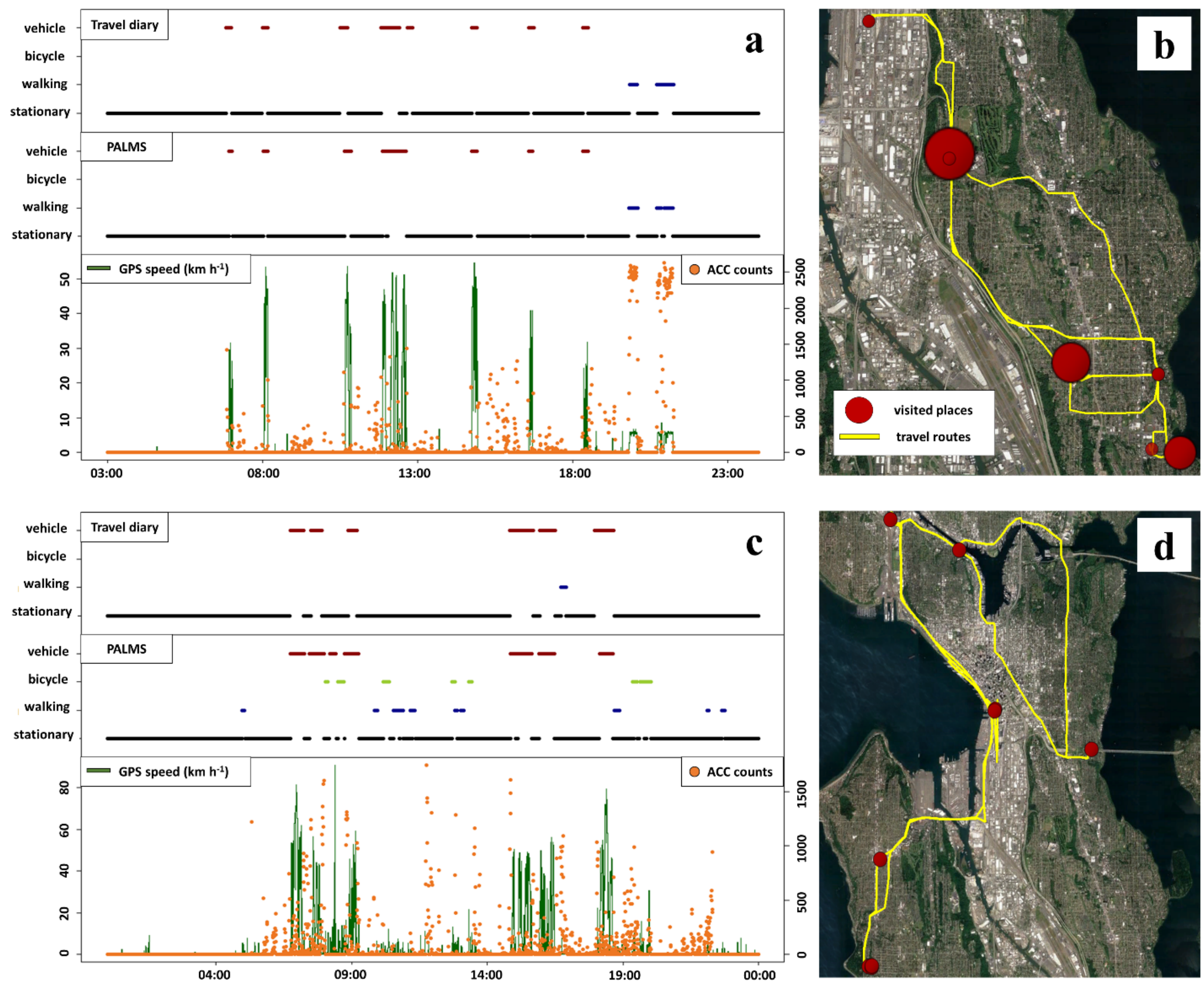


Figure 9 (a) and (c) show an example of high and low agreement between travel diary and PALMS outcomes, respectively. (b) and (d) represent places and travel routes identified by PALMS for each case. A bigger circle means high frequency of visit in the map

Trip level performance

Global Performance

There were 2,100 trips in the travel diaries of the sixty subjects across all modes. The days on which these trips were recorded had 2,483 trips across all modes derived from PALMS. 1,233 trips from the travel diary were identified by PALMS as trips (regardless of mode) (58.7%), 598 (28.5%) travel diary trips were not matched to a PALMS trip, and 269 (12.8%) travel diary trips were split into more than one trip in PALMS. Of the 598 travel diary trips that were not captured by PALMS, 56% were reported as vehicular trips, 41.1% were reported as walking, and the remaining 2.9% were reported as bicycle trips.

Figure 10 shows the agreement rates (AR_t) by varying the percentage of overlap time (OT) (0-100%) between PALMS and the travel diary. As expected, agreement rates were greater for lower values of OT percentage. The overall agreement rate (OT > 0%; where there was any temporal overlap between PALMS and travel diary trips) across the three travel modes was 46.4%. The agreement rate for vehicle, bicycle and pedestrian trips was 52.5%, 35.7%, and 34.8%, respectively.

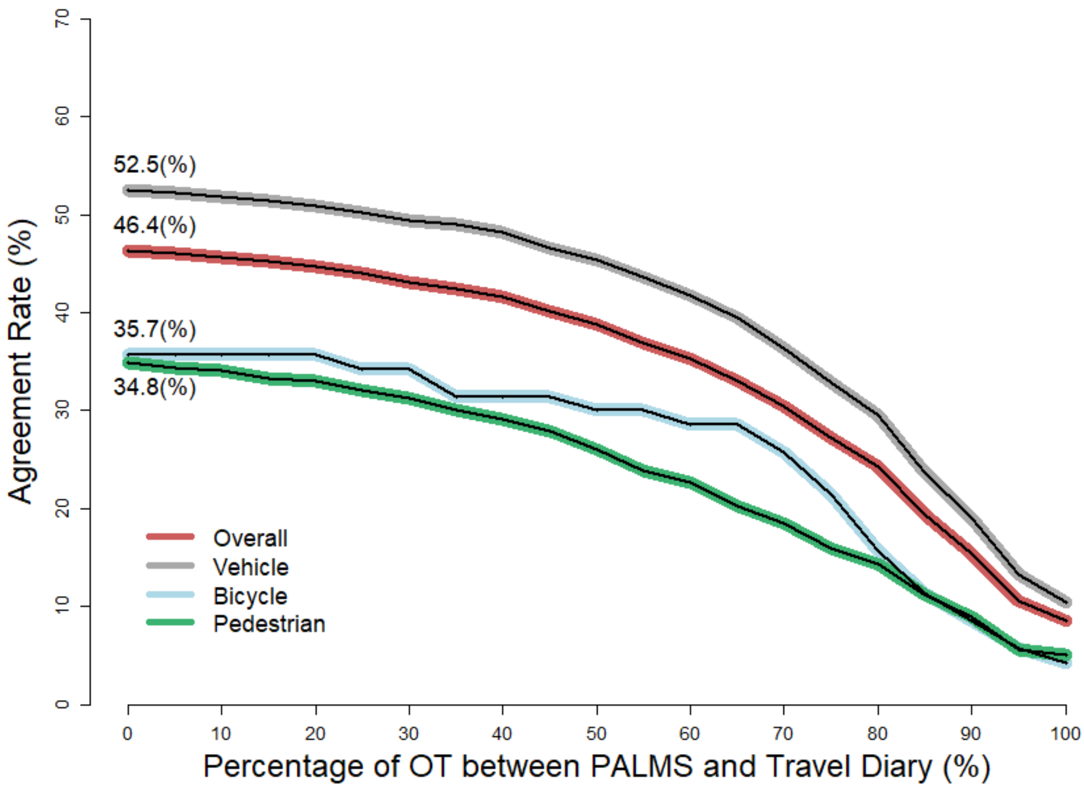


Figure 10 Agreement rate in the different modes by percentage of overlapped time (OT)

<u>Subject level performance</u>

To assess subject-level correspondence between PALMS and travel diary trip identification, we calculated agreement rates for each subject (AR_i). The total number of correctly identified trips per subject was first calculated by applying the same methods used in global performance analyses. This number of matched trips was divided by the total number of trips recorded in the subject's travel diary.

Figure 11 shows inter-subject variation in agreement rates based on having any OT. Of the 60 subjects, 10 (17%) had an agreement rate that was lower than 20%. Among these subjects, three had 0% agreement rates. Proportions of subjects that had an agreement rate of 20-40%, 40-

60%, 60-80%, 80-100% were 27.1%, 32.2%, 20.3% and 3.4%, respectively. Based on these subject-level agreement rates, the mean agreement rate was calculated as 42.4% (OT > 0%).

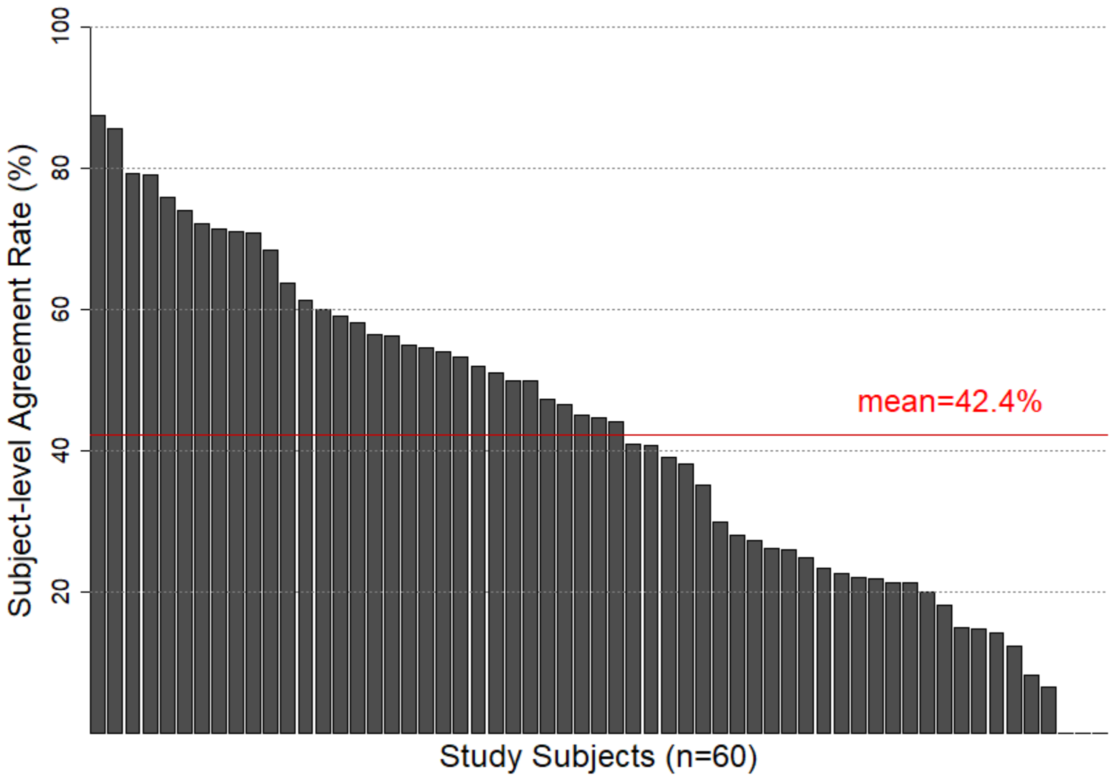


Figure 11 Subject-level agreement rates (OT > 0%)

Statistical modeling for PALMS subject-level performance

Four statistical models were applied to investigate variations in subject-level performance (OT > 0%) (Table 11). Models I and II show OLS and negative binomial modeling results. Also, since all subjects were clustered in three primary travel modes, we applied mixed-effects models to reflect hierarchical data structure in models III and IV. The average subject-level agreement rate was 59.4%, 20.5%, 48.2% for driver, transit user, and walker groups, respectively.

Car ownership was statistically significant and positively associated with PALMS subject-level performance in all models with 95% confidence interval except for model II (p-value=0.057). Primary travel mode was also statistically significant in models I and II. Being in the driver or walker group (with the transit user group as the reference category) was related to higher performance in PALMS. Primary travel mode was also used as a random intercept in mixed-effects models (III, IV). The intra-class correlation coefficient (ICC) was higher in model III. Other personal characteristics were not related to PALMS subject-level performance.

The hierarchical data structure was better explained in the linear mixed-effects model (model III), which is summarized in Figure 12. Bars that do not span 0 (e.g., travel mode and car ownership) indicate a significant relationship. For comparison, OLS model (model I) results were also included in the figure. With car ownership being the only statistically significant variable in model III, it was visualized using random intercepts by setting all other covariates to zero in Figure 13, which shows both of random intercepts and the fixed coefficient of car ownership. Agreement rates increased with car ownership for all subjects, but transit users had significantly lower agreement rates compared to the high drivers and high walkers.

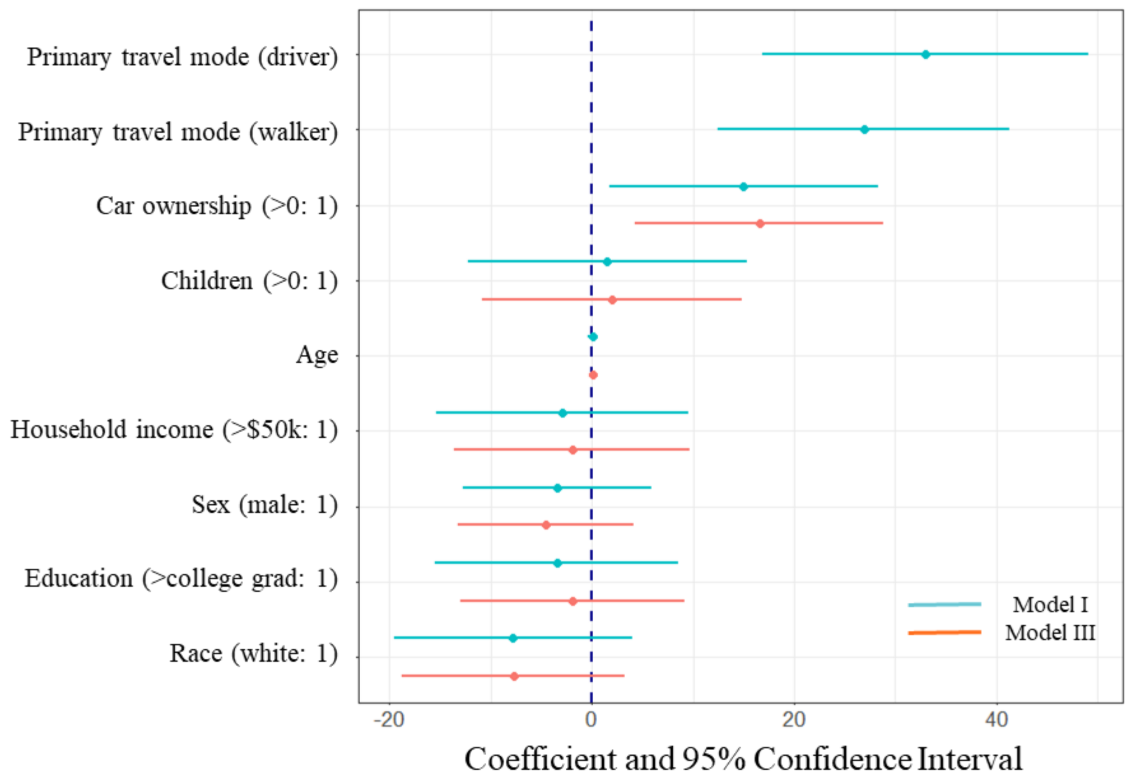


Figure 12 Coefficient and confidence interval plot for subject-level agreement rate regression (Model I and Model III)

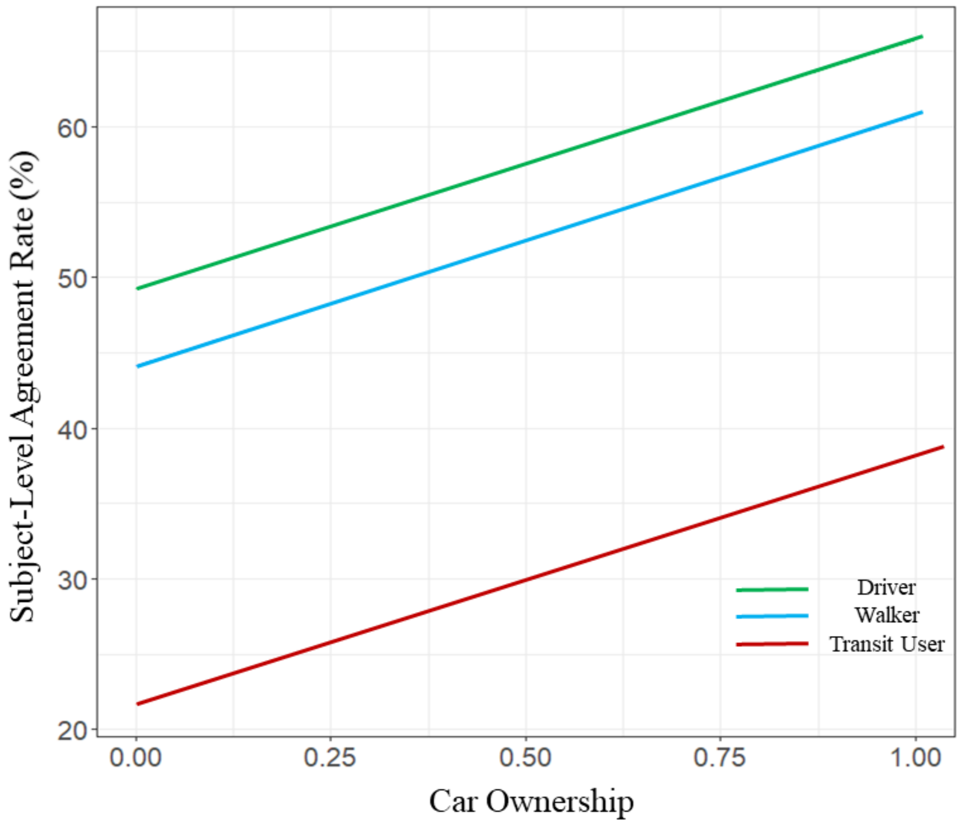


Figure 13 The effect of car ownership on subject-level agreement rate by primary travel mode

Table 11 Statistical modeling results

	Model I (OLS)			Model II (Negative Binomial)			Model III (Linear Mixed Effects)			Model IV (Mixed Effects Negative Binomial)		
Response Variable (PALMS subject-level performance; OT > 0%)	Subject-level Agreement Rate			# of Matching Trips			Subject-level Agreement Rate			# of Matching Trips		
	B	CI	p	IRR	CI	p	B	CI	p	IRR	CI	p
							Fixed Parts					
(Intercept)	20.78	-5.15 – 46.72	0.114	0.21	0.11 – 0.40	<.001	38.38	10.40 – 66.36	0.007	0.38	0.18 – 0.80	0.011
Age	0.07	-0.37 – 0.50	0.758	1	0.99 – 1.01	0.961	0.07	-0.32 – 0.47	0.719	1	0.99 – 1.01	0.907
Sex (ref. female)	-3.46	-12.85 – 5.93	0.462	0.95	0.79 – 1.15	0.622	-4.56	-13.04 – 3.92	0.292	0.93	0.76 – 1.13	0.469
Race (ref. non-white)	-7.78	-19.65 – 4.08	0.193	0.87	0.67 – 1.14	0.311	-7.73	-18.53 – 3.08	0.161	0.89	0.67 – 1.17	0.394
Education (ref. <college grad)	-3.46	-15.54 – 8.61	0.567	0.95	0.73 – 1.24	0.71	-1.9	-12.78 – 8.98	0.732	0.98	0.75 – 1.29	0.909
Household Income (ref. ≤$50k)	-2.92	-15.49 – 9.66	0.643	0.92	0.72 – 1.19	0.531	-1.93	-13.34 – 9.49	0.741	0.94	0.72 – 1.22	0.628
Children (ref. no children)	1.51	-12.36 – 15.38	0.828	1.02	0.77 – 1.36	0.869	2	-10.63 – 14.63	0.757	1.02	0.76 – 1.37	0.891
Car Ownership (ref. no car)	15.01	1.61 – 28.41	0.029	1.34	0.99 – 1.82	0.057	16.54	4.47 – 28.61	0.007	1.41	1.03 – 1.93	0.034
Primary Travel Mode (ref. transit user)							Random Parts					
driver	32.96^{***}	16.78 – 49.14	<.001	2.62	1.74 – 3.97	<.001		$\sigma^2 = 219.324$			-	
walker	26.87^{***}	12.38 – 41.36	<.001	2.34	1.60 – 3.45	<.001		$\tau_{00} = 154.055$			$\tau_{00} = 0.138$	
-	-	-	-	-	-	-		N = 3			N = 3	
-	-	-	-	-	-	-		ICC = 0.413			ICC = 0.056	
Observations		57			57			57			57	
R^2 / adj. R^2		.581 / .501			-			R^2 / Ω_0^2 = .578 / .577			Deviance = 51.43	

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Discussion

This study compared PALMS place, trip, and travel mode identification results with travel diary data from sixty adults living in a mid-sized U.S. metropolitan area, the greater Seattle area, and selected for having different primary travel modes. Compared to previous GPS algorithm assessment studies [36,95,98,114], the study used a larger sample and examined travel mode agreement among 2,100 recorded trips with 987,500 GPS point observations. The two data sets offered two levels of analysis.

At the GPS point level analysis, PALMS trip and place classification agreement rates were quite different at 56.0% and 94.7%, respectively. The place identification agreement rate was similar to that of a study which compared PALMS outcomes with SenseCam images sampled every minute (93.4%) [114], but the trip identification agreement rate was lower (88.5%) [114]. Since the present study used the self-reported travel diary as reference data, the results cannot be directly compared with objectively captured photographic data. However, the difference between the two studies suggest that diary data for places is more precisely recorded than it is for trips. The selected population mixing primarily drivers, transit users, and walkers could also explain the different results.

Agreement rates varied significantly by travel mode. While the rate was high at 77.4% for all observations, they were lower for bicycle (53.5%) and walking trips (58.2%) compared to vehicular trips (84.5%) at the GPS point level. Notably, 45.6% of bicycling trip GPS points recorded in the travel diary were classified by PALMS as driving. Distinguishing between driving and bicycling can be difficult due to primary reliance on speed recorded in GPS data. In PALMS, the cut-off speed used for differentiating driving from bicycling was 25 $km \cdot h^{-1}$. Results would likely be different using higher or lower cut-off speeds. In addition, the number of observations for bicycling trips in our sample (2,246 GPS points from 70 trips) was relatively small compared to other travel modes (vehicle: 30,714 GPS points from 1,367 trips; walking: 8,535 GPS points from 663 trips). It is also likely that distinguishing between motor vehicle travel and bicycling is more challenging in higher traffic and congested areas due to lower vehicular speeds. The use of ancillary data, such as measured heart rate, which was not available for this study, would enhance the performance of PALMS for separating bicycle trips from driving [127].

The difficulties in identifying trips and travel modes were also confirmed through individual participant level analyses. PALMS identified an average of 23.9 more trips per subject than were recorded in the travel diary, and 71.7% of subjects had more trips identified by PALMS than were recorded in the travel diary. Disparities in trip counts were much higher than the discrepancies seen for PALMS place identification algorithm results (16.3 more places per subject on average; 55% of subjects with more places in PALMS versus the travel diary). Visualizing mapped data from two subjects helped points to known limitations in travel diary data, which are prone to errors and omissions in recorded trips [95,104,106]. Evidence of unrecorded trips can be found in GPS and accelerometer data. Also, many single trips as recorded in the travel diary were split into multiple trips by PALMS. However, these errors could be minimized by careful selection of parameter settings in PALMS and by varying tolerances for acceptable short-duration stops within trips.

Since the GPS point level analyses revealed some limitations, we reconstructed our data at the trip level and compared PALMS outcomes with the travel diary. We took a novel approach of reporting global and subject-level performance. In the global performance analyses, we found that 12.8% of trips recorded in the travel diary were split into multiple trips in PALMS. This result confirmed one of our speculations about why PALMS identified more trips than travel diary records. For trip level analyses, we took the most conservative perspective by excluding these divided trips from correctly identified trip group.

In the global performance analyses, we advanced the agreement analyses by introducing the percentage of OT between PALMS and travel diary. In the analyses, the overall agreement rate was lower (46.4%, OT > 0%) than the results from GPS point level analyses (77.4%). Since we used strict rules by removing divided trips from the correctly identified trip group, many GPS points previously labeled as matched points were classified as unmatched trips in this level of analysis.

The agreement rate between PALMS and travel diary travel modes at the trip level was highest for vehicular trips (52.5%, OT > 0%), consistent with GPS point level analysis results. Bicycling had the second highest agreement rate (35.7%, OT > 0%), with walking trip agreement rate being slightly lower (34.8%, OT > 0%). At the GPS point level, agreement for walking was higher than for bicycling (58.2% versus 53.5%). However, the agreement rate for vehicular trips was much higher (84.5%). This discrepancy is likely due to the parameters for speed used to identify walking, bicycling, driving, which were specified as 1-10 $km \cdot h^{-1}$, 10-25 $km \cdot h^{-1}$, and ≥ 25 $km \cdot h^{-1}$, respectively. Walking and bicycling usually do not exceed 25 $km \cdot h^{-1}$. However, motor vehicles often drive at 25 $km \cdot h^{-1}$ for extended intervals, particularly in urban areas with high traffic congestion. The overlap in actual speeds increases the difficulty in differentiating different travel modes. The agreement rate was also highest in vehicular trips in a previous study that assessed PALMS [114].

In the subject-level performance analyses, the overall agreement rate (42.4%, OT > 0%) was slightly lower than in the global performance analyses (46.4%, OT > 0%). Subject-level agreement rates were highly variable across subjects, with a range of 0% to 87.5%. To further investigate this discrepancy, we included primary travel mode and socioeconomic information as covariates in subject-level models.

Primary travel mode was found to be the strongest explanatory variable examined for the discrepancy between PALMS and travel diary travel modes. Subject-level agreement rates were highest among those identified primarily as drivers and lowest among primarily public transit users. A previous study showed that transit users have more walking trips than non-users, which will ostensibly affect their individual-level agreement rates [128]. Also, transit trips have spatial and temporal characteristics that may make them more difficult to identify: they have more short-duration stops in designated areas than driving trips; public transit vehicles mostly run along arterial roads with stop-and-go traffic signals and congestion; and they may be more subjected to the urban canyon effects than driving trips, which can lower the quality of GPS data [95]. Owning a car was also found to be significantly related to higher subject-level agreement rates, in line with the global performance analyses results which showed that PALMS can best detect vehicular trips. This was notable because car ownership was significant in these models even after accounting for the primary mode grouping (driver, transit user, walker) and the high level of vehicle ownership in the sample.

Overall, the results from this study showed lower agreement rates than previous studies [98,114] because we took a more conservative approach and applied strict rules to assess the algorithm. However, the detailed investigation methods will provide new information and help developers and users to improve their future algorithms. Our study also has limitations. First, there is currently no gold standard criterion measure for travel behavior identification, and it is likely that both the PALMS algorithm and the travel dairy had errors and omissions. Future studies could be aided by the use of additional objective reference data (e.g., heart rate monitoring for walking and bicycling) to evaluate the algorithm. Second, we used default parameter settings in PALMS for our analyses. Agreement rates for active travel modes may be improved through careful selection of PALMS parameters.

Third, the TRAC study participants were sampled from the greater Seattle area, which contains pockets of high-density development and experiences severe commute-time traffic congestion on expressways and on arterial roads. The quality of individual-level GPS data might be better in suburban and rural areas where the configuration of built environments is less complex and traffic congestion is less frequent. Fourth, the TRAC sample did not include people with physical disabilities, which affect the travel mode choice. Still, anyone can suffer temporary physical or mental disabilities during the study period. Future studies may consider these confounders to better explain the subject-level performance of PALMS. Fifth, although the descriptive statistics of study participants showed that the sample is not biased, there is a possibility of self-selection in survey samples. To mitigate the problem, we applied a stratified random sampling method to proportionally include under-represented travel mode groups (e.g., public transit user). The performance of PALMS needs to be tested in different geographic areas where a higher proportion of the population uses transit or non-motorized modes. Such areas will likely also have different built environments and socio-demographics. Lastly, compared to previous studies, our larger sample still did not provide ample power for statistical modeling. Future studies may include a larger population.

Conclusions

The use of objectively measured data from such devices as GPS and accelerometers has dramatically changed research methods in the transportation planning and public health fields during the last fifteen years. Automated algorithms have enabled researchers to process large data sets and to identify fine-scaled and active travel behaviors in less time and with fewer errors. This study investigated the performance of PALMS, a widely used travel behavior identification system.

For a thorough investigation, we used multiple methods to compare PALMS outcomes with travel diary data. In analyses at the GPS point level, PALMS showed moderate agreement for trips and travel modes. Trip level analyses yielded moderate agreement rates for vehicular trips, but walking and bicycling trips had lower agreement rates. At the subject level, agreement rates were lower for the public transit user group than for the driver and the walker groups. It appears that trips taken by individuals who are primarily transit users are more difficult to identify: transit users take more walking trips and their many transit trips occur in densely developed areas with more traffic congestion, stop-and-go traffic flow, and lower quality GPS data due to poor reception in canyon-like settings.

PALMS appears to perform well in identifying vehicular trips and corresponding travel routes, indicating that the algorithm can help reduce the burden of processing GPS data into travel behavior data. To enhance the overall performance of the algorithm, further efforts are needed to identify non-motorized travel modes. The use of shorter measurement intervals for GPS and accelerometer data, as well as adjustments to speed thresholds suggested as defaults in the algorithm, may improve differentiation between walking and other modes. Ancillary data sources, such as heart rate monitoring, may help differentiate active (walking, bicycling) and passive (vehicular) modes. Efforts to discriminate between public transit and automobile trips will further expand the usefulness of PALMS. Geographic information systems (GIS) data already in existence, such as transit routes, stops, and stations may play a strong role. Finally, systematic evaluation of different parameter settings in PALMS could lead to better detection of places versus trips and short-duration stops within trips.

PALMS is a promising tool that serves to link data capturing health and transportation behaviors. Its performance in detecting active travel modes will be improved with further testing of the different parameter settings. Also, testing and evaluating PALMS internationally in a range of urbanized areas where travel modes other than driving are common will broaden the applicability of the algorithm.

Chapter 5 Conclusion

Safe active transportation can be achieved and promoted by increasing the number of pedestrian-friendly environments.[14,129] Quantifying built environment attributes around crash-risk locations (e.g., residential density within intersection buffers) are popular in active transportation safety study. However, measuring the built environment characteristics along the actual walking path using objective data (e.g., residential density along the individual walking path identified by GPS data) may improve the research methods in the field. In this dissertation, analyses first used a location-based pedestrian safety approach that combined built environment and crash data to identify crash-risk locations. More specifically, unique pedestrian crash-risk locations at intersection and non-intersection areas were detected by introducing a new protocol. Second, the factors affecting pedestrian crashes were evaluated using fine-grained built environment data within crash-risk locations. Lastly, the automated travel behavior detection algorithm PALMS (Personal Activity Location Measurement System) was evaluated to explore the potential of GPS and accelerometer data for the measurement of built environments along the pedestrian route. These analyses document how the now increasingly used sensor-based mobility data calls for reliable methods to translate the objectively measured activity and location information into trips that are actually taken by pedestrians.

In Chapter 2, a protocol was introduced to provide a useful tool for identifying unique crash-risk locations at intersection and non-intersection areas. Since it takes a lot of time and effort for researchers to manually classify uncorrelated crash-risk locations (e.g., collecting exclusively located intersection data by conducting a field investigation or comparing with satellite maps) [19,82], the protocol developed in the study can reduce labor hours and human errors to clean the data. The reliability of the protocol was assessed in terms of sample size, coverage, and accuracy by comparing with pedestrian crash data as a reference. A set of search radii was tested to check the sensitivity of the results, and the protocol showed excellent performance in identifying crash-risk locations. Search radius was the main parameter of the protocol. Its dimension can be selected based on the study of interest. In general, the search radius of 20 m showed high performance with regard to sample size, coverage, and accuracy. However, the 40 m search radius can be a better option for a study requiring a larger sample at intersection locations. The protocol provides researchers with a cost-effective method to identify unique intersection and non-intersection locations. It is also an efficient and reliable method for creating spatial analysis units for pedestrian crash modeling, allowing researchers to make comparisons between studies. Further investigation is necessary to improve the performance of the protocol for identifying unique crash-risk locations. Testing more granular and subdivided search radii will refine the capture of intersection and non-intersection locations, leading to create better spatial analysis units for pedestrian crash modeling

In Chapter 3, the risk factors of pedestrian-motor vehicle crashes were evaluated and compared between intersection and non-intersection locations identified from Chapter 2. Key findings include: (1) Wide roadways and principal arterial classification on intersection locations showed high number of pedestrian collisions; (2) Proxy measures of pedestrian volume showed high number of collisions at both intersection and non-intersection locations, to include bus ridership, residential density, employment density, park and ride, residential area, and service area; (3) Household income showed robust protective effects of pedestrian collisions. Built environment factors of pedestrian collision identified from the study should be prioritized to reduce the number of pedestrian crashes with the limited resources. Estimating the number of pedestrian motor-vehicle collisions in a single all-location model may not capture interaction effects of location type (intersection/non-intersection) with BE variables (e.g., roadway width) [42]. In this study, based on the refined all-location model, two different models were developed for each crash-risk location type. Factors affecting pedestrian motor-vehicle collisions were different between two locations, which suggests different safety programs and strategies for each location type. Pedestrian volumes are found to be correlated with land use, development density, and street network in previous studies [130,131]. Although methods have been developed to estimate pedestrian volumes, they do not have a fine-grained geographic scale [130,132,133], which made it difficult to be applied in this study. Reliable data on pedestrian volumes at a small geographical scale are needed to improve the modeling results.

In Chapter 4, an automated travel behavior detection algorithm PALMS (Personal Activity Location Measurement System) was evaluated; the algorithm used GPS and accelerometer data from the Travel Assessment and Community (TRAC) project which were compared with travel diaries from the same study. Key findings include: (1) The algorithm showed moderate identification power at the GPS point level analyses. The overall agreement rate of travel mode detection was 77.4% between PALMS and travel logs; (2) The trip level analyses showed lower agreement rates (42.4 – 46.4%) between two measures than those of the GPS point level analyses. Using additional objective data (e.g., combining GPS with heart rate monitoring to measure physical activity [134]) may improve the performance of the algorithm. (3) The agreement rate for vehicular trip detection was higher than for bicycling and walking in all analyses. The performance of identifying active travel modes might be improved by further testing of the different algorithm parameter settings. (4) Study participants' primary travel mode and car ownership were significantly related to the agreement rates. Because GPS and accelerometer data help identify individual travel routes, and may lead to improving the detection and quantification of micro and macro environmental characteristics directly associated with individual travel, further studies should evaluating the algorithm in a range of urbanized areas where a higher proportion of the population uses transit or non-motorized travel modes will broaden the applicability of the algorithm. Overall, the improvement of tools for identifying individual travel modes and routes may offer a new opportunity for better understanding of individual exposures to the built environments.

The protocol developed to create consistent spatial analysis units for location-based pedestrian-motor vehicle crash models, makes it possible for empirical results to be comparable. Also, it is a time saver which provides researchers a cost-effective method to identify unique crash-risk locations. The dissertation also contributes to the literature on factors affecting pedestrian crashes by comparing differences in crash-risk at intersection and non-intersection locations and by providing advanced visualization of empirical model results which can be used to prioritize safety countermeasures according to the characteristics of potential crash locations. The dissertation also highlights the potential of device-collected mobility data from GPS and accelerometers for identifying individual travel modes and routes that can be used for identifying the built environments along the pedestrian route. Lastly, the methods developed to assess the PALMS algorithm can be generalized and can serve to evaluate different approaches to travel mode classification. Detecting active transportation using mobility big data remains a challenging task. However, it is a crucial component for innovating the research methods in active transportation safety.

References

1. Millward H, Spinney J, Scott D. Active-transport walking behavior: destinations, durations, distances. Journal of Transport Geography. 2013 Apr 1;28:101–10.

2. Ogilvie D, Foster CE, Rothnie H, Cavill N, Hamilton V, Fitzsimons CF, et al. Interventions to promote walking: systematic review. BMJ. 2007 Jun 7;334(7605):1204.

3. Khan M, M. Kockelman K, Xiong X. Models for anticipating non-motorized travel choices, and the role of the built environment. Transport Policy. 2014 Sep;35:117–26.

4. Kuzmyak JR, Walters J, Bradley M, Kockelman KM. Estimating Bicycling and Walking for Planning and Project Development: A Guidebook. NCHRP Report [Internet]. 2014 [cited 2019 Apr 13];(770). Available from: https://trid.trb.org/view/1320568

5. Sahlqvist S, Song Y, Ogilvie D. Is active travel associated with greater physical activity? The contribution of commuting and non-commuting active travel to total physical activity in adults. Preventive medicine. 2012;55(3):206–11.

6. Tudor-Locke C, Ainsworth BE, Popkin BM. Active commuting to school. Sports medicine. 2001;31(5):309–13.

7. Parkkari J, Natri A, Kannus P, Mänttäri A, Laukkanen R, Haapasalo H, et al. A controlled trial of the health benefits of regular walking on a golf course∗∗Access the “Journal Club” discussion of this paper at http://www. elsevier.com/locate/ajmselect/. The American Journal of Medicine. 2000 Aug 1;109(2):102–8.

8. Lee IM, Buchner DM. The importance of walking to public health. Med Sci Sports Exerc. 2008 Jul;40(7 Suppl):S512-8.

9. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, et al. The Physical Activity Guidelines for Americans. JAMA. 2018 Nov 20;320(19):2020–8.

10. Murtagh EM, Murphy MH, Boone-Heinonen J. Walking – the first steps in cardiovascular disease prevention. Curr Opin Cardiol. 2010 Sep;25(5):490–6.

11. Davison RCR, Grant S. Is Walking Sufficient Exercise for Health? Sports Medicine. 1993 Dec 1;16(6):369–73.

12. Tight M, Timms P, Banister D, Bowmaker J, Copas J, Day A, et al. Visions for a walking and cycling focussed urban transport system. Journal of Transport Geography. 2011 Nov;19(6):1580–9.

13. Pucher J, Dijkstra L. Promoting Safe Walking and Cycling to Improve Public Health: Lessons From The Netherlands and Germany. Am J Public Health. 2003 Sep 1;93(9):1509–16.

14. Quistberg DA, Howard EJ, Hurvitz PM, Moudon AV, Ebel BE, Rivara FP, et al. The Relationship Between Objectively Measured Walking and Risk of Pedestrian–Motor Vehicle Collision. Am J Epidemiol. 2017 May 1;185(9):810–21.

15. Quistberg DA, Howard EJ, Ebel BE, Moudon AV, Saelens BE, Hurvitz PM, et al. Multilevel models for evaluating the risk of pedestrian–motor vehicle collisions at intersections and mid-blocks. Accident Analysis & Prevention. 2015 Nov 1;84:99–111.

16. Moudon AV, Lin L, Jiao J, Hurvitz P, Reeves P. The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. Accident Analysis & Prevention. 2011 Jan 1;43(1):11–24.

17. Lee C, Abdel-Aty M. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accident Analysis & Prevention. 2005 Jul 1;37(4):775–86.

18. Pulugurtha SS, Sambhara VR. Pedestrian crash estimation models for signalized intersections. Accident Analysis & Prevention. 2011 Jan 1;43(1):439–46.

19. Schneider RJ, Diogenes MC, Arnold LS, Attaset V, Griswold J, Ragland DR. Association between Roadway Intersection Characteristics and Pedestrian Crash Risk in Alameda County, California. Transportation Research Record: Journal of the Transportation Research Board. 2010 Jan;2198(1):41–51.

20. Yu C-Y. How Differences in Roadways Affect School Travel Safety. Journal of the American Planning Association. 2015 Jul 3;81(3):203–20.

21. Zegeer CV, Bushell M. Pedestrian crash trends and potential countermeasures from around the world. Accident Analysis & Prevention. 2012 Jan 1;44(1):3–11.

22. Federal Highway Administration. Toolbox of Countermeasures and Their Potential Effectiveness for Pedestrian Crashes [Internet]. 2008. Available from: https://safety.fhwa.dot.gov/ped_bike/tools_solve/ped_tctpepc/index.cfm

23. Huang H, Song B, Xu P, Zeng Q, Lee J, Abdel-Aty M. Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of Transport Geography. 2016 Jun;54:248–56.

24. Abdel-Aty M, Siddiqui C, Huang H, Wang X. Integrating Trip and Roadway Characteristics to Manage Safety in Traffic Analysis Zones. Transportation Research Record. 2011 Jan 1;2213(1):20–8.

25. Rydin Y, Bleahu A, Davies M, Dávila JD, Friel S, Grandis GD, et al. Shaping cities for health: complexity and the planning of urban environments in the 21st century. The Lancet. 2012 Jun 2;379(9831):2079–108.

26. Milton J, Mannering F. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. :19.

27. Huang H, Abdel-Aty MA, Darwiche AL. County-Level Crash Risk Analysis in Florida: Bayesian Spatial Modeling. Transportation Research Record. 2010 Jan 1;2148(1):27–37.

28. Xie Z, Yan J. Kernel Density Estimation of traffic accidents in a network space. Computers, Environment and Urban Systems. 2008 Sep;32(5):396–406.

29. Mitra S. Spatial Autocorrelation and Bayesian Spatial Statistical Method for Analyzing Intersections Prone to Injury Crashes. Transportation Research Record: Journal of the Transportation Research Board. 2009 Dec 1;2136:92–100.

30. Guo F, Wang X, Abdel-Aty MA. Modeling signalized intersection safety with corridor-level spatial correlations. Accident Analysis & Prevention. 2010 Jan 1;42(1):84–92.

31. Lightstone AS, Dhillon PK, Peek-Asa C, Kraus JF. A geographic analysis of motor vehicle collisions with child pedestrians in Long Beach, California: comparing intersection and midblock incident locations. Injury Prevention. 2001 Jun 1;7(2):155–60.

32. Chen P. Built environment factors in explaining the automobile-involved bicycle crash frequencies: A spatial statistic approach. Safety Science. 2015 Nov;79:336–43.

33. Miranda-Moreno L, Fernandes D. Modeling of Pedestrian Activity at Signalized Intersections. Transportation Research Record: Journal of the Transportation Research Board. 2011 Dec 1;2264:74–82.

34. Miranda-Moreno LF, Morency P, El-Geneidy AM. The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections. Accident Analysis & Prevention. 2011 Sep 1;43(5):1624–34.

35. Kestens Y, Thierry B, Chaix B. Re-creating daily mobility histories for health research from raw GPS tracks: Validation of a kernel-based algorithm using real-life data. Health & place. 2016;40:29–33.

36. Thierry B, Chaix B, Kestens Y. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. International journal of health geographics. 2013;12(1):1.

37. Carlson JA, Jankowska MM, Meseck K, Godbole S, Natarajan L, Raab F, et al. Validity of PALMS GPS scoring of active and passive travel compared with SenseCam. Medicine and science in sports and exercise. 2015;47(3):662–7.

38. Pedestrian Traffic Fatalities by State: 2017 Preliminary Data | GHSA [Internet]. [cited 2018 Oct 17]. Available from: https://www.ghsa.org/resources/spotlight-pedestrians18

39. Moudon AV, Lin L, Jiao J, Hurvitz P, Reeves P. The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. Accident Analysis & Prevention. 2011 Jan 1;43(1):11–24.

40. Al-Ghamdi AS. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention. 2002 Nov;34(6):729–41.

41. Abdel-Aty M. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research. 2003 Jan 1;34(5):597–603.

42. Moore DN, Schneider WH, Savolainen PT, Farzaneh M. Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. Accident Analysis & Prevention. 2011 May 1;43(3):621–30.

43. Quistberg DA, Howard EJ, Ebel BE, Moudon AV, Saelens BE, Hurvitz PM, et al. Multilevel models for evaluating the risk of pedestrian–motor vehicle collisions at intersections and mid-blocks. Accident Analysis & Prevention. 2015 Nov;84:99–111.

44. Nie K, Wang Z, Du Q, Ren F, Tian Q, Nie K, et al. A Network-Constrained Integrated Method for Detecting Spatial Cluster and Risk Location of Traffic Crash: A Case Study from Wuhan, China. Sustainability. 2015 Mar 4;7(3):2662–77.

45. Wier M, Weintraub J, Humphreys EH, Seto E, Bhatia R. An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. Accident Analysis & Prevention. 2009 Jan;41(1):137–45.

46. 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. 2017;37(8):1459–76.

47. Hancock MW, Wright B. A Policy on Geometric Design of Highways and Streets. 2013.

48. Wang X, Abdel-Aty M, Brady PA. Crash Estimation at Signalized Intersections: Significant Factors and Temporal Effect. Transportation Research Record. 2006 Jan 1;1953(1):10–20.

49. Pietrucha MT, Opiela KS. Safe accommodation of pedestrians at intersections. Transportation Research Record. 1993;12–12.

50. Forsyth A, Van Riper D, Larson N, Wall M, Neumark-Sztainer D. Creating a replicable, valid cross-platform buffering technique: The sausage network buffer for measuring food and physical activity built environments. International Journal of Health Geographics. 2012 Jul 9;11(1):14.

51. Poch M, Mannering F, Member A. Negative Binomial Analysis of Intersection-Accident Frequencies. Journal of Transportation Engineering. 1996;105–113.

52. Forsyth A, Schmitz KH, Oakes M, Zimmerman J, Koepp J. Standards for Environmental Measurement Using GIS: Toward a Protocol for Protocols. Journal of Physical Activity and Health. 2006 Feb;3(s1):S241–57.

53. Shankar VN, Ulfarsson GF, Pendyala RM, Nebergall MB. Modeling crashes involving pedestrians and motorized traffic. Safety Science. 2003 Aug;41(7):627–40.

54. Satria R, Castro M. GIS Tools for Analyzing Accidents and Road Design: A Review. Transportation Research Procedia. 2016;18:242–7.

55. Ma L, Yan X, Qiao W. A Quasi-Poisson Approach on Modeling Accident Hazard Index for Urban Road Segments. Discrete Dynamics in Nature and Society. 2014;2014:1–8.

56. Aguero-Valverde J, Jovanis PP. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis & Prevention. 2006 May 1;38(3):618–25.

57. Song JJ, Ghosh M, Miaou S, Mallick B. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis. 2006 Jan;97(1):246–73.

58. Quddus MA. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. Accident Analysis & Prevention. 2008 Jul 1;40(4):1486–97.

59. Moudon AV, Kang M. Safe Main Street Highways Part II: Analyses of Collisions Involving Pedestrians and Bicyclists in Washington State. Washington State Department of Transportation, Office of Research & Library; 2017.

60. Litman T. Integrating Public Health Objectives in Transportation Decision-Making. American Journal of Health Promotion. 2003 Sep;18(1):103–8.

61. Liu C, Shen Q. An empirical analysis of the influence of urban form on household travel and energy consumption. Computers, Environment and Urban Systems. 2011 Sep;35(5):347–57.

62. Health USD of, Services H. Physical activity and health: a report of the Surgeon General. DIANE Publishing; 1996.

63. Giles-Corti B, Donovan RJ. Relative Influences of Individual, Social Environmental, and Physical Environmental Correlates of Walking. American Journal of Public Health. 2003 Sep;93(9):1583–9.

64. Rafferty AP, Reeves MJ, McGee HB, Pivarnik JM. Physical activity patterns among walkers and compliance with public health recommendations. Med Sci Sports Exerc. 2002 Aug;34(8):1255–61.

65. Gilmour H. Physically active Canadians. Health Reports. 2007 Aug;18(3):45–65.

66. Murtagh EM, Nichols L, Mohammed MA, Holder R, Nevill AM, Murphy MH. Walking to improve cardiovascular health: a meta-analysis of randomised control trials. The Lancet. 2014 Nov 19;384:S54.

67. Gregg EW, Gerzoff RB, Caspersen CJ, Williamson DF, Narayan KMV. Relationship of Walking to Mortality Among US Adults With Diabetes. Arch Intern Med. 2003 Jun 23;163(12):1440–7.

68. Orsini N, Mantzoros CS, Wolk A. Association of physical activity with cancer incidence, mortality, and survival: a population-based study of men. British Journal of Cancer. 2008 Jun;98(11):1864–9.

69. Lee I-M, Paffenbarger RS. Associations of Light, Moderate, and Vigorous Intensity Physical Activity with LongevityThe Harvard Alumni Health Study. Am J Epidemiol. 2000 Feb 1;151(3):293–9.

70. McTiernan A, Kooperberg C, White E, Wilcox S, Coates R, Adams-Campbell LL, et al. Recreational Physical Activity and the Risk of Breast Cancer in Postmenopausal Women: The Women's Health Initiative Cohort Study. JAMA. 2003 Sep 10;290(10):1331–6.

71. Huang HF, Stewart JR, Zegeer CV. Evaluation of Lane Reduction "Road Diet" Measures on Crashes and Injuries. Transportation Research Record: Journal of the Transportation Research Board. 2002 Jan;1784(1):80–90.

72. Ewing R, Dumbaugh E. The Built Environment and Traffic Safety: A Review of Empirical Evidence. Journal of Planning Literature. 2009 May;23(4):347–67.

73. Hadayeghi A, Shalaby AS, Persaud BN. Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. Accident Analysis & Prevention. 2010 Mar 1;42(2):676–88.

74. Zahabi SAH, Strauss J, Manaugh K, Miranda-Moreno LF. Estimating Potential Effect of Speed Limits, Built Environment, and Other Factors on Severity of Pedestrian and Cyclist Injuries in Crashes. Transportation Research Record. 2011 Jan 1;2247(1):81–90.

75. LaScala EA, Gerber D, Gruenewald PJ. Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. Accident Analysis & Prevention. 2000 Sep 1;32(5):651–8.

76. Washington S, Metarko J, Fomunung I, Ross R, Julian F, Moran E. An inter-regional comparison: fatal crashes in the southeastern and non-southeastern United States: preliminary findings. Accident Analysis & Prevention. 1999 Jan 1;31(1):135–46.

77. Li Z, Wang W, Liu P, Bigham JM, Ragland DR. Using Geographically Weighted Poisson Regression for county-level crash modeling in California. Safety Science. 2013 Oct 1;58:89–97.

78. Haynes R, Jones A, Kennedy V, Harvey I, Jewell T. District Variations in Road Curvature in England and Wales and their Association with Road-Traffic Crashes. Environ Plan A. 2007 May 1;39(5):1222–37.

79. Ukkusuri S, Miranda-Moreno LF, Ramadurai G, Isa-Tavarez J. The role of built environment on pedestrian crash frequency. Safety Science. 2012 Apr;50(4):1141–51.

80. Levine N, Kim KE, Nitz LH. Spatial analysis of Honolulu motor vehicle crashes: II. Zonal generators. Accident Analysis & Prevention. 1995 Oct 1;27(5):675–85.

81. Chen P, Zhou J. Effects of the built environment on automobile-involved pedestrian crash frequency and risk. Journal of Transport & Health. 2016 Dec;3(4):448–56.

82. Poch M, Mannering F. Negative Binomial Analysis of Intersection-Accident Frequencies. Journal of Transportation Engineering. 1996 Mar;122(2):105–13.

83. MacNab YC. Bayesian spatial and ecological models for small-area accident and injury analysis. Accident Analysis & Prevention. 2004 Nov;36(6):1019–28.

84. Kang B, Scully JY, Stewart O, Hurvitz PM, Moudon AV. Split-Match-Aggregate (SMA) algorithm: integrating sidewalk data with transportation network data in GIS. International Journal of Geographical Information Science. 2015 Mar 4;29(3):440–53.

85. Hurvitz PM, Moudon AV. Home Versus Nonhome Neighborhood: Quantifying Differences in Exposure to the Built Environment. American Journal of Preventive Medicine. 2012 Apr 1;42(4):411–7.

86. Moudon AV, Sohn DW, Kavage SE, Mabry JE. Transportation-Efficient Land Use Mapping Index (TELUMI), a Tool to Assess Multimodal Transportation Options in Metropolitan Regions. International Journal of Sustainable Transportation. 2011 Jan 4;5(2):111–33.

87. Ryan S, Frank L. Pedestrian Environments and Transit Ridership. Journal of Public Transportation. 2009 Mar;12(1):39–57.

88. Cervero R. Walk-and-Ride: Factors Influencing Pedestrian Access to Transit. Journal of Public Transportation [Internet]. 2001 Jan 1;3(4). Available from: https://scholarcommons.usf.edu/jpt/vol3/iss4/1

89. Noland RB, Quddus MA. Analysis of Pedestrian and Bicycle Casualties with Regional Panel Data. Transportation Research Record: Journal of the Transportation Research Board. 2004 Jan;1897(1):28–33.

90. Laflamme L, Diderichsen F. Social differences in traffic injury risks in childhood and youth—a literature review and a research agenda. Injury Prevention. 2000 Dec 1;6(4):293–8.

91. LaScala EA, Gruenewald PJ, Johnson FW. An ecological study of the locations of schools and child pedestrian injury collisions. Accident Analysis & Prevention. 2004 Jul 1;36(4):569–76.

92. Koh PP, Wong YD. Comparing pedestrians' needs and behaviours in different land use environments. Journal of Transport Geography. 2013 Jan 1;26:43–50.

93. Palmer JR, Espenshade TJ, Bartumeus F, Chung CY, Ozgencil NE, Li K. New approaches to human mobility: Using mobile phones for demographic research. Demography. 2013;50(3):1105–28.

94. Stopher PR, Greaves SP. Household travel surveys: Where are we going? Transportation Research Part A: Policy and Practice. 2007;41(5):367–81.

95. Gong H, Chen C, Bialostozky E, Lawson CT. A GPS/GIS method for travel mode detection in New York City. Computers, Environment and Urban Systems. 2012;36(2):131–9.

96. Wolf J, Guensler R, Bachman W. Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data. Transportation Research Record: Journal of the Transportation Research Board. 2001;1768:125–34.

97. Kahlmeier S, Racioppi F, Cavill N, Rutter H, Oja P. "Health in all policies" in practice: guidance and tools to quantifying the health effects of cycling and walking. Journal of Physical Activity and Health. 2010;7(s1):S120–5.

98. Ellis K, Godbole S, Marshall S, Lanckriet G, Staudenmayer J, Kerr J. Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms. Frontiers in Public Health [Internet]. 2014 Apr;2(36). Available from: https://www.frontiersin.org/article/10.3389/fpubh.2014.00036

99. Cavill N, Rutter H, Hill A. Action on cycling in primary care trusts: results of a survey of Directors of Public Health. Public health. 2007;121(2):100–5.

100. Kerr J, Duncan S, Schipperjin J. Using global positioning systems in health research: a practical approach to data collection and processing. American Journal of Preventive Medicine. 2011;41(5):532–40.

101. Kang B, Moudon AV, Hurvitz PM, Reichley L, Saelens BE. Walking Objectively Measured: Classifying Accelerometer Data with GPS and Travel Diaries. Med Sci Sports Exerc. 2013 Jul;45(7):1419–28.

102. Cetateanu A, Luca B-A, Popescu AA, Page A, Cooper A, Jones A. A novel methodology for identifying environmental exposures using GPS data. International Journal of Geographical Information Science. 2016;1–17.

103. Lubans DR, Boreham CA, Kelly P, Foster CE. The relationship between active travel to school and health-related fitness in children and adolescents: a systematic review. International Journal of Behavioral Nutrition and Physical Activity. 2011;8(1):5.

104. Kelly P, Krenn P, Titze S, Stopher P, Foster C. Quantifying the difference between self-reported and global positioning systems-measured journey durations: a systematic review. Transport Reviews. 2013;33(4):443–59.

105. Wesolowski A, Stresman G, Eagle N, Stevenson J, Owaga C, Marube E, et al. Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. Scientific reports. 2014;4:5678.

106. Scully JY, Vernez Moudon A, Hurvitz PM, Aggarwal A, Drewnowski A. GPS or travel diary: Comparing spatial and temporal characteristics of visits to fast food restaurants and supermarkets. PloS one. 2017;12(4):e0174859.

107. Bricka SG, Sen S, Paleti R, Bhat CR. An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. Transportation research part C: emerging technologies. 2012;21(1):67–88.

108. Hurvitz P, Moudon A, Kang B, Saelens B, Duncan G. Emerging Technologies for Assessing Physical Activity Behaviors in Space and Time. Frontiers in Public Health [Internet]. 2014 Jan;2(2). Available from: https://www.frontiersin.org/article/10.3389/fpubh.2014.00002

109. Paz-Soldan VA, Reiner Jr RC, Morrison AC, Stoddard ST, Kitron U, Scott TW, et al. Strengths and weaknesses of Global Positioning System (GPS) data-loggers and semi-structured interviews for capturing fine-scale human mobility: findings from Iquitos, Peru. PLoS Negl Trop Dis. 2014;8(6):e2888.

110. Shareck M, Kestens Y, Gauvin L. Examining the spatial congruence between data obtained with a novel activity location questionnaire, continuous GPS tracking, and prompted recall surveys. International Journal of Health Geographics. 2013;12(1):40.

111. Krenn PJ, Titze S, Oja P, Jones A, Ogilvie D. Use of global positioning systems to study physical activity and the environment: a systematic review. American journal of preventive medicine. 2011;41(5):508–15.

112. Stopher P, FitzGerald C, Zhang J. Search for a global positioning system device to measure person travel. Transportation Research Part C: Emerging Technologies. 2008;16(3):350–69.

113. Zhou JJ, Golledge R. Real-time tracking of activity scheduling/schedule execution within a unified data collection framework. Transportation Research Part A: Policy and Practice. 2007;41(5):444–63.

114. Carlson JA, Jankowska MM, Meseck K, Godbole S, Natarajan L, Raab F, et al. Validity of PALMS GPS scoring of active and passive travel compared with SenseCam. Medicine and science in sports and exercise. 2015;47(3):662–7.

115. Jankowska MM, Schipperijn J, Kerr J. A Framework For Using GPS Data In Physical Activity And Sedentary Behavior Studies. Exerc Sport Sci Rev. 2015 Jan;43(1):48–56.

116. Vazquez-Prokopec GM, Stoddard ST, Paz-Soldan V, Morrison AC, Elder JP, Kochel TJ, et al. Usefulness of commercially available GPS data-loggers for tracking human movement and exposure to dengue virus. International journal of health geographics. 2009;8(1):68.

117. Rodríguez DA, Brown AL, Troped PJ. Portable global positioning units to complement accelerometry-based physical activity monitors. Medicine and science in sports and exercise. 2005;37(11 Suppl):S572-81.

118. Almanza E, Jerrett M, Dunton G, Seto E, Pentz MA. A study of community design, greenness, and physical activity in children using satellite, GPS and accelerometer data. Health & place. 2012;18(1):46–54.

119. Stopher P, FitzGerald C, Xu M. Assessing the accuracy of the Sydney Household Travel Survey with GPS. Transportation. 2007;34(6):723–41.

120. Chen C, Gong H, Lawson C, Bialostozky E. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transportation Research Part A: Policy and Practice. 2010;44(10):830–40.

121. Schuessler N, Axhausen K. Processing raw data from global positioning systems without additional information. Transportation Research Record: Journal of the Transportation Research Board. 2009;(2105):28–36.

122. Bohte W, Maat K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. Transportation Research Part C: Emerging Technologies. 2009;17(3):285–97.

123. Shen L, Stopher PR. Review of GPS Travel Survey and GPS Data-Processing Methods. Transport Reviews. 2014 May 4;34(3):316–34.

124. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. International Journal of Behavioral Nutrition and Physical Activity [Internet]. 2015 Dec [cited 2018 Nov 12];12(1). Available from: http://www.ijbnpa.org/content/12/1/159

125. Hurvitz PM, Moudon AV, Kang B, Fesinmeyer MD, Saelens BE. How Far from Home? The Locations of Physical Activity in an Urban US Setting. Preventive medicine. 2014;

126. Meseck K, Jankowska MM, Schipperijn J, Natarajan L, Godbole S, Carlson J, et al. Is missing geographic positioning system data in accelerometry studies a problem, and is imputation the solution? Geospatial health. 2016;11(2).

127. Brage S, Brage N, Franks P, Ekelund U, Wareham N. Reliability and validity of the combined heart rate and movement sensor Actiheart. European journal of clinical nutrition. 2005;59(4):561.

128. Saelens BE, Vernez Moudon A, Kang B, Hurvitz PM, Zhou C. Relation Between Higher Physical Activity and Public Transit Use. Am J Public Health. 2014 May;104(5):854–9.

129. O′Hern S, Oxley J. Understanding travel patterns to support safe active transport for older adults. Journal of Transport & Health. 2015 Mar 1;2(1):79–85.

130. Yin L, Cheng Q, Wang Z, Shao Z. ‘Big data’ for pedestrian volume: Exploring the use of Google Street View images for pedestrian counts. Applied Geography. 2015 Sep 1;63:337–45.

131. Hajrasouliha A, Yin L. The impact of street network connectivity on pedestrian volume. Urban Studies. 2015 Oct 1;52(13):2483–97.

132. Schneider RJ, Arnold LS, Ragland DR. Pilot Model for Estimating Pedestrian Intersection Crossing Volumes. Transportation Research Record. 2009 Jan 1;2140(1):13–26.

133. Landis BW. Bicycle system performance measures. ITE Journal (Institute of Transportation Engineers). 1996;66(2):18–26.

134. Duncan JS, Badland HM, Schofield G. Combining GPS with heart rate monitoring to measure physical activity in children: A feasibility study. Journal of Science and Medicine in Sport. 2009 Sep;12(5):583–5.