STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

TR0003 (REV 10/98)

Lock Data on Form

ADA Notice

For individuals with sensory disabilities, this document is available in alternate formats. For information call (916) 654-6410 or TDD (916) 654-3880 or write Records and Forms Management, 1120 N Street, MS-89, Sacramento, CA 95814.

1. REPORT NUMBER CA18-2452	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER

4. TITLE AND SUBTITLE Pedestrian Safety Improvement Program: Phase 2	5. REPORT DATE January 31, 2018 6. PERFORMING ORGANIZATION CODE
7. AUTHOR Julia Griswold, Aditya Medury, Louis Huang, David Amos, Jiajian Lu, Robert Schneider, and Offer Grembek	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS UC Berkeley Safe Transportation Research & Education Center 2614 Dwight Way, #7374 Berkeley, CA 94720-7374	10. WORK UNIT NUMBER 11. CONTRACT OR GRANT NUMBER 65A0547
12. SPONSORING AGENCY AND ADDRESS California Department of Transportation Division of Research and Innovation and System Information, MS-83 1727 30th Street Sacramento CA 95816	13. TYPE OF REPORT AND PERIOD COVERED Final Report 14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The Pedestrian Safety Improvement Program is an effort of the California Department of Transportation (Caltrans) to identify and address problems with regard to pedestrian safety in California, with the long-term goal of substantially reducing pedestrian fatalities and injuries in California. The efforts and findings presented in this report reflect the work of a team of experts in transportation engineering, transportation planning, public health, geographic information systems, and urban design from the UC Berkeley Safe Transportation Research & Education Center. In particular, Pedestrian Safety Improvement Program: Phase 2 is focused on three distinct areas pertaining to Caltrans' pedestrian safety monitoring program: (i) pedestrian exposure modeling, (ii) contextualized hotspot development, and (iii) pedestrian safety toolkit development.

17. KEY WORDS	18. DISTRIBUTION STATEMENT No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161

19. SECURITY CLASSIFICATION *(of this report)* Unclassified	20. NUMBER OF PAGES 69	21. COST OF REPORT CHARGED N/A

Reproduction of completed page authorized.

DISCLAIMER STATEMENT

This document is disseminated in the interest of information exchange. The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California or the Federal Highway Administration. This publication does not constitute a standard, specification or regulation. This report does not constitute an endorsement by the Department of any product described herein.

For individuals with sensory disabilities, this document is available in Braille, large print, audiocassette, or compact disk. To obtain a copy of this document in one of these alternate formats, please contact: the Division of Research and Innovation, MS-83California Department of Transportation, P.O. Box 942873, Sacramento, CA 94273-0001

Pedestrian Safety Improvement Program

Phase 2

Final Technical Report

Prepared by the University of California
Safe Transportation Research and Education Center

for the

California Department of Transportation

January 31, 2018

Acknowledgments

Many thanks go to Rachel Carpenter, Ted Davini, and Mary Hartegan of Caltrans for reaching out to local agencies on our behalf to request pedestrian count data. We are grateful to the many agencies and organizations that responded to our data request, including Alameda County Transportation Commission, City of Menlo Park, City of Rialto, City of Riverside Public Works, City of Roseville, City of San Jose, City of San Luis Obispo, City of Santa Ana, City of Solvang, City of South Gate, County of San Luis Obispo Department of Public Works, County of Tuolumne Community Resources Agency, El Dorado County Transportation Commission, Los Angeles County Department of Public Health, Newport Beach Department of Public Works, Orange County Transportation Authority, Presidio Trust, San Bernardino Association of Governments, San Bernardino Regional Parks, San Luis Obispo Council of Governments, San Mateo County Department of Public Works, Santa Clara Valley Transportation Authority, Santa Cruz County Regional Transportation Commission, Southern California Association of Governments, Tahoe Regional Planning Agency, Town of Los Gatos, Transportation Agency for Monterey County, Trinity County, Tulare County Association of Governments, and Willdan Engineering. We also thank Eric Fischer for sharing his Github repositories of count data scraped from municipal websites and Melanie Curry for spreading the word on Streetsblog.

We would especially like to thank Rachel Carpenter, John Ensch, Dean Samuelson, Robert Kim, Joel Retanan, Jerry Kwong, and other Caltrans staff members for providing constructive and invaluable guidance and comments to improve our work and to make sure it can be applied to improve pedestrian safety in California.

TABLE OF CONTENTS

Chapter 1. Introduction

The Pedestrian Safety Improvement Program is an effort of the California Department of Transportation (Caltrans) to identify and address systemic problems of pedestrian safety in California, with the long-term goal of substantially reducing pedestrian fatalities and injuries in California. The efforts and findings presented in this report reflect the work of a team of experts in traffic engineering, transportation planning, public health, geographic information systems, and urban design from the UC Berkeley Safe Transportation Research & Education Center.

The effort is well-timed. Available data indicate that pedestrians are 37 times more vulnerable than the rest of roadway users in California—that is, they suffer 37 times more injuries than they inflict on others. Additionally, while California has seen major gains in traffic safety over the last few years, these gains disproportionately reflect improvements in passenger vehicle safety. For example, while there was a nearly 10% decrease in overall traffic fatalities from 2007-2016, the gains were mostly realized for motorized modes (19% reduction in fatalities) but pedestrian deaths increased by 33. If we compare trends from 2010 (the end of the most recent financial crisis) there is an overall increase across all fatalities of about 33%, but again, this reflects a 29% increase in fatalities of motorized modes and an increase of 44% for pedestrian fatalities. Thus, pedestrians need more protection and investment but receive less of both than motorized users.

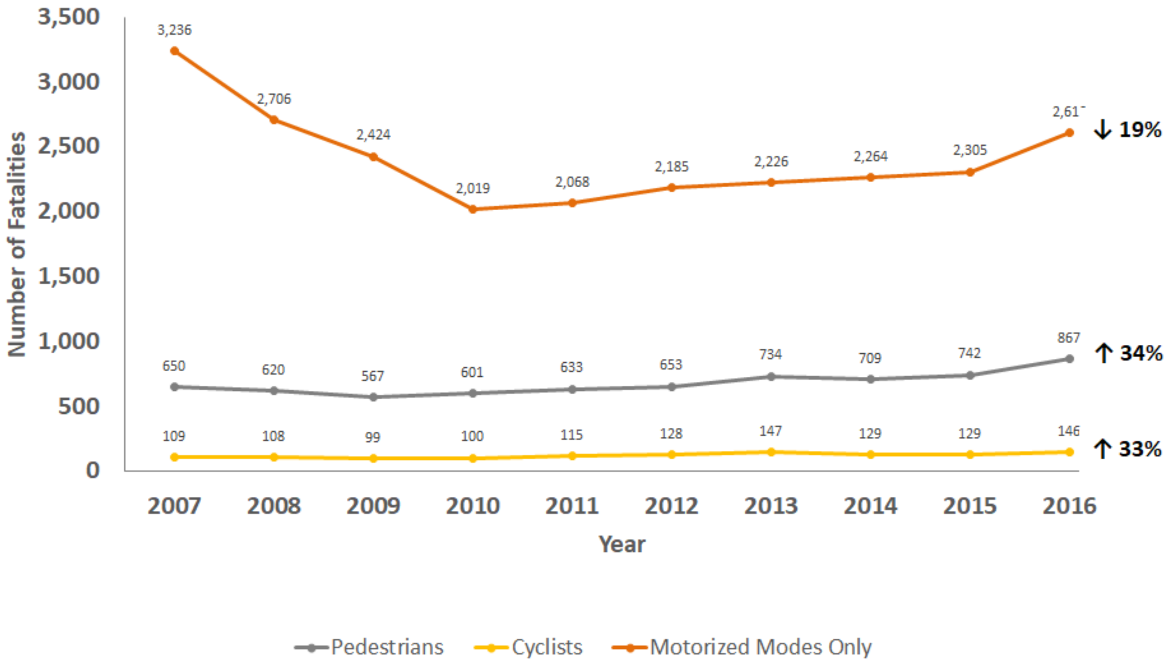


Figure 1-1. Fatal crash trends in California (2007-2016)

This report represents an effort to provide the knowledge and identify the resources needed to address this imbalance between pedestrians and motorized roadway users in California. The approach presented here is intentionally pragmatic, aiming not for an ideal plan, but for one that can help Caltrans and the State make gradual progress toward goals to improve pedestrian safety in California.

1.1. Key Components

The report is divided into three core chapters (excluding this chapter) that describe the overall project and findings. In particular, these chapters focus on three distinct focus areas pertaining to Caltrans' pedestrian safety monitoring program: (i) pedestrian exposure modeling, (ii) contextualized hotspot development, and (iii) pedestrian safety toolkit development. Collectively, these three areas of emphasis balance the need for ready-to-use, user-friendly decision-support tools for identifying pedestrian high collision concentration locations (HCCLs) in the near term, while also making significant advances that lay the foundation for more statistically rigorous network screening approaches in future extensions of this work:

Chapter 2 – Pedestrian Exposure Model describes the process to develop a pedestrian exposure model for the California State Highway System (SHS), explains the scope of the model in application to the SHS and summarizes the annual volume estimates.

Chapter 3 – Contextualized Hotspot Clustering describes a clustering approach to develop a pedestrian crash typology for the state highway system and evaluates the distribution of the proposed crash types within the crash population as well as within pedestrian HCCLs.

Chapter 4 – Pedestrian Safety Toolkit summarizes the enhancements made to the pedestrian safety monitoring report tool, along with the modifications made to the crash data import and SWITRS matching processes.

Overall, this report represents a tremendous amount of analysis and exploration of pedestrian safety in California. It is hoped that this analysis will provide Caltrans and stakeholders with the information they need to address current challenges and develop plans to continue progress in the future.

Chapter 2. Pedestrian Exposure Model

Pedestrian volume data are important for safety analysis because they can be used as a basic measure of exposure at a specific location. For example, the relative risk of pedestrian crashes for people traveling along state highways can be estimated as the number of pedestrian crashes per million pedestrian crossings. Further, using pedestrian volume as a variable in safety performance functions can show which roadway design features or other characteristics of a location should be modified to reduce pedestrian crashes and injuries. Volume data can also be used to identify how common pedestrian activity is on the State Highway System, indicating the importance of designing roadways for safe and convenient pedestrian access.

It is impractical to count pedestrians at every intersection and along every segment of the 15,000-mile State Highway System on a routine basis. This problem can be addressed by applying statistical models to estimate volumes at specific locations.

This chapter describes the process to develop a pedestrian exposure model for the California State Highway System. First, as described in the following section, we conducted a literature review of previous pedestrian models, the methods used, and potential explanatory variables. In the Model Development section, we describe the chosen explanatory variables, the processing of the count data, and the model estimation. The final section explains the scope of the model in application to the SHS and summarizes the annual volume estimates.

2.1. Background and Literature Review

This section summarizes existing research on pedestrian volume models, highlighting variables that could be used to estimate pedestrian volumes at specific locations on the California State Highway System. Existing literature includes pedestrian volume studies from California as well as other parts of North America.

2.1.1. Previous Pedestrian Demand Models

NCHRP Report 770 (1) provides a summary of pedestrian demand modeling research, highlighting three general categories of models that can provide facility-level volume estimates at roadway intersections and pedestrian network segments.

- "Trip generation and flow" models. This approach estimates the number of pedestrian trips between small areas, such as block faces or pedestrian analysis zones. These models follow a traditional traveling modeling approach, since they estimate trip generation, trip distribution, and network assignment. Ultimately, trips assigned to the pedestrian network are totaled for specific roadway crossings and sidewalk segments. One study applied the traditional travel model approach to block-sized pedestrian analysis zones in central Baltimore (2), and a similar approach is being developed using small grid cell pedestrian analysis zones in Portland (3).
- "Network simulation" models. This category of models, including Space Syntax, develops volume estimates for each part of a pedestrian network based on network characteristics such as connectivity and sight lines. In some cases, these network variables are combined with land use

variables to estimate pedestrian volumes (4,5). However, these models are often complex, take time, and require special programs time to apply.

- "Direct demand" models. These models estimate pedestrian volumes along roadway segments and intersections using site and surrounding area characteristics. Street block face or mid-block count data have been used to model pedestrian volumes in New York, NY (6), Milwaukee, WI (7), and Minneapolis, MN (8,9). However, more recent direct demand pedestrian volume models have been developed from intersection crossing counts (10-17). Intersection crossing counts can provide a more direct representation of pedestrian exposure for safety analyses.

We recommended the direct demand modeling approach for Caltrans because these models are simple to understand, do not require complex computer applications to execute, and are straightforward to apply. Typical steps used in the direct demand approach are listed below.

1. Pedestrian counts are taken at a sample of locations in a community. These counts are often collected manually over short periods of time, but automated detection techniques that collect data over weeks, months, or even years can also be used.
2. Short-period counts may be expanded to represent annual volume estimates (annual volume estimates can be compared with crash data that is reported on a yearly basis).
3. The annual (or other duration) pedestrian volumes are used as the dependent variable in a predictive model. Statistical software is used to identify significant relationships between the pedestrian volumes at each study location and explanatory variables describing the characteristics of the study location (e.g., land use characteristics, transportation system features, demographic factors, or any other factors thought to be relevant to pedestrian volumes).

Finally, the preferred statistical model equation can be used to estimate pedestrian volumes in other locations throughout the community.**Error! Reference source not found.Error! Reference source not found.**summarizes several recent direct demand pedestrian volume models. All were developed from counts collected within specific geographic areas, so they may not provide accurate pedestrian volume estimates in other communities. Many of these models are based on short counts (ranging from 2 to 12 hours) and are only appropriate for estimating pedestrian volumes during the specific times of day (e.g., afternoon peak period) or seasons of the year when the counts were taken. However, models developed for San Francisco (16) and the California State Highway System (17) are based on counts that were extrapolated to annual volumes, so they produce estimates of full-year pedestrian volumes. Many early applications of this approach used linear regression modeling. While simple to understand, this approach can produce unrealistic, negative volume estimates, so most recent studies have used loglinear and negative binomial model structures.

Table 2-1. Direct Demand Pedestrian Volume Models

General information		Pedestrian Count Information				Statistically-Significant Explanatory Variables				Model Information	
Model Location	**Source**	**Locations Used for Model**	**Pedestrian Count Description**	**Type of Count Sites**	**Count Period(s) Used for Model**	**Land Use**	**Transportation System**	**Socioeconomic Characteristics**	**Other**	**Model Output**	**Model Type**
Charlotte, NC	UNC Charlotte (Pulugurtha & Repaka 2008) (10)	176	Pedestrians counted each time they arrived at the intersection from any direction	Signalized intersections	7 am-7 pm	• Pop. within 0.25 mi. • Jobs within 0.25 mi. • Mixed land use within 0.25 mi. • Urban residential area within 0.25 mi.	• Number of bus stops within 0.25 mi.			Total pedestrians approaching intersections from 7 am to 7 pm (shorter periods also modeled)	Linear
Alameda County, CA	UC Berkeley SafeTREC (Schneider, Arnold, & Ragland 2009) (11)	50	Pedestrians counted every time they crossed a leg of the intersection (pedestrians within 50 feet of the crosswalk were counted)	Signalized and unsignalized arterial and collector roadway intersections	Tu, W, or Th: 12-2 pm or 3-5 pm; Sa: 9-11 am, 12-2 pm, or 3-5 pm	• Population within 0.5 mi. • Employment within 0.25 mi. • Commercial properties within 0.25 mi.	• BART (regional transit) station within 0.1 mi.			Total pedestrian crossings at arterial and collector roadway intersections during a typical week	Linear
San Francisco, CA	San Francisco State (Liu & Griswold 2009) (12)	63	Pedestrians counted each time they crossed a leg of the intersection (no distance to crosswalk specified)	Signalized and unsignalized intersections	Weekdays 2:30-6:30 pm	• Population density (net) within 0.5 mi. • Employment density (net) within 0.25 mi. • Patch richness density within 0.063 mi. • Residential land use within 0.063 mi.	• MUNI (light-rail transit) stop density within 0.38 mi. • Presence of bike lane at intersection		• Mean slope within 0.063 mi.	Total pedestrian crossings at intersections from 2:30-6:30 pm on typical weekday	Linear
Santa Monica, CA	Fehr & Peers (Haynes *et al*. 2010) (13)	92	Pedestrians counted each time they crossed a leg of the intersection (no distance to crosswalk specified)	Signalized and unsignalized intersections	Weekdays 5-6 pm	• Employment density within 0.33 mi. • Within a commercially-zoned area	• Afternoon bus frequency • Average speed limit on the intersection approaches		• Distance from Ocean	Total pedestrian crossings at intersections from 5-6 pm on typical weekday	Linear

General information		Pedestrian Count Information				Statistically-Significant Explanatory Variables				Model Information	
Model Location	**Source**	**Locations Used for Model**	**Pedestrian Count Description**	**Type of Count Sites**	**Count Period(s) Used for Model**	**Land Use**	**Transportation System**	**Socioeconomic Characteristics**	**Other**	**Model Output**	**Model Type**
San Diego, CA	Alta Planning + Design (Jones *et al.* 2010) (14)	80	Pedestrians counted each time they arrived at the intersection from any direction	Signalized and unsignalized intersections (includes some trail/roadway intersections)	Weekdays 7-9 am	• Population density within 0.25 mi. • Employment density within 0.5 mi. • Presence of retail within 0.5 mi.	• Greater than 6,000 transit ridership at bus stops within 0.25 mi. • 4 or more Class I bike paths within 0.25 mi.	• More than 100 households without vehicles within 0.5 mi.		Total pedestrians approaching intersections from 7 am to 9 am	Loglinear
Montreal, Quebec	McGill University (Miranda-Moreno & Fernandes 2011) (15)	1018	Pedestrians counted each time they crossed a leg of the intersection (no distance to crosswalk specified)	Signalized intersections	Weekdays 6-9 am, 11 am-1 pm, and 3:30-6:30 pm	• Population within 400 m • Commercial space within 50 m • Open space within 150 m • Schools within 400 m	• Subway within 150 m • Bus station within 150 m • % Major arterials within 400 m • Street segments within 400 m • 4-way intersection		• Distance to downtown • Daily high temp. >32°C	Total pedestrian crossings at intersections over 8 count hours (shorter periods also modeled)	Loglinear (also used Negative binomial)
San Francisco, CA	UC Berkeley SafeTREC (Schneider, Henry, Mitman, Stonehill, & Koehler 2012) (16)	50	Pedestrians counted every time they crossed a leg of the intersection (pedestrians within 50 feet of the crosswalk were counted)	Signalized and unsignalized intersections	Tu, W, or Th: 4-6 pm, extrapolated to annual volumes	• Households within 0.25 mi. • Employment within 0.25 mi. • Within high-activity zone (with parking meters) • Within 0.25 mi. of university campus	• Intersection is controlled by a traffic signal		• Maximum slope of any approach leg	Total pedestrian crossings at intersections during a full year	Loglinear
Minneapolis, MN	University of Minnesota (Hankey *et al*. 2012) (8)	259	Pedestrians were counted each time they crossed a screenline in the middle of a block	Midblock locations along sidewalks and multi-use trails	September 12-hour (6:30 am-6:30 pm) counts and 2 hour counts (4-6 pm) extrapolated to 12-hour counts	• Distance to the central business district	• Intersection is on a principal arterial roadway • Intersection is on an arterial roadway • Intersection is on a collector roadway	• Percent of neighborhood residents who are non-White • Percent of neighborhood residents with a college education	• Distance to the nearest body of water • Precipitation during count period	Total pedestrians using roadway and trail segments during a 12-hour period in September	Negative binomial

General information		Pedestrian Count Information				Statistically-Significant Explanatory Variables				Model Information	
Model Location	**Source**	**Locations Used for Model**	**Pedestrian Count Description**	**Type of Count Sites**	**Count Period(s) Used for Model**	**Land Use**	**Transportation System**	**Socioeconomic Characteristics**	**Other**	**Model Output**	**Model Type**
California	UC Berkeley SafeTREC (Grembek *et al*. 2014) (17)	66	Pedestrians counted each time they crossed a leg of the intersection (no distance to crosswalk specified)	Intersections along urban arterials in the State Highway System	Various two-hour and four-hour periods on weekdays and weekends, extrapolated to annual volumes	• Households within 0.1 mi.	• Sum of mainline and minor street vehicle volumes • Number of lanes on minor street			Total pedestrian crossings at intersections during a full year	Loglinear
Minneapolis, MN	University of Minnesota (Hankey & Lindsey 2016) (9)	471	Pedestrians were counted each time they crossed a screenline in the middle of a block	Midblock locations along sidewalks and multi-use trails	September days: 4-6 pm	• Population density within 1250 m network buffer • Retail area within 200 m network buffer • Open space area within 1000 m network buffer	• Transit stops within 1000 m network buffer • Off-street trail within 3000 m network buffer (only in 1 of 2 models) • Major roads within 200 m network buffer (only in 1 of 2 models)			Total pedestrians using roadway and trail segments from 4-6 pm on September days	Loglinear

2.1.2. Potential Explanatory Variables

Using the direct demand modeling approach, pedestrian volumes are assumed to be a function of the characteristics at and around specific locations on the California State Highway System. These characteristics will be represented by a set of explanatory variables. Previous research suggests explanatory variables representing land use, transportation system, socioeconomic, and several other characteristics are associated with pedestrian volumes. While there are many possible pedestrian model inputs, some explanatory variables are easier than others to gather statewide. For example, population density is provided by the U.S. Census Bureau's American Community Survey at the block group level for the entire country, so this information would be relatively easy to obtain for any location along the State Highway System. In contrast, there are no statewide databases of commercial property locations (this information has been gathered in previous studies through special requests to county tax assessors). Lists of potential explanatory variables and the assumed ease of collecting these variables are provided in Table 2-2, Table 2-3, and Table 2-4. Ease of collecting each variable is classified into the following categories:

- Easy. Data are available statewide from an existing data source. The variable can be created through basic GIS analysis.
- Moderate. Data are available for most or all of the state from existing data sources, but the data may be in different formats in different jurisdictions. The variable may require more sophisticated GIS analysis to create.
- Difficult. Data are not available from existing data sources. Field data collection or manual data collection from aerial or street-level imagery may be needed to create the variable.

Table 2-2. Potential Pedestrian Volume Model Inputs and Ease of Data Collection: Land Use Variables

Variable	Study (buffer area used)	Ease of Collection
Population within a given distance	Charlotte, NC (10) (0.25 mi.); Alameda County (11) (0.5 mi.); Montreal, QC (15) (400 m)	Easy (American Community Survey block group data)
Population density within a given distance	Minneapolis, MN (9) (1250 m network buffer); San Francisco (12) (0.5 mi.); San Diego County (14) (0.25 mi.)	Easy (American Community Survey block group data)
Jobs within a given distance	Charlotte, NC (10) (0.25 mi.); Alameda County (11) (0.25 mi.); San Francisco (16) (0.25 mi.)	Easy (Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics block data)
Employment density within a given distance	San Francisco (12) (0.25 mi.); Santa Monica (13) (0.33 mi.); San Diego County (14) (0.5 mi.)	Easy (Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics block data)
Households within a given distance	San Francisco (16) (0.25	Easy (American Community Survey

Variable	Study (buffer area used)	Ease of Collection
	mi.); California (17) (0.1 mi.)	block group data)
Schools within a given distance	Montreal, QC (15) (400 m)	Easy (California Department of Education GIS data)
Presence of retail within 0.5 mi.	San Diego County (14)	Moderate (County tax assessor parcel data; need to look to each jurisdiction)
Commercial space within a given distance	Minneapolis, MN (9); 200 m network buffer); Montreal, QC (15) (50 m)	Moderate (County tax assessor parcel data; jurisdiction-specific)
Commercial properties within a given distance	Alameda County (11) (0.25 mi.)	Moderate (County tax assessor parcel data; jurisdiction-specific)
Residential land use within a given distance	San Francisco (12) (0.063 mi.)	Moderate (County tax assessor parcel data; jurisdiction-specific)
Urban residential area within a given distance	Charlotte, NC (10) (0.25 mi.)	Moderate (County tax assessor parcel data; jurisdiction-specific and also need to define “urban”)
Within a commercially zoned area	Santa Monica (13)	Moderate (County tax assessor parcel data; jurisdiction-specific)
Open Space within a given distance	Minneapolis, MN (9); (1000 m network buffer); Montreal, QC (15) (150 m)	Moderate (County tax assessor parcel data; jurisdiction-specific and also need to define “open space”)
Distance to central business district (CBD) or downtown	Minneapolis, MN (8); Montreal, QC (15)	Moderate (US Census Bureau GIS data; need to define CBD location(s) within each region)
Mixed land use within a given distance	Charlotte, NC (10) (0.25 mi.)	Difficult (County tax assessor parcel data; jurisdiction-specific and also requires complex calculation)
Patch richness density within a given distance	San Francisco (12) (0.063 mi.)	Difficult (Requires complex calculation and multiple data sources)

Table 2-3. Potential Pedestrian Volume Model Inputs and Ease of Data Collection: Transportation System Variables

Variable	Study (buffer area used)	Ease of Collection
Street segments within a given distance	Montreal, QC (15) (400 m)	Easy (US Census GIS data or Caltrans GIS data)
4-way intersection	Montreal, QC (15)	Easy (US Census GIS data or Caltrans GIS data)
Signalized intersection	San Francisco (16)	Easy (Caltrans TASAS data)

Variable	Study (buffer area used)	Ease of Collection
Major arterials or major roadways within a given distance	Minneapolis, MN (9) (200 m network buffer); Montreal, QC (15) (400 m)	Moderate (US Census GIS data or Caltrans TASAS data; need reliable classification of arterial roadways)
Intersection is on a principal arterial roadway	Minneapolis, MN (8)	Moderate (US Census GIS data or Caltrans TASAS data; need reliable classification of arterial roadways)
Intersection is on an arterial roadway	Minneapolis, MN (8)	Moderate (US Census GIS data or Caltrans TASAS data; need reliable classification of arterial roadways)
Intersection is on a collector roadway	Minneapolis, MN (8)	Moderate (US Census GIS data or Caltrans TASAS data; need reliable classification of arterial roadways)
Regional transit station within a given distance	Alameda County (11) (0.1 mi.); Montreal, QC (15) (150 m)	Moderate (Metropolitan Planning Organization or Regional Transit Authority; jurisdiction-specific)
Number of transit stops within a given distance	Minneapolis, MN (9) (1000 m network buffer); Charlotte, NC (10) (0.25 mi.)	Moderate (Metropolitan Planning Organization or Regional Transit Authority; jurisdiction-specific)
Bus station within a given distance	Montreal, QC (15) (150 m)	Moderate (Metropolitan Planning Organization or Regional Transit Authority; jurisdiction-specific)
Sum of mainline and minor street vehicle volumes	California (17)	Moderate (Caltrans TASAS data; needs to be connected from segments to intersections and may not be available for all roads)
Number of lanes on minor street	California (17)	Moderate (Caltrans TASAS data; needs to be connected from segments to intersections and may not be available for all roads)
Average speed limit on the intersection approaches	Santa Monica (13)	Moderate (Caltrans TASAS data; needs to be connected from segments to intersections and may not be available for all roads)
Multi-use trail density	Minneapolis, MN (9) (3000 m network buffer)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
4 or more multi-use trails within a given distance	San Diego County (14) (0.25 mi.)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)
Presence of bike lane at intersection	San Francisco (12)	Difficult (Pedestrian and bicycle facility inventories do not exist statewide)

Variable	Study (buffer area used)	Ease of Collection
Afternoon bus frequency	Santa Monica (13)	Difficult (Metropolitan Planning Organization or Regional Transit Authority; jurisdiction-specific and requires frequency in addition to location data)
Greater than 6,000 transit ridership at bus stops within 0.25 mi.	San Diego County (14)	Difficult (Metropolitan Planning Organization or Regional Transit Authority; jurisdiction-specific and requires ridership in addition to location data)
Parking meters on at least one approach to intersection ("high-activity zone")	San Francisco (16)	Difficult (Parking facility inventories do not exist statewide)
MUNI (light rail) stop density within a given distance	San Francisco (12) (0.38 mi.)	Location specific (San Francisco)

Table 2-4. Potential Pedestrian Volume Model Inputs and Ease of Data Collection: Socioeconomic and Other Variables

Variable	Study (buffer area used)	Ease of Collection
Socioeconomic Variables		
Percent of neighborhood residents who are non-White	Minneapolis, MN (8)	Easy (American Community Survey block data)
Percent of neighborhood residents with a college education	Minneapolis, MN (8)	Easy (American Community Survey block data)
More than 100 households without vehicles within a given distance	San Diego County (14) (0.5 mi.)	Easy (American Community Survey block data)
Other Variables		
Mean slope within a given distance	San Francisco (12) (0.063 mi.)	Easy (US Geological Survey National Elevation Dataset)
Maximum slope of any intersection approach	San Francisco (16)	Easy (US Geological Survey National Elevation Dataset)
Distance to ocean	Santa Monica (13)	Easy (US Census Bureau GIS data)
Distance to the nearest body of water	Minneapolis, MN (8)	Easy (US Census Bureau GIS data)
Precipitation during count period	Minneapolis, MN (8)	Easy (National Oceanic and Atmospheric Administration weather data or data collector records)
Daily high temperature > 32C	Montreal, QC (15)	Easy (National Oceanic and Atmospheric Administration weather data)

2.2. Model Development

As previouslymentioned, the direct demand approach that we selected assumes that pedestrian volumes are a function of the built environment and demographic attributes of the surrounding area. To develop the predictive model, we needed to collect data for and process the explanatory variables that describe the surrounding area and the dependent variable, or annual pedestrian intersection volumes. These efforts are described in the next two subsections and followed by explanation of the model estimation process.

2.2.1. Explanatory Variables

The complete list of explanatory variables is shown in Table 2-5. For variables that pertain to an area around the intersection, such as population, the value was calculated at three different buffer distances– half-mile, quarter-mile, and tenth-mile. The scale of these variables is described as “buffer” in the scale column of Table 2-5. Other variables are related to the specific attributes of the intersection location and are described as “intersection” in Table 2-5. Explanation for how we calculated each variable is described below.

Table 2-5. Explanatory variables

Description	Scale	Data Source
Demographics		
Population	Buffer	U.S. Census ACS
Number of households	Buffer	U.S. Census ACS
Population that is white alone	Buffer	U.S. Census ACS
Number of walk commuters	Buffer	U.S. Census ACS
Number of transit commuters	Buffer	U.S. Census ACS
Number of households with no vehicle	Buffer	U.S. Census ACS
Number of college degree holders	Buffer	U.S. Census ACS
Percent of population that is white alone	Buffer	U.S. Census ACS
Walk commute mode share	Buffer	U.S. Census ACS
Transit commute mode share	Buffer	U.S. Census ACS
Percent of households with no vehicle	Buffer	U.S. Census ACS
Percent of population with a college degree	Buffer	U.S. Census ACS
Infrastructure		
Intersection is on a principal arterial	Intersection	CRS
Intersection is on a minor arterial	Intersection	CRS
Intersection is on a collector street	Intersection	CRS
Four-way intersection	Intersection	CRS
Intersection has a signal	Intersection	Open Street Map
Network Connectivity		
Number of meters of streets	Buffer	U.S. Census TIGER

Description	Scale	Data Source
Number of street segments	Buffer	U.S. Census TIGER
Transit		
All Transit Metric	Buffer	Center for Neighborhood Technology
Number of jobs within 30 minutes on transit	Buffer	Center for Neighborhood Technology
Number of transit commuters	Buffer	Center for Neighborhood Technology
Number of transit trips per week	Buffer	Center for Neighborhood Technology
Number of routes	Buffer	Center for Neighborhood Technology
Employment/Land Use		
Employment square footage of foot traffic land uses	Buffer	ESRI Business Analyst
Number of employees	Buffer	ESRI Business Analyst
Climate		
Number of days with more than 0.5 inch rain	Intersection	CA Energy Commission
Number of days with more than 1 inch rain	Intersection	CA Energy Commission
Number of days with more than 10 inches snow	Intersection	CA Energy Commission
Number of days with temp over 90F	Intersection	CA Energy Commission
Other		
Distance to water body	Intersection	ESRI Business Analyst
Number of schools	Buffer	CA Dept of Education
Maximum slope of intersecting road segment	Intersection	Google Elevation

2.2.1.1. Population and demographics

We used U.S. Census American Community Survey data to develop the demographic variables. The five-year dataset, collected from 2011 to 2015, provides sample-based estimates at the block group level; block groups are smaller than tracts but larger than blocks. The spatial resolution worked well for our tenth, quarter-mile, and half-mile buffer distances.

We collected demographic data on race, education, households, and commute mode. For each attribute, we calculated both the total and the percent of the block group population.

Our analysis required us to take the Census data and analyze it spatially, near the count locations. We used the Census GIS shapefile and joined the columns in Table 2-6 to it.

Table 2-6. Census variables used in analysis

Variable	Census Description
B02001e1	Race: Total: Total population -- (Estimate)
B02001e2	Race: White alone: Total population -- (Estimate)
B08301e1	Means of Transportation to Work: Total: Workers 16 years and over -- (Estimate)

B08301e19	Means of Transportation to Work: Walked: Workers 16 years and over -- (Estimate)
B08301e10	Means of Transportation to Work: Public transportation (excluding taxicab): Workers 16 years and over -- (Estimate) (includes bus, streetcar, subway, railroad, and ferryboat)
B11001e1	Household Type (Including living alone): Total: Households -- (Estimate)
B15003e1	Educational Attainment for the Population 25 Years and Over: Total: Population -- (Estimate)
B15003e22	Educational Attainment for the Population 25 Years and Over: Bachelors degree -- (Estimate)
B25044e1	Tenure by Vehicles Available: Total: Occupied housing units -- (Estimate)
B25044e3	Tenure by Vehicles Available: Owner occupied: No vehicle available: Occupied housing units -- (Estimate)
B25044e10	Tenure by Vehicles Available: Renter occupied: No vehicle available: Occupied housing units -- (Estimate)

We wrote a Python script using the ArcPy library to process the variables. For each buffer distance, the script clipped the block groups by the appropriate buffer, calculated the area of the clipped block groups, and then divided that area by the total area of the original block groups to determine the percentage of block groups that fall within the buffer. The calculations for the different types of variables were as follows:

- For the number variables, like population, the percentage was used to scale down the total block group population to an area-based estimate of the population that falls within the clipped block group. Summing the population estimates of the block groups by buffer, produced estimates of the total population falling within each buffer.
- For the percentage variables, like percent of population that is white alone, we followed the steps described above for both the numerator variable, white-alone population, and the denominator variable, total population. The final variable was the ratio of the two.

Table 2-7below lists the input variables used to make the calculations for each demographic variable.

Table 2-7. Input variables used for calculation for each Census variable

Variable Name	Numerator Variable	Denominator Variable
Population	B02001e1	n/a
Number of households	B11001e1	n/a
Population that is white alone	B02001e2	
Number of walk commuters	B080301e19	n/a
Number of transit commuters	B08301e10	n/a
Number of households with no vehicle	B25044e3 + B25044e10	n/a
Number of college degree holders	B15003e22	n/a
Percent of population that is white alone	B02001e2	B02001e1
Walk commute mode share	B08301e19	B08301e1
Transit commute mode share	B08301e10	B08301e1
Percent of households with no vehicle	B25044e3 + B25044e10	B25044e1
Percent of population with a college degree	B15003e22	B15003e1

2.2.1.2. Transit

There is not a freely available comprehensive dataset for the transit systems in the state. Much of the data are available through Google's General Transit Feed Specification (GTFS), though not all transit agencies publish their stops, routes, and schedules using that specification. We purchased data from AllTransit, a service provided by the Center for Neighborhood Technology (CNT). The AllTransit data also uses GTFS, but CNT staff contacted transit agencies that did not publish using GTFS and compiled a comprehensive database of transit information available for purchase. The AllTransit data is aggregated at the Census tract level. CNT processed the raw stop, route, and schedule data and provided us with a dataset that includes useful metrics, which include:

- Transit trips per week within one half mile
- Transit routes within one half mile
- Jobs accessible within a 30 minute trip
- Percent of commuters who use transit
- AllTransit Performance Score

The AllTransit Performance Score is an aggregate value from 1 to 10 that "measures more than just access to transit. It considers the performance of transit - connections to other routes, jobs accessible in a 30-minute transit ride, and the number of workers using transit to travel" (https://alltransit.cnt.org/methods/).

Although a few of the descriptions of the metrics provided in this dataset, like transit trips per week within one half mile, sound like they are site-specific, all the variables were provided at the tract-level. This scale is not conducive to identifying subtle differences in transit access between nearby locations, but it was the best available data source. For each metric except for AllTransit, we used the approach described in the demographics section to estimate the values within each buffer. For the AllTransit metric, we took an area-based weighted average of the tracts falling within the buffer.

2.2.1.3. Employment

We selected two employment metrics: employees and square footage of traffic-generating commercial uses. The first metric attempts to capture the contribution to pedestrian exposure from people working near the relevant intersections. The second measure captures the scale of businesses that generate walking trips by attracting customers. The data source for both metrics is ESRI Business Analyst software. ESRI sourced the data from Infogroup. The software mapped every business in the United States, complete with the number of employees that work there and the approximate size, in square feet, of the business. To determined the number of employees near relevant intersections, we conducted a GIS analysis that selected all businesses within our chosen buffer distance and summed the number of employees in those businesses. We did not discriminate based on the type of business.

We used the same data set for business square footage, but we filtered the businesses by type. Warehouses, for example, do not generate significant foot traffic outside of their employees, and that foot traffic is captured in the metric above. Each of the businesses in the ESRI Business Analyst dataset has a corresponding North American Industry Classification System (NAICS) code that categorizes the business by type. For our analysis, we only considered businesses from the following categories:

- 44-45: Retail Trade
- 522: Banks
- 54: Professional, Scientific, and Technical Services
- 62: Health Care and Social Assistance
- 71: Arts, Entertainment, and Recreation
- 72: Accommodation and Food Services
- 812: Personal and Laundry Services
- 813: Religious, Grantmaking, Civic, Professional, and Similar Organizations

The ESRI Business Analyst data do not provide exact square foot measurements for each business, but instead categorizes them into one of four ranges:

- A: 1 - 2,499 square feet
- B: 2,500 - 9,999 square feet
- C: 10,000 - 39,999 square feet
- D: 40,000 square feet and above

We used the middle of each of the A-C ranges and the lowest value for range D when summing the total amount of square feet within the buffer distance. Therefore, we applied the value 1,250 for A, 6,250 for B, 25,000 for C, and 40,000 for D. We summed all of the square footages for the businesses within the buffers for our metric.

2.2.1.4. Infrastructure

The functional classification variables indicated whether the streets making an intersection were either principal arterials, minor arterials, or collectors. For example, the value for principal arterial would be 0 if neither street was a principal arterial, 1 if one of the streets was, and 2 if both were. The four-way intersection and signal variables could have values of 0 or 1, if the intersection had the respective feature.

2.2.1.5. Network Connectivity

We developed a script in ArcMap that counted all street segments within the buffer distances from count locations. The script also computed the total meters of street centerlines within each of the buffer distances for the count locations. We used the California Road System (CRS) street GIS dataset when computing the street segments and street lengths.

2.2.1.6. Climate data

Although the annual volume estimates we used as the dependent variable are based on short-term counts conducted during varying weather conditions, we were not able to account for weather when making those estimates. We assumed, however, that the varying climates throughout the large state of California would affect the total pedestrian activity at different locations, particularly in locations and months of the year with extreme weather.

We used the California Energy Commission's (CEC) Building Climate Zone Areas map to define the 16 climate zones of California and define their boundaries. While this map was not created explicitly for transportation applications, the CEC needs accurate climate data to assign energy budgets to new and retrofitted buildings. The CEC also identified representative cities within each of the sixteen climate zones, and temperature and precipitation measurements were taken from each of those locations as representative of the entire zone.

The CEC website provided GIS map data of the zones, but did not provide the temperature and precipitation data publicly. To obtain this information for each of the representative cities, we relied on data from the National Oceanic and Atmospheric Administration (NOAA). NOAA's National Centers for Environmental Information provides an online climate data search. Using the tool to specify location, date range and data type, we downloaded all of the data we needed. NOAA did not have climate data for all of the cities chosen as representative cities within each climate zone. For zones 1, 8, and 16, we chose alternative cities within each climate zone with available NOAA data (see Table 2-8).

Table 2-8. Representative cities from each climate zone

Climate Zone	CEC City	SafeTREC City
1	Arcata	Eureka
2	Santa Rosa	Santa Rosa
3	Oakland	Oakland
4	San Jose	San Jose
5	Santa Maria	Santa Maria
6	Torrance	Torrance
7	San Diego	San Diego
8	Fullerton	Anaheim
9	Burbank-Glendale	Burbank-Glendale
10	Riverside	Riverside
11	Red Bluff	Red Bluff
12	Sacramento	Sacramento
13	Fresno	Fresno
14	Palmdale	Palmdale
15	Palm Springs	Palm Springs
16	Blue Canyon	Tahoe City

For each representative city, we gathered four weather data points (See Table 2-9):

Table 2-9. Weather data gathered for each city

Contents	Data Code	Year
Long-term averages of number of days during the year with precipitation >= 0.50 inches	ANN-PRCP-AVGNDS-GE050HI	2010
Long-term averages of number of days during the year with precipitation >= 1.00 inches	ANN-PRCP-AVGNDS-GE100HI	2010
Long-term averages of number of days during the year with snowfall >= 10.0 inches	ANN-SNOW-AVGNDS-GE100TI	2010
Long-term average number of days per year where tmax is greater than or equal to 90F	ANN-TMAX-AVGNDS-GRTH090	2010

The first two data points measure the average number of days a city accumulates precipitation. We hypothesized that areas with significant snowfall would see less pedestrian activity in winter months, so we gathered data on snowfall. Finally, we gathered data for extremely high temperatures. We chose data from 2010 as it is the most recent date NOAA made data available.

We assigned each count location to the climate zone they resided in and associated the corresponding climate data for the representative city of the climate zone. For example, all count locations within climate zone 12 inherited climate data from Sacramento. These data are granular in scale, but do account for differences in activity between different regions of the state.

2.2.1.7. Distance to nearest body of water

California contains tens of thousands of bodies of water, as well as hundreds of miles of coastline along the Pacific Ocean and the San Francisco Bay. Many of the bodies of water are small streams, ponds, and seasonal wetlands that likely would not have an effect on nearby pedestrian volumes. We used the definition of a lake from the “National Lakes Assessment” published by the EPA to narrow down the bodies of water and eliminate small ponds (Source). Our dataset only includes the 3,417 lakes in California 10 acres or larger. The GIS data originated from the California Department of Fish and Wildlife (DFW) from 2013.

We used 54 of the largest rivers and tributaries in the state and we eliminated all smaller streams and tributaries from our dataset. The dataset is published by Natural Earth, who sourced the data from World Data Bank 2, published by the U.S. Central Intelligence Agency. We obtained coastal GIS data from the DFW’s Marine Region’s GIS resources.

Once we compiled the coast, river, and lake dataset, we conducted a “Near Table” analysis on all of the pedestrian count locations. The analysis determines the closest body of water to the count location and reports that distance in miles.

2.2.1.8. Schools

We prepared a dataset of all primary and secondary schools, public and private, in California. We obtained the private school directory from the California Department of Education (DOE) and geocoded the addresses found in the directory (http://www.cde.ca.gov/ds/si/ps/). The public school GIS was also found on the DOE website. The DOE private school data was not geocoded, so we performed that task so all school sites could be used in GIS. We then simply tallied the number of schools within a given set of buffer distances from count locations. We ensured that the schools were classified as “open” and not “closed” or “pending.” We also processed the data so that multiple schools at the same location were counted as one school. This was necessary, as the data often included many schools at the same site, even though they could be considered the same school. This includes evening, vocational, and adult education programs; district offices; and multiple programs in the same high school.

2.2.1.9. Slope

We measured the maximum slope of streets leading to relevant intersections. To do this, we applied a buffer of 330 feet from the center of the intersection, then gathered the latitude/longitude pairs from the points where the buffers intersected the streets. We wrote a script that fetched the elevations from the points using the Google Maps Elevation API. We compared the elevations for the street points to the elevations from the intersection to determine the slope for each street leading into the intersections. We then found the maximum slope for each intersection and used that data as the slope variable.

2.2.2. Dependent Variable

The dependent variable for the pedestrian exposure model is the annual pedestrian crossing volume at each intersection count site. Because agencies cannot afford to conduct long-term crossing counts, either manually or using automated counting technology, we created annual count estimates by expanding short-term crossing counts using expansion factors in a process explained under Processing Count Data.

This approach required that we compile large amounts of short-term count data, as the dependent variable for the model, and long-term count data to create the expansion factors. The count data processing involved two main tasks. First, we used the long-term count data to develop expansion factors, and second, we applied the expansion factors to the short-term counts to create the annual volume estimates.

2.2.2.1. Count Data Compilation

Each Caltrans district has a budget for collecting video-based count data through Miovision, and these data, collected at several hundred locations, formed the basis of the short-term pedestrian intersection count data. Among the 583 count studies, durations ranged from 1 hour to 96 hours. Count durations longer than 12 hours were generally multiple daytime counts, such as 7AM to 7PM on consecutive days.

To acquire more data, Caltrans Local Assistance emailed a count data request to a list of previous applicants for Active Transportation Program (ATP) grant funding, and Streetsblog shared the request on their website. A number of agencies shared their pedestrian and bicycle count data sets. Table 2-10describes the short-term pedestrian intersection count studies that we received in this outreach effort. The count data under “Other” came from a transportation enthusiast who had scraped multiple municipal websites and planning reports to extra count data (https://github.com/ericfischer?tab=repositories). Most counts were conducted during morning and afternoon 2-hour peaks and some also included midday peak.

Table 2-10. Sources of short-term pedestrian intersection count data provided by local jurisdictions

Agency/Jurisdiction	Number of Count Studies	Number of Unique Locations Used in Model
Alameda County Transportation Commission	101	59
City of Brea	6	0
El Dorado County	16	0
Los Angeles County	5	0

Town of Los Gatos	5	3
City of Newport Beach	17	0
Orange County Transportation Authority	122	97
City of Rialto	35	10
City of Riverside	3	0
City of Santa Ana	71	68
City of San Luis Obispo	99	11
San Luis Obispo County	189	68
San Luis Obispo Council of Governments	91	87
Sonoma County	2	0
South Gate	14	5
Tulare County	81	61
Other	657	479
Caltrans - Miovision	538	360
Total	**2052**	**1308**

We compiled a total of 2,052 short-term count studies from throughout the state. This number was narrowed down to a total of 1,308 intersection locations for several reasons:

1. Some locations had multiple studies so these were consolidated into a single location for the model.
2. Some Miovision studies on the SHS counted no pedestrians during the duration of the study. We assumed that these locations had negligible pedestrian activity are were not appropriate for inclusion in the model.
3. Since the exposure model is intended for estimating pedestrian volumes on the SHS, and California’s roads are not representative of the SHS, we screened intersections by functional classification of the intersecting roads. We eliminated intersections whose two intersecting roads were some combination of local and minor collector and did not include a higher level functional classification. We used the functional classification data from the CRS dataset.

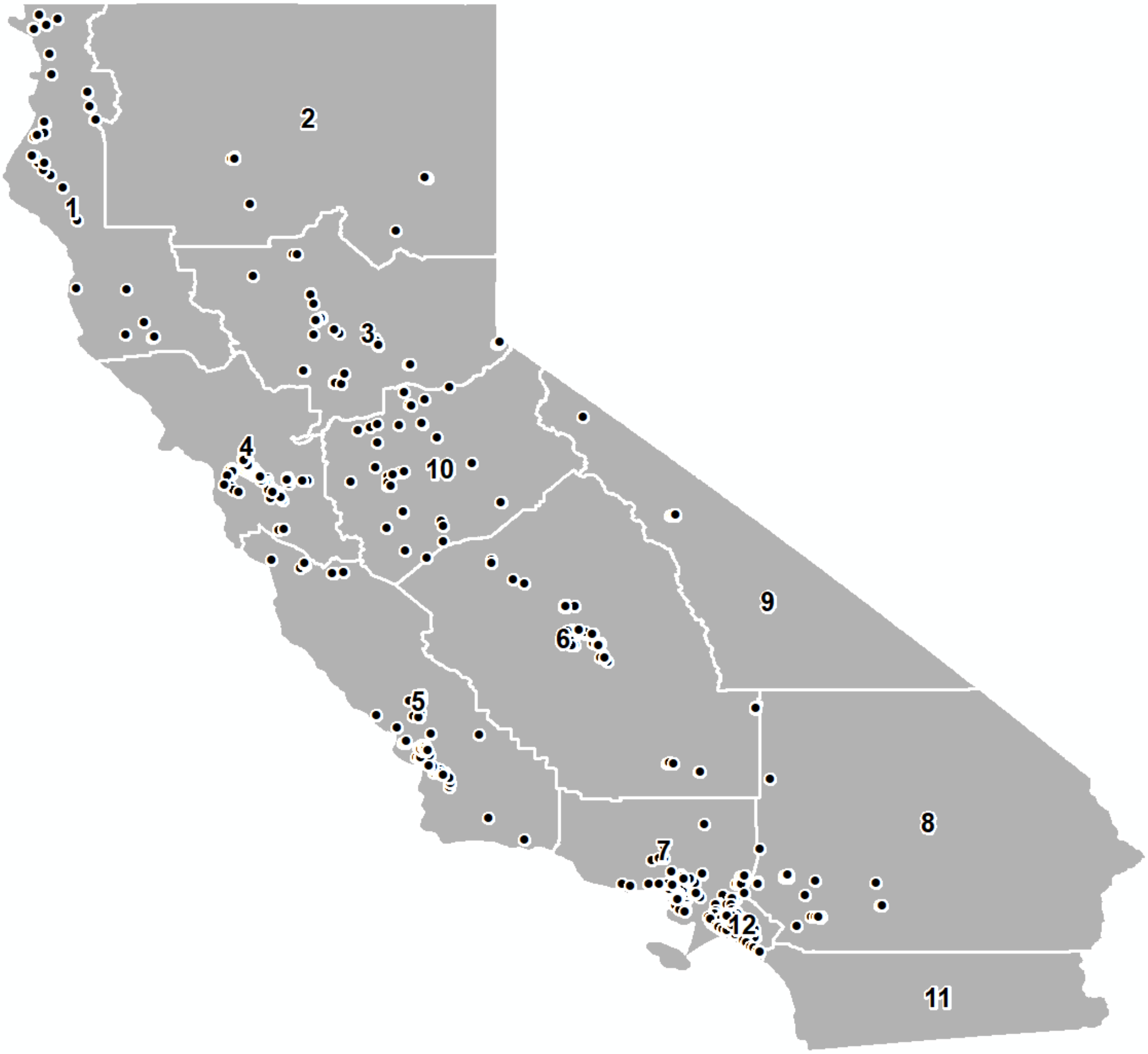


Figure 2-1. Short-term count locations included in model

Among the 1,308 short-term count locations included in the study (Figure 2-1), count durations ranged between 1 and 86 usable hours for developing the annual estimates. The average count duration was 7.9 hours and the median was 6 hours. Most of the 6-hour counts were a combination of 2-hour morning, midday, and afternoon peak counts.Figure 2-2 shows the number of locations by count duration.

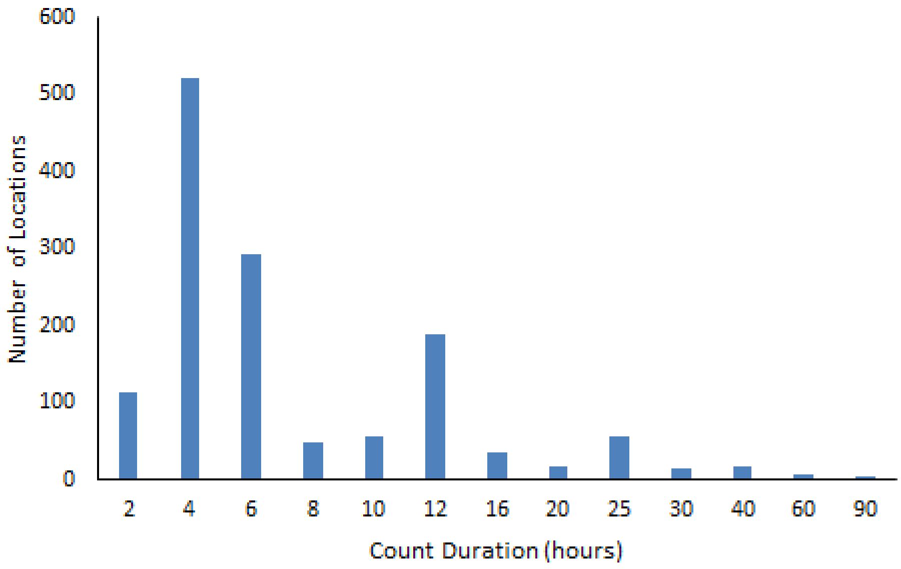


Figure 2-2. Histogram of short-term count durations

In addition to the short-term count data, we compiled 133 long-term count studies with count durations ranging between 1 week and 5 years. Automated passive infrared counters from the same vendor were set up at sidewalk and trail locations in Alameda, Fresno, Los Angeles, San Diego, and San Francisco Counties between 2008 and 2017 as part of unrelated research and planning programs. These counters collect screenline volumes and cannot capture intersection crossing volumes, but when set up on a sidewalk near an intersection, they likely capture similar activity patterns to the crosswalks. The counters aggregated counts to 15-minute or 1-hour intervals, and we analyzed all counts at 1-hour intervals.

Figure 2-3 shows the number of sites by annual volume estimate and district. Districts 4 and 7, the most urban districts, have the highest volume locations, as expected. District 12 also has some higher volume sites.

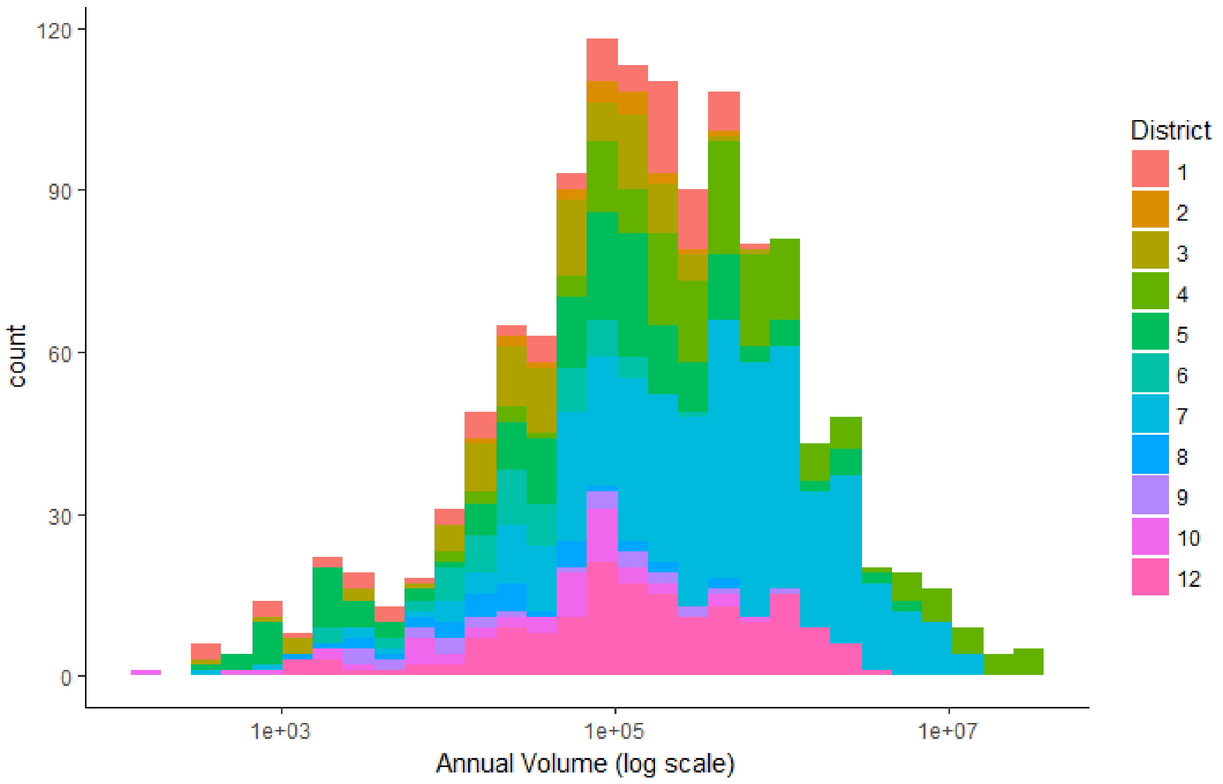


Figure 2-3. Histogram of annual volume estimates by district

2.2.2.2. Creating Expansion Factors

A number of pedestrian and bicycle studies have used automated count data from a limited number of locations to develop expansion factors that can be used to expand short-term counts to long-term volume estimates (8,16,18,19,20,21,22,23). This process is based on the assumption that similar sites will have similar patterns in hourly, weekly, and monthly volume trends. Even if the magnitude of volumes is different, the proportional trends can be similar.

Before calculating the expansion factors, we cleaned the long-term count data. For each location, we graphed the count values over time and flagged observations that were outliers. These outliers included intervals with spikes in counts that were more than double the values at corresponding times for surrounding weeks and extended periods of zero counts, both indicating counter errors or aberrations from normal activity (e.g., athletic event, parade, street festival). We also excluded national holidays.

To account for human error, we also developed algorithms to automate the error checking. We flagged count periods that violated a check that values were within a certain number of standard deviations of the expected value. We tested two methods of establishing the threshold:

- Single observation threshold. For the count in question, consider the counts taken at the same time of the week in the previous 4 weeks and in the following 4 weeks. The count is “probably incorrect” if it is more than two standard deviations above or below the average of the eight same-time-of-week counts. The same-time-of-week counts should exclude holidays.

- Multiple observation thresholds (four consecutive count periods). For the four consecutive count periods in question, consider the counts taken at the same time of the week in the previous 4 weeks and in the following 4 weeks. For each of the four periods, calculate the average and standard deviation of the eight corresponding same-time-of-week counts. The four consecutive hours of counts are “probably incorrect” if each individual count in the series is more than one standard deviation above or below the average of its eight corresponding same-time-of week counts. The same-time-of-week counts should exclude holidays.

After flagging the periods with potential counter errors or outliers, we graphed the counts for the surrounding two weeks to visually evaluate whether the count looked incorrect. We determined that the automated approach did not flag appropriate outliers, so we did not incorporate these checks into the final data cleaning process.

After cleaning the data, we compiled all the long-term counts into two spreadsheets. The first spreadsheet was a location reference that included the intersection ID, city, county, state, latitude, and longitude. The second sheet included all of the counts aggregated into hourly time periods.

For each counter, we calculated hour-to-weekday, day-to-week, and week-to-year expansion factors as the proportion of the total daily, weekly, or annual pedestrian traffic that a given smaller period is expected to represent. One option for combining the individual locations is to take the average of the expansion factors at all sites in what is called the single factor approach. Previous research has shown, however, that sites with different land use have different activity patterns, which supports the use of factor groups for the expansion factors. We used the approach described in (24) to define the hour-to-weekday factor groups based on the land use at the long-term count site as defined in Table 2-11. We could not identify land use categories that correlated with weekend activity patterns, so weekend counts were not used in the annual volume estimates. The hour-to-weekday expansion factors are shown in Table 2-12.

Table 2-11. Land use definitions for factor groups

Category	Definition
Central Business District	In the downtown area as labelled on Google Maps or from expert knowledge of the area
School	School facility on adjacent block or yellow crosswalk present at intersection
Trail	Count location within a block of trail access point
Other	All other sites

Table 2-12. Hour-to-weekday expansion factors by land use factor group

Hour	Other	CBD	School	Trail
12:00 AM	0.009	0.011	0.003	0.001
1:00 AM	0.006	0.009	0.002	0.001
2:00 AM	0.005	0.006	0.001	0.000
3:00 AM	0.004	0.006	0.002	0.001
4:00 AM	0.005	0.005	0.003	0.001
5:00 AM	0.011	0.009	0.012	0.030
6:00 AM	0.024	0.020	0.023	0.061
7:00 AM	0.055	0.044	0.118	0.084
8:00 AM	0.059	0.054	0.080	0.098
9:00 AM	0.053	0.060	0.047	0.098
10:00 AM	0.056	0.070	0.046	0.070
11:00 AM	0.059	0.077	0.053	0.044
12:00 PM	0.062	0.081	0.058	0.031
1:00 PM	0.063	0.076	0.074	0.034
2:00 PM	0.070	0.070	0.088	0.034
3:00 PM	0.083	0.070	0.110	0.064
4:00 PM	0.074	0.062	0.061	0.068
5:00 PM	0.073	0.059	0.060	0.084
6:00 PM	0.065	0.052	0.050	0.114
7:00 PM	0.055	0.049	0.046	0.049
8:00 PM	0.041	0.037	0.025	0.016
9:00 PM	0.031	0.031	0.018	0.008
10:00 PM	0.022	0.023	0.012	0.007
11:00 PM	0.015	0.019	0.007	0.003
No. of Locations	**55**	**25**	**15**	**7**

Whether the location was near a school was the main feature distinguishing between the day-to-week activity patterns. Schools typically had less activity on weekends. The day-to-week expansion factors are shown in Table 2-13.

Table 2-13. Day-to-week expansion factors by land use factor group

Day	**Non-School**	**School**
Mon	0.146	0.154
Tue	0.149	0.154
Wed	0.149	0.156
Thu	0.148	0.169
Fri	0.149	0.154
Sat	0.138	0.114
Sun	0.121	0.098
No. of Locations	**83**	**20**

There were 8 locations with at least one year of continuous counts. These factors (Table 2-14) do not sum to 1 because they are meant to take one week of counts in a given month and expand to an annual estimate.

Table 2-14. Week-to-year expansion factors

Month	Group Average
January	0.079
February	0.082
March	0.085
April	0.086
May	0.084
June	0.082
July	0.092
August	0.088
September	0.086
October	0.083
November	0.080
December	0.074
No. of Locations	**8**

We applied the expansion factors to calculate the annual volume estimates based on hourly counts. First, we used the hour-to-weekday factors to estimate daily volumes using the following formula:

$$\widehat{V_{daily}} = \frac{\sum_i^n V_i}{\sum_i^n \alpha_i}$$

where V_i is the observed volume in hour i, and α_i is the hour-to-day expansion factor calculated for hour i. We used a corresponding formula for the day-to-week and week-to-year expansion. The resulting number was the annual estimate for the year the short-term counts were conducted. We normalized these counts to 2016 using adjustment factors developed from Census ACS estimates of total population for California by year (Table 2-15).

Table 2-15. Annual adjustment factors to normalize count estimates to 2016

Year	Total CA Population	Annual Adjustment Factor
2006	35,980,000	1.09
2007	36,230,000	1.08
2008	36,580,000	1.07
2009	36,960,000	1.06
2010	37,330,000	1.05
2011	37,680,000	1.04
2012	38,010,000	1.03
2013	38,340,000	1.02
2014	38,680,000	1.01
2015	38,990,000	1.01
2016	39,250,000	1.00

2.2.3. Exposure Model Estimation

The final dataset included 75 explanatory variables, when accounting for the different buffer sizes for appropriate variables, and the dependent variable (annual volume). It was not practical to test all the possible combinations for this many variables, so we performed some preliminary analyses to identify variables that may be more significant, avoid problems with collinearity, and determine if transformations may be appropriate.

2.2.3.1. Variable Selection and Transformation

Scatter plots of the dependent and independent variables are a useful tool for visually evaluating correlation and patterns that may suggest the value of transforming the variables. For example, number of employees with a quarter mile showed a dispersed relationship when graphed against annual volume (see Figure 2-4). Taking a log transformation of annual volume narrows the band, but the relationship is still not linear (seeFigure 2-5). Finally, graphing the log of both number of employees and annual volume demonstrates a more linear relationship (see Figure 2-2)Figure 2-5. The second two graphs also show how the inclusion of locations with zero values for the annual volume is problematic, since the trends are not consistent with the other locations. Dropping these locations brings the sample down to 1,270 locations.

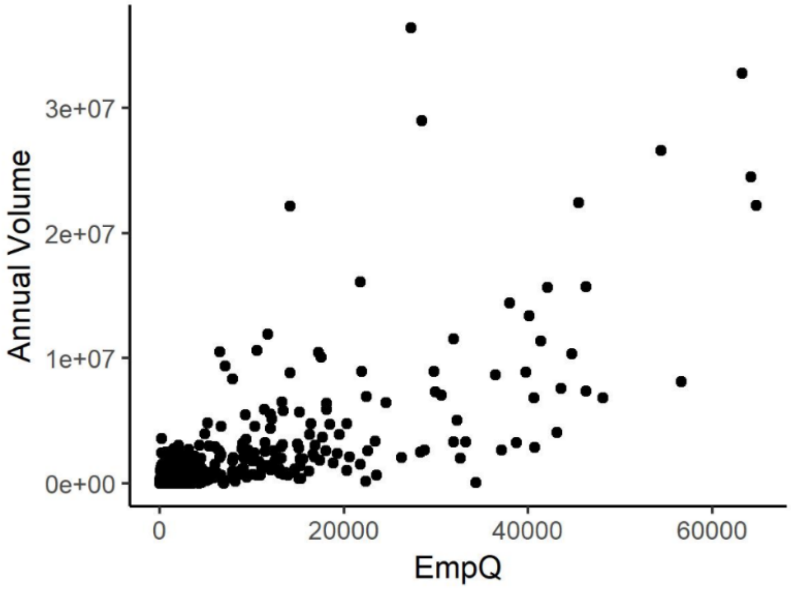


Figure 2-4. Scatter plot of number of employees within a quarter mile and annual volume

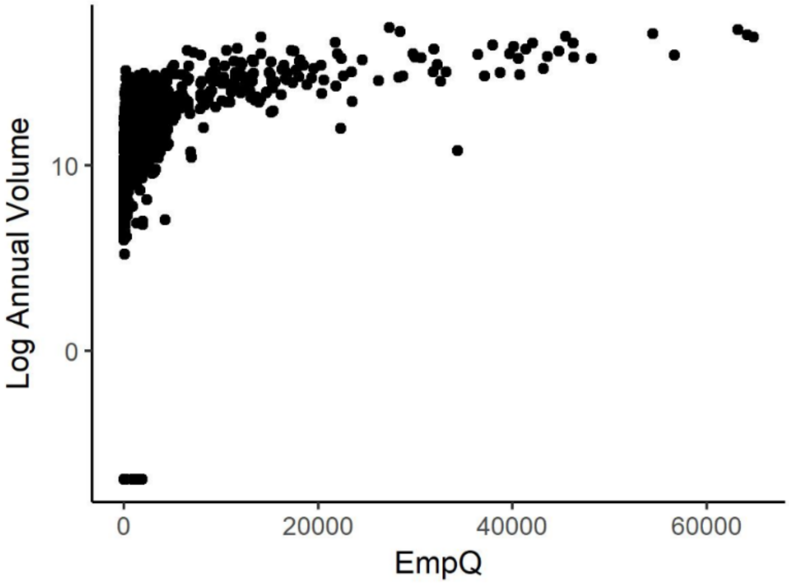


Figure 2-5. Scatter plot of number of employees within a quarter mile and log of annual volume

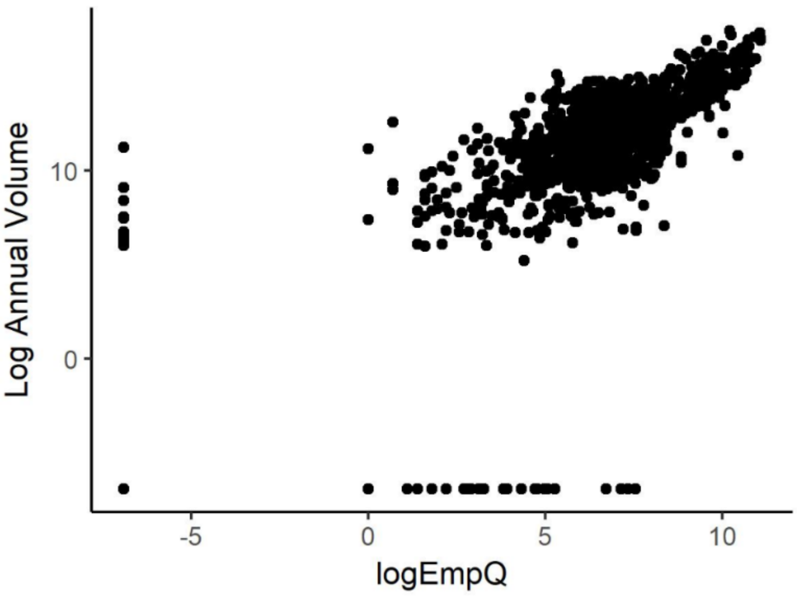


Figure 2-6. Scatter plot of log of number of employees within a quarter mile and log of annual volume

We produced corresponding scatter plots for all 75 explanatory variables. Evaluation of these graphs helped us determine that using a log transformation of the dependent variable was most appropriate. We also calculated Pearson and Spearman correlation coefficients for the complete set of variables, including log transformations of all variables. Pearson correlation evaluates the linear relationship between variables, whereas Spearman compares the ranks of the values and does not use the raw values. Spearman is less vulnerable to being pulled up by very high outliers, of which there are a few in this dataset. The correlation analysis helped us to identify the set of variables that were strongly correlated with the log of annual volume, such as log of the number of routes within a tenth mile (logRoutesT) (ρ_p=0.70, ρ_s=0.73). Additionally, for each variable highly correlated with the dependent variable, we needed to avoid including other explanatory variables in the model that were strongly correlated and a cause of collinearity. For instance, logRoutesT is strongly correlated with population within a half mile (PopH) (ρ_p=0.74, ρ_s=0.88).

2.2.3.2. Initial Estimation

The model was specified as a loglinear regression and estimated using ordinary least squares regression. The model structure was as follows:

$$\ln(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_j X_{ji} + \epsilon_i$$

where:

Y_i = annual pedestrian crossing volume at intersection i;
X_{ji} = value of explanatory variable j at intersection i; and
B_j = model coefficient for explanatory variable j,
ϵ_i = error term for the log-linear model, which is normally distributed.

Using the logarithm of the dependent variable prevents the model from returning negative values. Linear and negative binomial models were also tested, but the loglinear model had the best fit when evaluating both training and test data. To develop estimates from a loglinear model, we use the following equation:

$$\widehat{Y_i} = e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_j X_{ji}}$$

For the model estimation, we removed a random subset (10 percent) of the locations for testing and used the remaining 90 percent for the training. The testing subset was used to test and compare the predictive power of models estimated using the training data. We tested 300 different iterations of the training and testing data to be sure that the goodness-of-fit (adjusted R-squared) and residual sum of squares (RSS) were consistent. We began each model with a set of 15 to 20 non-collinear explanatory variables and ran stepwise backward regression to automatically remove the least significant variable in each step until all the variables were significant (p-value<.01). To improve the goodness-of-fit, we tested trading out similar explanatory variables, including different buffer distances for the same variables and strongly correlated other variables. We tested removing variables to see the impact on the adjusted R-squared and dropped variables that had an impact of less than 0.01. We also dropped variables if the direction of the coefficient did not make intuitive sense.

2.2.3.3. Manipulation of Variables

At intermediate stages in the modeling process, we plotted residuals and compared predicted and observed values to understand how the model was performing (Figure 2-7). We found that one particular low observed volume location had a very high predicted location. This location at Hyde and Turk in San Francisco has very high population (a positively correlated variable), but does not otherwise have conditions as favorable to walking as nearby neighborhoods, partly due to higher crime and fewer businesses in the immediate vicinity.

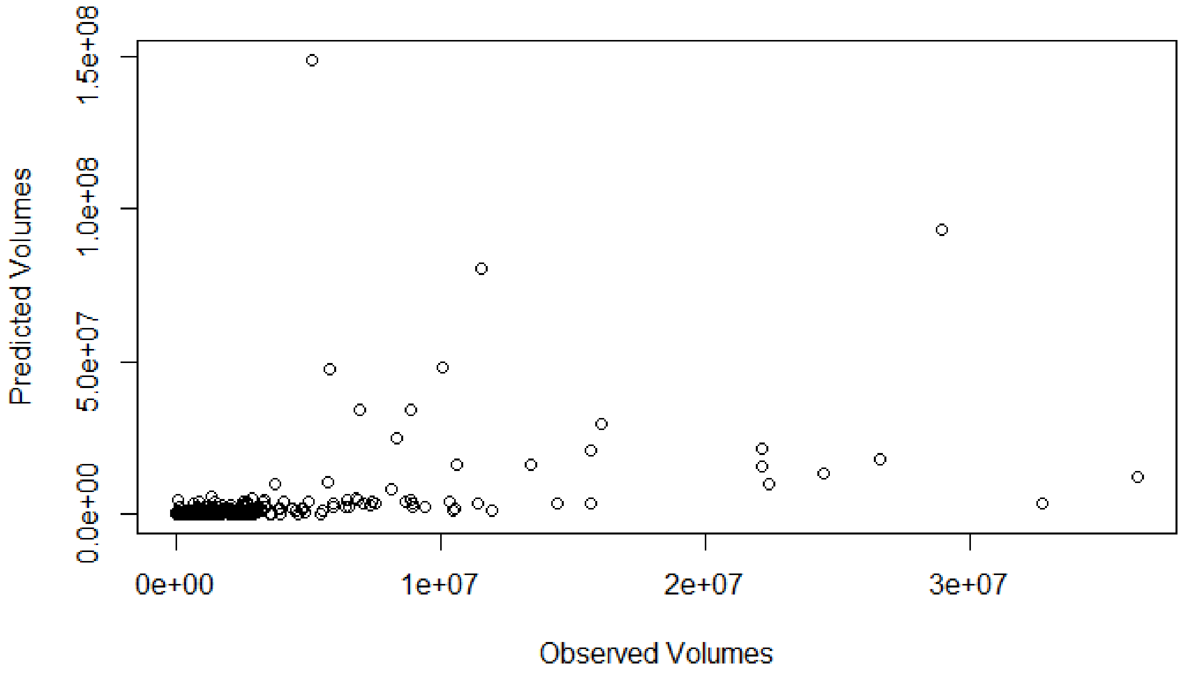


Figure 2-7. Observed vs. predicted pedestrian volumes (training data)

To further examine this error, we plotted the residuals vs. population within a half mile (Figure 2-8). In this plot, we see that for population values greater than 20 thousand the residuals drop off, indicating that there is a threshold beyond which greater population does not point towards greater walking activity. As a result, we truncated the population within a half mile variable at 20 thousand. Replacing the original population variable with the truncated version in the model removed the outlier. There were only 20 locations that required truncation, 14 of which were in District 4 and 6 of which were in District 7.

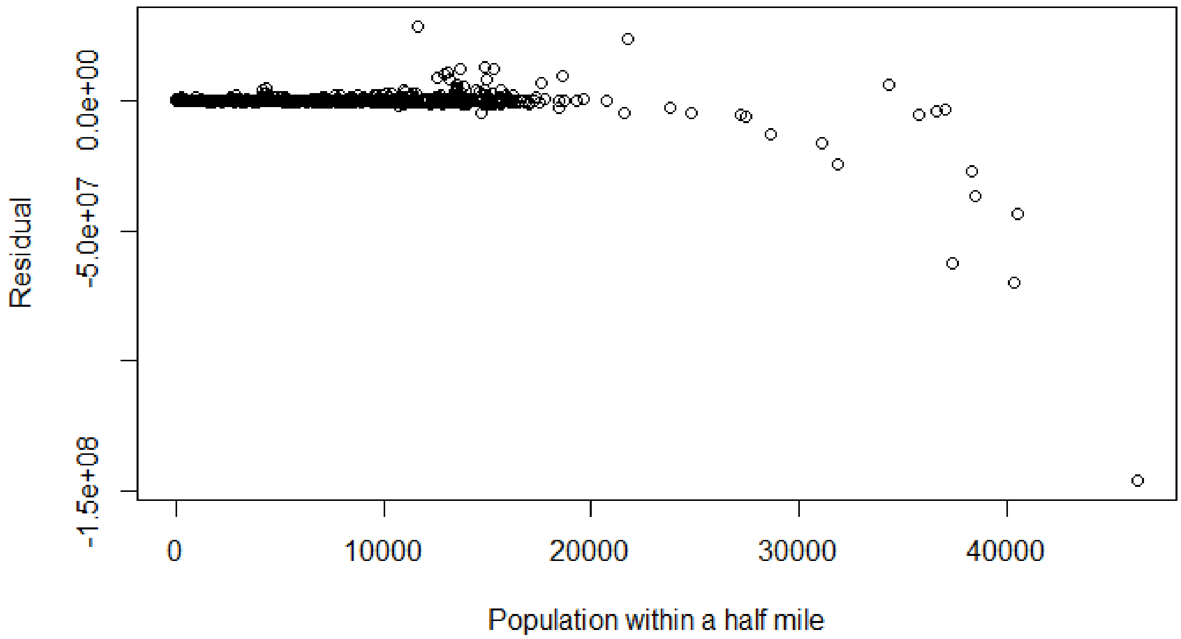


Figure 2-8. Residuals as a function of population within half a mile

Several of the variables that performed well in the models tested were best with the larger buffer scales (half and quarter mile). We theorized that greater concentration of a variable closest to the count location may have more impact on volume than when it is evenly dispersed or has greater concentrations further from the site. To account for this we modified two buffer variables, employees and street segments, to weight them according to high concentrations within a tenth of a mile by multiplying the variable by (1 + Value_Tenth/Value_X), where X is the scale of the given variable. This change improved the R-squared by approximately 0.01. Figure 2-9 and

Figure2-10 repeat the observed vs predicted and residuals as a function of population within a half mile after the variables were manipulated. It is apparent that the major outlier is gone from the predicted values and the residuals are more linear as a function of the truncated population with a half mile.

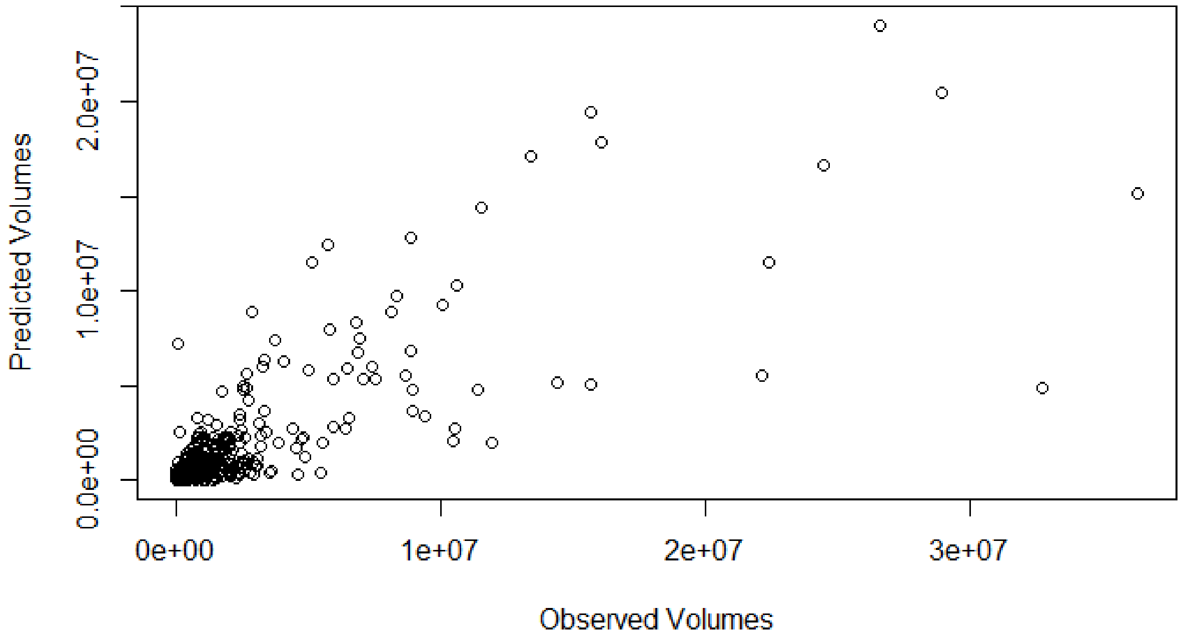


Figure 2-9. Observed vs. predicted pedestrian volumes (training data) with truncated population variable and weighted number of employees and street segments

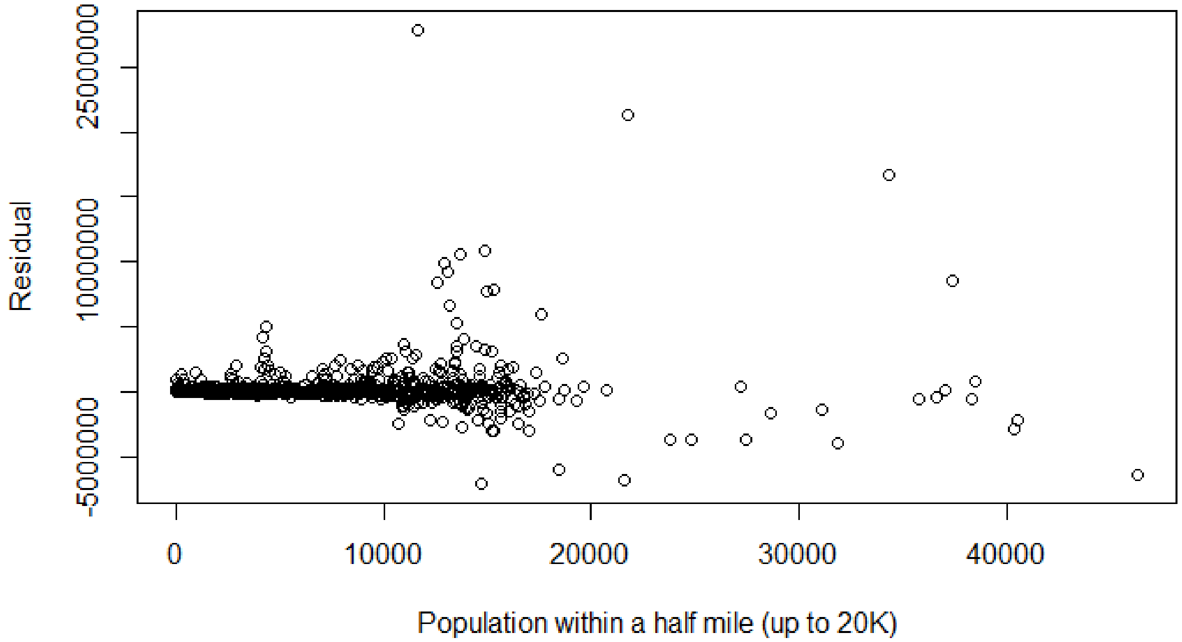


Figure 2-10. Residuals as a function of truncated population within half a mile

2.2.3.4. Final Model

The final model included eight explanatory variables and the specification is described inTable 2-16. All variables were highly significant (p-values << 0.001). The adjusted R-squared was 0.714, meaning that the model is able to explain 71.4 percent of the variability of the log of pedestrian volumes.

Since the linear regression models the log of the pedestrian volumes using a normally distributed error term, the untransformed volume estimates (obtained as exponential of the log-linear model's output) follow a log-normal distribution. The coefficient of variation, which is a standardized measure of dispersion, for the log-normally distributed pedestrian volume estimates was 1.5, which indicates a 50% error relative to the mean estimate. The root mean squared error (RMSE) for the untransformed estimates was 1,582,572.

Log of the weighted number of employees within a quarter mile had a positive coefficient, indicating that presence of jobs nearby is positively associated with pedestrian activity. This result is consistent with previous research. Work locations are attractors for commuters, who may finish their commute of any mode on foot, even if they do not walk the whole way. Additionally, businesses and government offices may attract visitors seeking services.

Truncated population within a half mile also had a positive coefficient, a logical result, indicating that the presence of more residents is positively associated with pedestrian activity. Residences are potential sources of pedestrians, who may be conducting commute or other trip on foot or connecting to other modes of transportation. Even if walk mode share is low, more population means more potential pedestrians. Similar variables have been significant in many previous pedestrian models.

Log of weighted street segments within a half mile had a positive coefficient, indicating that a greater density of streets is associated with greater pedestrian activity. This result is consistent with previous research on street connectivity, which theorizes that people are more likely to walk when there are more route options that are more direct.

Walk commute mode share within a half mile had a positive coefficient, an expected result. Although this variable is similar to our dependent variable, it does not cause endogeneity problems because it is an aggregate variable at a larger scale.

Log of number of schools within a half mile had a positive coefficient, indicating that presence of more schools is associated with greater pedestrian activity. This variable has performed well in previous models. Schools are a potential attractor for pedestrian activity, but they also may be a proxy for density or mix of land uses.

The principal and minor arterial variables both had positive coefficients. We theorized that these arterial roads may be locations where more businesses are located and more trips on transit and walking occur since there are more attractors. Principal arterials have a greater impact on pedestrian activity in the model, and this result is consistent with our expectation.

The four-way intersection dummy also had a positive coefficient. We expected that intersections with four legs would have more activity due to the greater connectivity as well as having more crosswalks.

Table 2-16. Final pedestrian exposure model

	Scale	Manipulation	Transformation	Estimate	Pr(>|t|)
Intercept				5.58	< 2e-16 ***
Number of employees	1/4 mile	weighted	log	0.390	< 2e-16 ***
Population	1/2 mile	truncated		0.000142	< 2e-16 ***
Number of street segments	1/2 mile	weighted	log	0.302	2.08e-05 ***
Walk commute mode share	1/2 mile			2.84	6.25e-08 ***
Number of schools	1/2 mile		log	0.0444	1.38e-05 ***
Principal arterial	Intersection			0.457	4.17e-16 ***
Minor Arterial	Intersection			0.384	6.23e-10 ***
Four-way intersection	Intersection			0.413	7.38e-09 ***
Dependent Variable: log(Annual Volume Estimate) Adj. R^2 = 0.714					

*** p-value < 0.001

Table 2-17shows the correlations between the dependent and explanatory variables. Many of the explanatory variables are moderately correlated with each other, but still provide significant explanatory power within the model. Principal arterial, minor arterial, and four-way intersection have lower correlations with the dependent variable because they have limited discrete values, whereas the other variables are continuous.

Table 2-17. Correlation matrix of dependent and explanatory variables

	Annual Volume[l]	Employees[l,w,4]	Population[t,2]	Street Segments[l,w,2]	Walk Mode Share[2]	Schools[l,2]	Principal Arterial	Minor Arterial	Four-Way
Annual Volume[l]	1								
Employees[l,w,4]	0.71	1							
Population[t,2]	0.70	0.47	1						
Street Segments[l,w,2]	0.64	0.66	0.58	1					
Walk Mode Share[2]	0.48	0.51	0.39	0.31	1				
Schools[l,2]	0.42	0.29	0.43	0.42	0.16	1			
Principal Arterial	0.24	0.14	0.16	0.17	0.01	0.05	1		
Minor Arterial	0.16	0.13	0.13	0.00	0.14	0.08	-0.49	1	
Four-Way	0.34	0.24	0.23	0.21	0.11	0.20	0.16	0.08	1

[l] log transformation; [w] weighted; [t] truncated; [2] ½ mile buffer; [4] ¼ mile buffer

2.3. Model Application

2.3.1. Model Scope

The initially stated scope of the pedestrian exposure model was the California SHS. There are practical reasons for limiting this scope to roads that are expected to have pedestrian activity:

1. Certain state highways, including freeways and some expressways, prohibit pedestrians. While this does not necessarily mean that there are not pedestrian collisions in these locations, the pedestrian exposure at these locations is likely to be extremely low and cannot be modeled in the same way as locations where pedestrians may legally travel. We do not know enough about the reasons why people walk on freeways unless they have a disabled car. This behavior is prohibited and is related more to random chance of a car collision or other problem than to the conditions of the surrounding area.
2. Certain state highways, primarily remote 2-lane rural highways, despite allowing pedestrians, have essentially no pedestrian activity because there are no nearby activity generating land uses. The variables we use in our model would not predict the very limited activity in these locations.

In the first step before defining the scope, we generated an intersection file for the SHS based on CRS street centerline data, using the following broad steps within a Model Builder tool in ArcMap.

1. Selected roads within 200 feet of state highways, excluding freeways.

2. Merged divided roads into single line segments with a merge distance of 40 meters.
3. Found the intersection of these road segments to create intersection points.
4. Deleted overlapping points.

This process created a set of points located at the intersection of street centerlines along the SHS.

The process required some extra steps to fix errors identified through visually checking. We used CRS data to identify roads with highway numbers (HWY_NUM) and exclude functional classification (FC_DRAFT) equal to 1 (freeways), but that was not enough information to generically identify expressways that serve essentially as freeways (like Hwy 24 in District 4). The speed limit data in ESRI Streets data and the design speed data in the TASAS highway table, which we have in a linear referenced layer, were useful in removing overpass and underpass locations that were falsely created as intersections in the above process. We could not use the linear referenced TASAS highway or intersection data for intersection generation because the linear referencing does not place intersections accurately enough based on the postmiles. None of the above mentioned streets datasets had coincidental lines, so we relied on proximity thresholds instead of intersecting features when removing roads from the dataset. Additionally, Merge divided tool did not successfully merge all divided roads, which meant that some intersections were represented by as many as 4 non-overlapping points. We used a 25-meter tolerance to remove some of these duplicates, but some remained. We could not increase the tolerance without removing some points for offset intersections. The final intersection file contains 23,017 points and has some errors, but was too large to manually check all locations.

We tested several methods of determining the scope of model application, and used 10 years (2004-2013) of pedestrian collision data from SWITRS to validate the different approaches. This was based on the assumption that collisions are a proxy for pedestrian activity. We determined that population density was better than roadway attributes, such as number of lanes, design speed, or access control, at identifying locations where pedestrian collisions occur, and thus where pedestrians likely travel. Using Census block-level population densities, we selected intersectionswithin 200 metersof a block with minimum 500 people per square mile. The final model scope file contains 12,414intersection points (shown in Figure 2-11). We validated the scope by examining pedestrian collision locations, assuming these locations are where people are walking. Of the 13,963 pedestrian collisions along the SHS between 2004 and 2013, 13,669 (98%) occurred within 200 meters of a scope intersection. Of the remaining collisions, only 51 (0.3%) were intersection collisions.

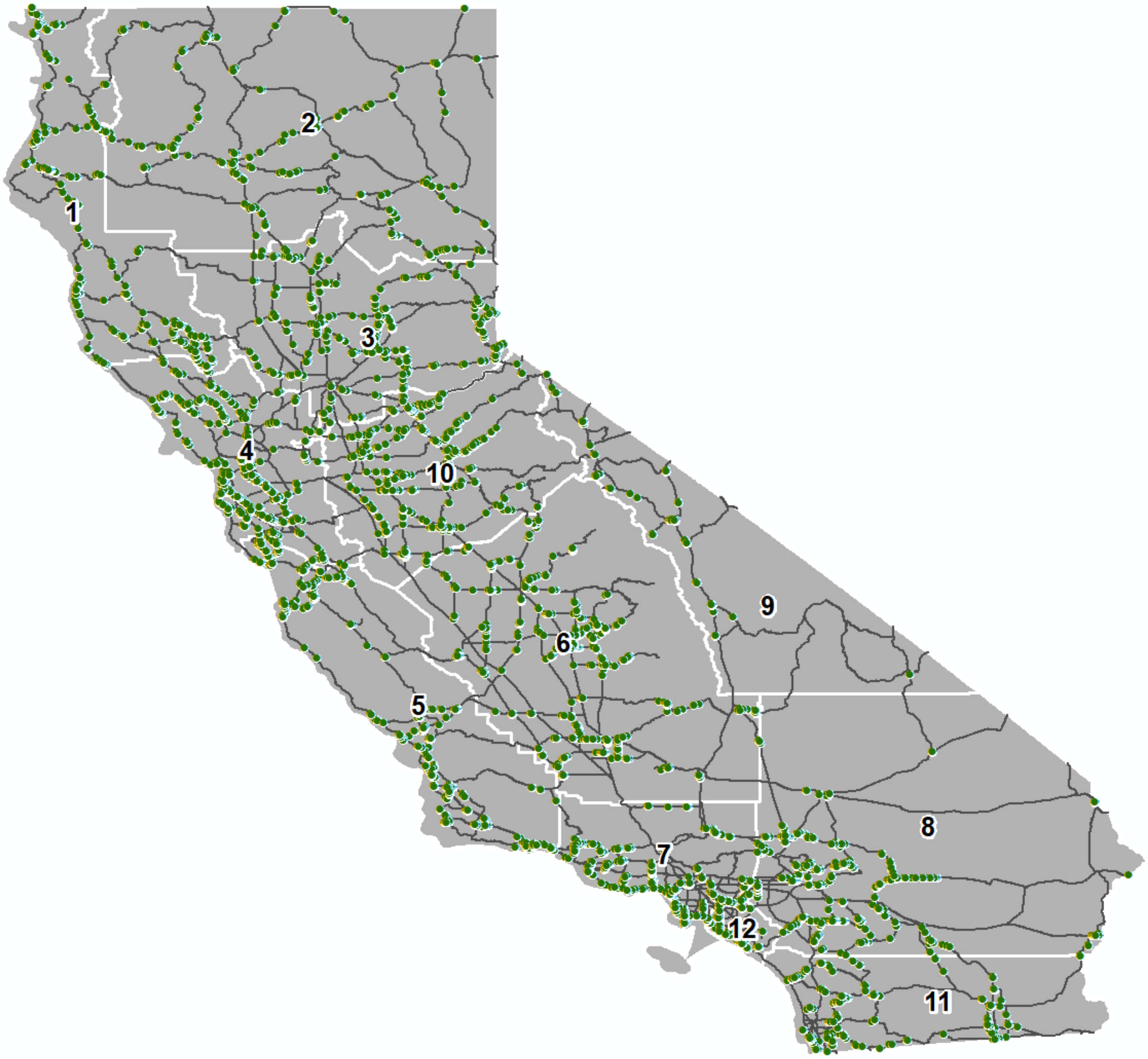


Figure 2-11. Final pedestrian exposure model scope – intersections with volume predictions

2.3.2. Pedestrian Volume Predictions

We calculated each of the explanatory variables from the final model for all the locations in the model scope and used these values along with the model results to predict annual pedestrian volumes at each location. Figure 2-12 shows the number of locations by predicted annual pedestrian volume and Caltrans district.

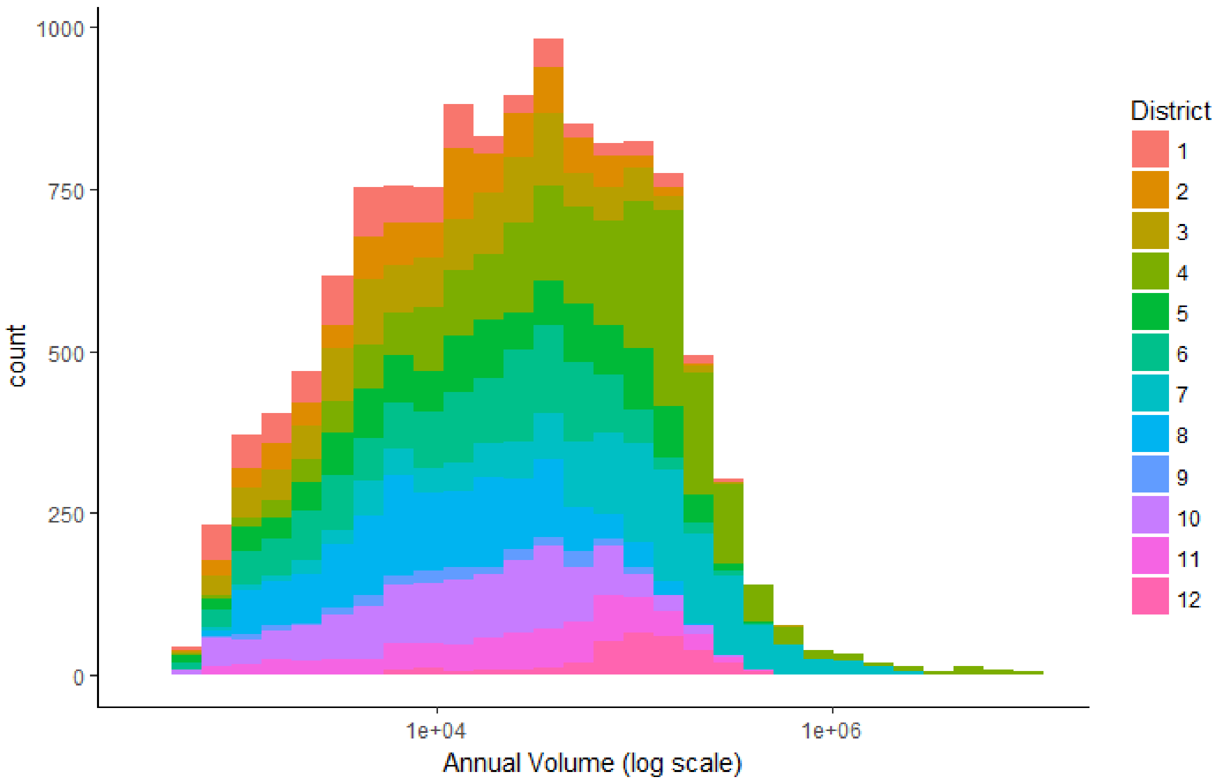


Figure 2-12. Histogram of the number of locations by predicted annual pedestrian volume and district

Chapter 3. Contextualized Hotspot Clustering

3.1. Motivation

Pedestrian hotspots, also referred to as high collision concentration locations (HCCLs), are identified within the pedestrian safety monitoring report (PSMR) tool by using crash frequencies and their associated injury severity distribution as the selection and prioritization criteria. When compared to techniques that are applicable to hotspots associated with automobile crashes (e.g., crash rates, safety performance functions and empirical Bayes), the use of crash frequencies can be restrictive, as it can suffer from regression-to-mean bias. In this regard, the estimation of intersection-level pedestrian volumes will be helpful in the future for developing more statistically robust methodologies for pedestrian hotspot identification.

But in cases when pedestrian exposure may not be readily available, (such as for ramps, segments), the use of only crash outcome-based metrics (injuries, fatalities) may not necessarily translate to recurring crash concerns with systematic, underlying patterns occurring at the hotspot location. In contrast, once an HCCL is available for investigation, traffic safety investigators are likely to look for such patterns within crash attributes and narratives when determining whether a site needs a specific countermeasure. The absence of such patterns may result in no recommendation for countermeasures which implies that the investigation resources were sub-optimally utilized. Thus, it would be helpful to bring some of the pattern matching capabilities upstream of the investigation process into the hotspot identification/prioritization process.

3.2. Clustering Methodology

Clustering is a form of unsupervised learning approach that seeks to group multidimensional data that "similar" to each other. In order to define what represents similarity in observations, either a distance-based approach or a probability-based approach can be undertaken. A distance-based approach is utilized when the input data represent counts or numbers, and the similarity can be quantified by evaluating which cluster is closest using a distance function (e.g., euclidean distance). An example of a popular distance-based clustering approach is k-Means clustering, wherein each observation is associated with a cluster which is closest in euclidean distance to its mean value. However, when the input variables are categorical in nature, it is not possible to quantify the difference between two observations. In such instances, a probability-based clustering approach can still be undertaken, wherein a joint probability function can be parameterized which determines the likelihood of certain attributes belong to one cluster or another. In this instance, as our variables of interest are categorical in nature, a probability-based clustering approach is more suitable. As a result, we utilized latent class analysis as the clustering methodology for this project. This approach was motivated by Depaire et al. (25) who applied LCA to analyze crashes in Brussels Capital Region between 1997 and 1999.

LCA is a probabilistic clustering framework every cluster has a different underlying probability distribution from which its data elements are generated(25). In order to identify the different clusters, the probability distribution functions are parameterized, after which the problem of finding the clusters reduces to a parameter estimation problem. In order to estimate the parameters, we used the Expectation-

Maximization (EM) algorithm, which uses an iterative method to find the parameter estimates that can maximize the log-likelihood function. For more details on the EM algorithm, please refer to Ng (2009) (26).

3.2.1. Crash Variables of Interest

In order to detect patterns in a standardized manner, it is necessary to reduce the dimensionality of various crash attributes available within the TASAS crash data. A meaningful representation of the dimensionality reduction is a pedestrian crash typology which can be defined by a simultaneous occurrence of certain crash characteristics. It is important to note that while more well-defined crash typologies exist for automobile crashes (e.g. rear-end, broadside, sideswipe, etc.), similar definitions do not exist for pedestrian crashes. The variable selection process for defining the crash types had two primary guiding principles: (i) the variables should help summarize the dynamics of the crash (interaction between the modes), (ii) relevance of the crash dynamics for countermeasures. Based on these considerations, we used the following set of crash variables for crash typology development:

1. Movement preceding collision for automobiles:
    a. Proceeding straight
    b. Turning left or right
    c. Others (e.g., lane change, stopped/parked vehicle, etc.)
2. Movement preceding collision for pedestrians:
    a. Crossing at crosswalk—intersection
    b. Crossing not at crosswalk
    c. Roadway (including shoulder)
    d. Others (e.g., crossing at crosswalk—not intersection, not in roadway, approach/leave school bus)
3. Location of collision:
    a. Extreme lanes (right or left lane)
    b. Beyond lanes (right shoulder area, beyond shoulder (left or right))
    c. Others (e.g., interior lanes, gore area, etc.)
4. Lighting:
    a. Lighting absent (dark—no street light)
    b. Others (daylight, dusk/dawn, dark—street light, dark—inopr. street light, etc.)

Collectively, the variables chosen above help summarize the *context* of the crash: where it happened (using location of collision and attributes of pedestrian movement), how it happened (by a combination of automobile and pedestrian movement), and if any potential countermeasures can be applied (in this case, street lighting).

The variables chosen above do not include any facility-type desciptors (e.g., freeway vs arterial streets, segments vs intersection), as we intended for the crash typology to be generic and comparable across multiple facility types.

3.2.2. Latent Class Analysis

The process of identify meaningful crash clusters was undertaken as a two-step process. Since LCA requires the number of crash clusters as inputs (k), the first step involved applying LCA for different number of cluster combinations. Since the output of LCA is probabilistic classification of a crash to different clusters, as part of the step, we need to find the most dominant trends within each cluster to convert the probabilistic cluster definitions into deterministic cluster definitions. As a result, while LCA provides a probability for each crash to belong to one of the k clusters, at the second step there may some percentage of crashes that may not belong to any cluster since they do not meet all the criteria associated with that cluster. For instance, if a cluster is defined to have the following crash characteristics: (i) the motor vehicle was proceeding straight, (ii) the pedestrian was crossing not at the crosswalk, (iii) the location of the collision was in one of the extreme lanes, and (iv) lighting was not an issue. In this case, if a crash satisfied criteria (i), (ii) and (iv), but not (iii) (e.g., the crash took place in one of the interior lanes), then while the LCA may still classify it within this cluster with high probability, to ensure that the patterns within each cluster are consistently defined, we would exclude such a crash from this cluster definition.

Lastly, we determined the optimal number of clusters in a manner that they covered at least 5% of the total crashes. While such a model selection approach is statistically less rigorous, it increases the probability of observing each crash type within the typology to be observed across different HCCLs.

3.3. Data Set

In order to find clusters of similar crashes within the pedestrian crashes occurring along the California state highway system, we applied LCA on 4289 pedestrian crashes obtained from TASAS for the years 2009-2013. However, this dataset excluded crashes occurring on freeway segments as they are also not being considered within the hotspot identification process owing to the lack of good countermeasures to address them.

3.4. Results

The LCA-based clustering process resulted in the identification of 8 distinct crash types, as defined in Table 3-1. The crash typology can be further sub-classified by whether the movement preceding collision is through, turning or others, whereas within those sub-categories additional differences exist with regards to the pedestrian movements. While location of collision was included as a defining characteristic of some crash types, in other instances, especially when the pedestrian is also identified to be moving along the roadway, it was not deemed as essential for inferring the crash dynamics. Finally, lighting was only identified to be of significance in only one straight movement cluster. Collectively, the crash types covered 86.6% of all crashes in the dataset.

Table 3-1. Crash Typology

Cluster No.	Auto Movement	Pedestrian Movement	Location of Collision	Light Issue	%
1	Straight	Xing not at Xwalk	Extreme Lanes	No	17

2		Roadway including Shoulder	-	No	9
3		Roadway including Shoulder	-	Yes	5.4
4		Xing Xwalk at Ixn	Extreme Lanes	No	12
5	Turning	Xing Xwalk at Ixn	Beyond Lanes	No	14
6		Xing not at Inxn/Xwalk	-	No	5
7	Right Turn	Xing Xwalk at Ixn	Extreme Lanes	No	9.2
8	Other	-	-	-	15
				Total	**86.6**

While crash types 1-7 are informative with regards to providing the crash context, crash type 8 is relatively minimal in its description, since it only includes crashes wherein the motor vehicles are neither turning nor going straight prior to the collision. However, it was included in the current version of the typology so as to analyze the implications of a commonly observed motor vehicle movement being absent.

A visual representation of the crash types are shown in Figure 3-1.

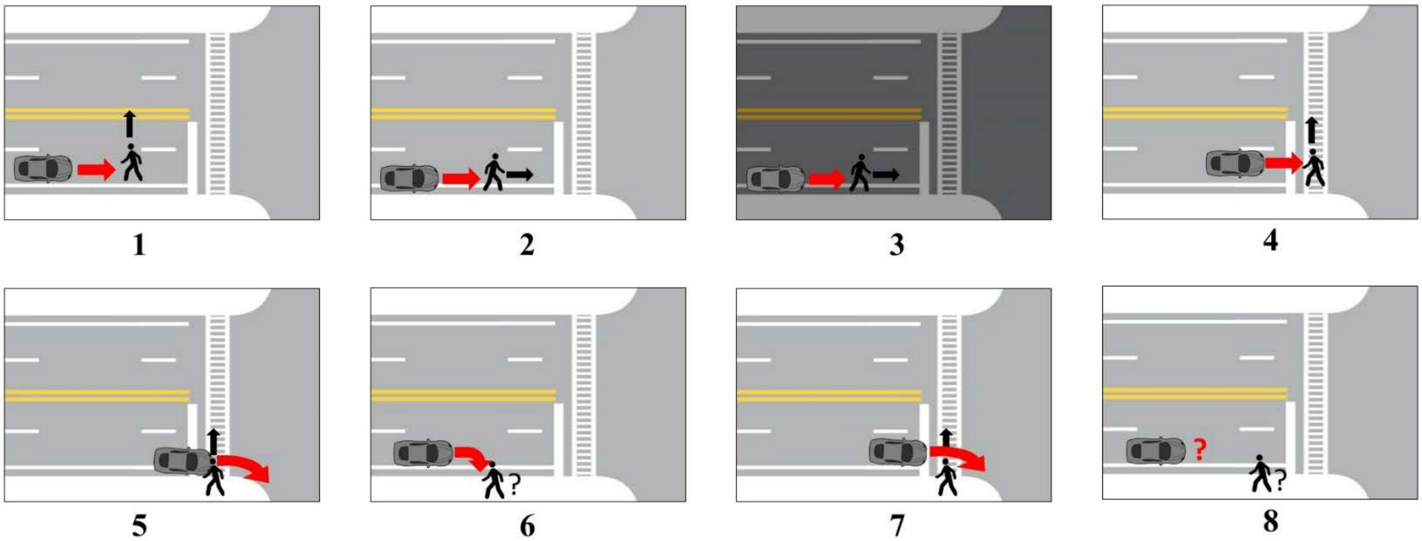


Figure3-1. Visual Representation of the pedestrian crash types

Tables 3-2 to 3-5 provide some summary statistics to supplement our understanding how these clusters differentiate themselves from each other. Table 3-2 shows the distribution of crashes within each crash type across different access controlled facilities. As is expected, a majority of pedestrian crashes across all crash types occur on conventional streets, where pedestrians typically have unrestricted access. The only exception to this trend is crash type 7, a turning movement cluster, which has 54% of crashes occurring at freeways. Considering that freeway segment crashes were excluded from the clustering process, it is expected that these freeway crashes occur along ramps. This hypothesis is confirmed by Table 3-3 which shows a distribution of highway segment/intersection/ramp crashes within each cluster. As per Table 3-3, 54% of all crashes in crash type 7 occur at ramps. Table 3-3 also reveals that most of the crashes in crash type 1-3 occur along highway segments which is consistent with the crash type definitions wherein the pedestrian movements are either crossing not in a crosswalk, or walking along the roadway. In general, the distributions reveal that a majority of pedestrian crashes occur along highway segments in conventional streets.

Table 3-2. Distribution of access controls across crash types

Crash Type	Conventional	Freeway	Expressway	One-Way City Streets	Grand Total
Not in a cluster	**71%**	25%	3%	2%	100%
1	**84%**	10%	4%	2%	100%
2	**65%**	30%	5%	0%	100%
3	**68%**	17%	15%	0%	100%
4	**77%**	17%	2%	4%	100%
5	**79%**	17%	2%	3%	100%
6	**66%**	30%	2%	2%	100%
7	46%	**54%**	0%	0%	100%
8	**69%**	25%	5%	1%	100%
Grand Total	**71%**	23%	4%	2%	100%

Table 3-3. Distribution of Highway-Int-Ramp across Clusters

Crash Type	H	I	R	Grand Total
Not in a cluster	41%	35%	24%	100%
1	**75%**	15%	10%	100%
2	**59%**	11%	30%	100%
3	**81%**	2%	17%	100%
4	46%	38%	17%	100%
5	41%	42%	17%	100%
6	39%	30%	31%	100%
7	8%	38%	**54%**	100%
8	**52%**	22%	25%	100%
Grand Total	**50%**	26%	23%	100%

Table 3-4 provides the distribution of different injury severity levels within each crash type. The results reveal that crash types involving motor vehicles traveling straight and pedestrians not crossing at intersections have a relatively higher share of fatal and severe injury collisions. In comparison, crashes that occur at/near intersections, which predominantly involve turning movements, involve relatively less severe injuries. We would expect such a difference to exist as there is a greater likelihood of vehicles traveling at higher speeds at the time of collision when traveling straight.

Table 3-4. Distribution of Injury Severity across Clusters

Crash Type	No Match	Fatal	Severe	Other Visible	Complaint of Pain	Grand Total
Not in a cluster	1%	6%	16%	36%	41%	100%
1	0%	**22%**	**28%**	35%	16%	100%
2	1%	**11%**	**25%**	34%	29%	100%
3	1%	**39%**	**22%**	23%	15%	100%
4	0%	6%	19%	36%	39%	100%
5	1%	1%	9%	38%	**50%**	100%
6	1%	3%	11%	37%	48%	100%
7	0%	0%	3%	41%	**55%**	100%
8	1%	7%	15%	35%	43%	100%
Grand Total	1%	10%	17%	35%	37%	100%

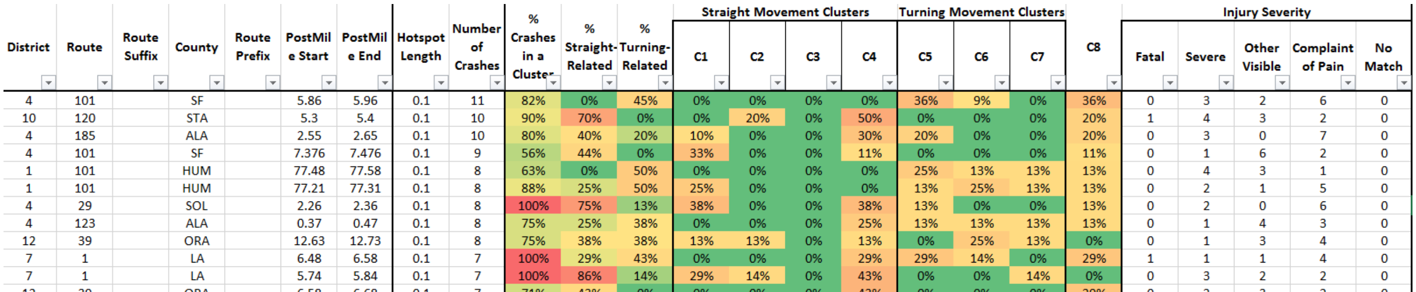
Finally, Table 3-5 shows the distribution of the crash types across each year of the crash data set. It reveals that the relative shares of the crash types are reasonably stable, which is encouraging as it indicates that the crash dynamics that are represented with each crash type is recurring in nature.

Table 3-5. Distribution of Clusters across Year

Crash Type	2009	2010	2011	2012	2013	Grand Total
Not in a cluster	23%	25%	23%	27%	26%	25%
1	19%	14%	16%	17%	20%	17%
2	7%	9%	10%	8%	9%	9%
3	5%	6%	5%	6%	7%	6%
4	5%	6%	6%	4%	4%	5%
5	13%	13%	12%	10%	11%	12%
6	6%	4%	5%	6%	5%	5%
7	7%	8%	7%	6%	5%	7%
8	15%	16%	16%	17%	12%	15%
Grand Total	100%	100%	100%	100%	100%	100%

3.5. Applying crash typology to HCCLs

Once we applied the crash typology to all pedestrian crashes, we analyzed their distribution within



												Straight Movement Clusters				Turning Movement Clusters				Injury Severity				
District	Route	Route Suffix	County	Route Prefix	PostMile Start	PostMile End	Hotspot Length	Number of Crashes	% Crashes in a Cluster	% Straight-Related	% Turning-Related	C1	C2	C3	C4	C5	C6	C7	C8	Fatal	Severe	Other Visible	Complaint of Pain	No Match
4	101		SF		5.86	5.96	0.1	11	82%	0%	45%	0%	0%	0%	0%	36%	9%	0%	36%	0	3	2	6	0
10	120		STA		5.3	5.4	0.1	10	90%	70%	0%	0%	20%	0%	50%	0%	0%	0%	20%	1	4	3	2	0
4	185		ALA		2.55	2.65	0.1	10	80%	40%	20%	10%	0%	0%	30%	20%	0%	0%	20%	0	3	0	7	0
4	101		SF		7.376	7.476	0.1	9	56%	44%	0%	33%	0%	0%	11%	0%	0%	0%	11%	0	1	6	2	0
1	101		HUM		77.48	77.58	0.1	8	63%	0%	50%	0%	0%	0%	0%	25%	13%	13%	13%	0	4	3	1	0
1	101		HUM		77.21	77.31	0.1	8	88%	25%	50%	25%	0%	0%	0%	13%	25%	13%	13%	0	2	1	5	0
4	29		SOL		2.26	2.36	0.1	8	100%	75%	13%	38%	0%	0%	38%	13%	0%	0%	13%	0	2	0	6	0
4	123		ALA		0.37	0.47	0.1	8	75%	25%	38%	0%	0%	0%	25%	13%	13%	13%	13%	0	1	4	3	0
12	39		ORA		12.63	12.73	0.1	8	75%	38%	38%	13%	13%	0%	13%	0%	25%	13%	0%	0	1	3	4	0
7	1		LA		6.48	6.58	0.1	7	100%	29%	43%	0%	0%	0%	29%	29%	14%	0%	29%	1	1	1	4	0
7	1		LA		5.74	5.84	0.1	7	100%	86%	14%	29%	14%	0%	43%	0%	0%	14%	0%	0	3	2	2	0

pedestrian HCCLs. While this feature is not currently available within the PSMR tool, we incorporated these modifications using MATLAB. We identified HCCLs using the sliding window method with a window length of 0.1 miles, and minimum crash threshold of 2 crashes. Based on the crashes identified within each HCCL, we can now summarize the distribution of different crash types within each HCCL, as shown in a sample output in Figure 3-2.

Figure3-2.A sample output of HCCLs sorted by fatal+injury along with the crash typology distribution

The application of the sliding window method on the 2009-2013 crash data yielded 815 HCCLs covering 2126 crashes. In comparison, the input crash dataset included 4289 crashes. Thus, it is possible that distribution of crash types, as observed in the crash population may differ from the distribution of crashes that are captured within HCCLs. Table 3-6 provides a comparison of the distribution of crash types observed in the HCCLs along with the total pedestrian crash population considered for this analysis. The comparison reveals that straight movement crash types are underrepresented in HCCLs, whereas the turning movement crash types as well as the crash type with neither straight nor turning movements have comparable representation in HCCLs.

Table 3-6. Comparing distribution of pedestrian crash types in HCCLs and crash population

Crash Type	In HCCLs	In Crash Population
1	16.2%	17.0%
2	6.4%	9.0%
3	2.8%	5.4%
4	5.9%	12.0%
5	13.7%	14.0%
6	4.8%	5.0%
7	7.8%	9.2%
8	13.9%	15.0%
All Straight Movement Crash Types	31.4%	43.4%
All Turning Movement Crash Types	26.2%	28.2%
All Crash Types	71.5%	86.6%

In the case of crash types 2 and 3, it is reasonable to expect that the likelihood of multiple crashes involving pedestrians walking along the roadway within the same segment of 0.1 miles is low when compared to the overall population which does not impose any spatial constraints. But in the case of crash type 4, which involves a pedestrian crossing at an intersection crosswalk, it is unclear why there would be a drop-off of 6%. The nature of this relationship between crash types in the population vs within HCCLs would be explored further in future extensions of this work.

Chapter 4. Pedestrian Safety Toolkit

This chapter provides an overview of the updates made to the pedestrian safety monitoring report (PSMR) tool during this phase of PSIP.

4.1. Roadmap of Pedestrian Safety Tools

Table 4-1 summarizes the roadmap of the different pedestrian safety-related tools that were developed during this project. PSMR tool is the primary tool for identifying and prioritizing pedestrian HCCLs. It imports TASAS crash data and conducts network screening using either sliding window method or dynamic programming. However, the research team has revised the functionality of the tool significantly, as indicated by the version number changing from v1.x to v2.x, based on feedback from Caltrans users as well as other issues identified by the research team. In particular, we made four significant changes to the overall functionality of the network screening process:

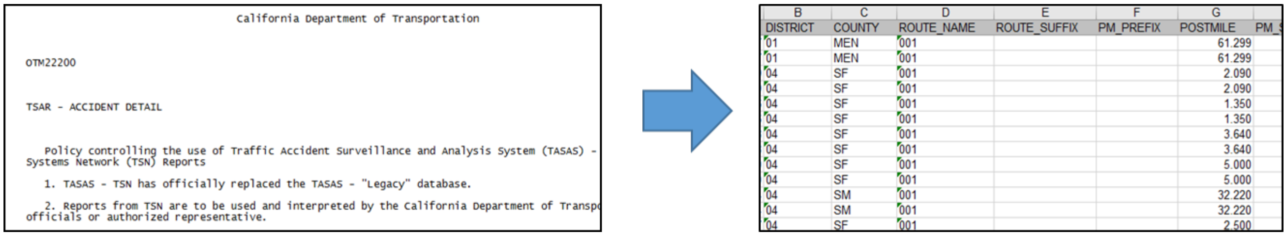
1. We divided the data crash data importing feature into two distinct processes:
   - When using Excel-based files that are obtained from the TASAS database directly, only the PSMR tool (v2.x) is required for importing the data (as well as conducting SWITRS matching).
   - When using text-based files obtained from TASAS Selective Accident Retrieval (TSAR), the TSAR2XLS tool will be required to first convert the text files into an Excel-based file, which can be then be used as an input into PSMR (v2.x).
2. We improved the error handling capabilities within TSAR2XLS tool to better address missing data within the TSAR files.
3. We modified the network screening process to analyze left and right independent alignments separately.
4. We improved the SWITRS matching functionality to speed it up when running large files, as well as allow matching to be undertaken without requiring postmile information as matching criterion.

Table 4-1. Roadmap of the various pedestrian safety-related tools developed within the project

	V 1.X			V 2.X		
	Data sources	**Functionality**	**Version/Date**	**Data sources**	**Functionality**	**Version/ Date**
PSMR	TSAR SWITRS (optional)	- Import data - SWITRS matching - Generate a list of district-level pedestrian HCCL's - Sliding Window - Dynamic Programming.	V1.3 07/14/17	**TASAS-TSN Input (preferred)** **TSAR-2-XLS Input** SWITRS	**- Improved DP** **- Combine D/UD** **- Highway Group**	V2.0 10/02/17 V2.1 *Modified SWITRS matching*
TSAR2XLS	TSAR SWITRS (optional)	- Converts TSAR text files to Excel - Error handling - Error log -Faster SWITRS matching	V1.0 10/06/17 V1.1 *Modified SWITRS matching*			

4.2. TSAR2XLS

The objective of creating a separate tool for importing TSAR files was to transition from using text-based TSAR files, which are themselves created from within TASAS, to using Excel-based pedestrian crash data files from TASAS directly. Such a transition is motivated by reducing the likelihood of error-prone data input which is possible when creating text files from a database, and then converting it back into a database-friendly Excel file. However, we recognize that generating TSAR files is relatively easier for the Caltrans stakeholders in comparison to obtaining TASAS data directly. Thus, TSAR2XLS provides this continuity in functionality of importing TSAR files, while transitioning the PSMR tool to also input TASAS-based Excel worksheets.

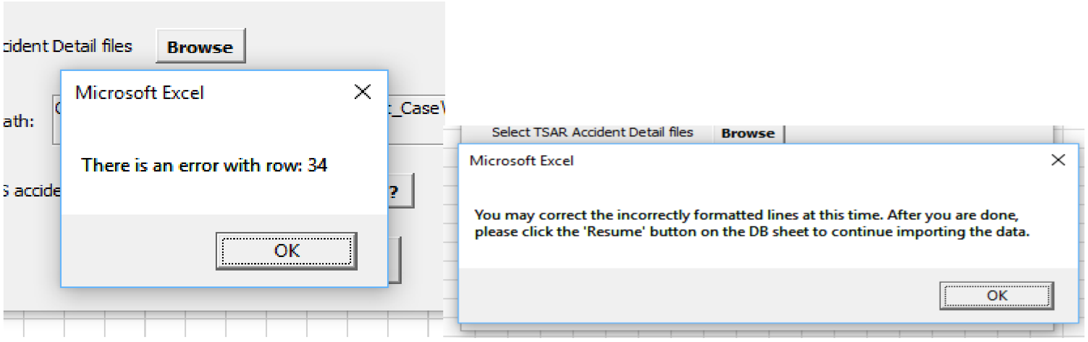


```
                                California Department of Transportation

OTM22200

TSAR - ACCIDENT DETAIL

   Policy controlling the use of Traffic Accident Surveillance and Analysis System (TASAS) -
Systems Network (TSN) Reports

   1. TASAS - TSN has officially replaced the TASAS - "Legacy" database.

   2. Reports from TSN are to be used and interpreted by the California Department of Transpo
officials or authorized representative.
```

B	C	D	E	F	G	
DISTRICT	COUNTY	ROUTE_NAME	ROUTE_SUFFIX	PM_PREFIX	POSTMILE	PM_
01	MEN	001			61.299	
01	MEN	001			61.299	
04	SF	001			2.090	
04	SF	001			2.090	
04	SF	001			1.350	
04	SF	001			1.350	
04	SF	001			3.640	
04	SF	001			3.640	
04	SF	001			5.000	
04	SF	001			5.000	
04	SM	001			32.220	
04	SM	001			32.220	
04	SF	001			2.500	

Figure 4-1. TSAR2XLS converts text-based TSAR files to an Excel worksheet

Figure 4-1 illustrates the key functionality of the TSAR2XLS tool, which is to convert text-based TSAR accident detail files into Excel files. However, while the importing of text data into Excel is largely error-free, it is possible that some missing data within a few columns of the text file may disrupt the structure of the data, thus interrupting the import process. The shortcomings of an error identified within PSMR v1.3, which provided the TSAR import feature, is shown in Figure 4-2. Herein, the user importing a TSAR file is alerted of an error in row 34 of the file, and is requested to intervene to make changes to the file so as to make corrections within the file. However, the intermediate worksheet shown to the user does not provide any capabilities regarding the type of formatting error or how to correct it.



Microsoft Excel

There is an error with row: 34

OK

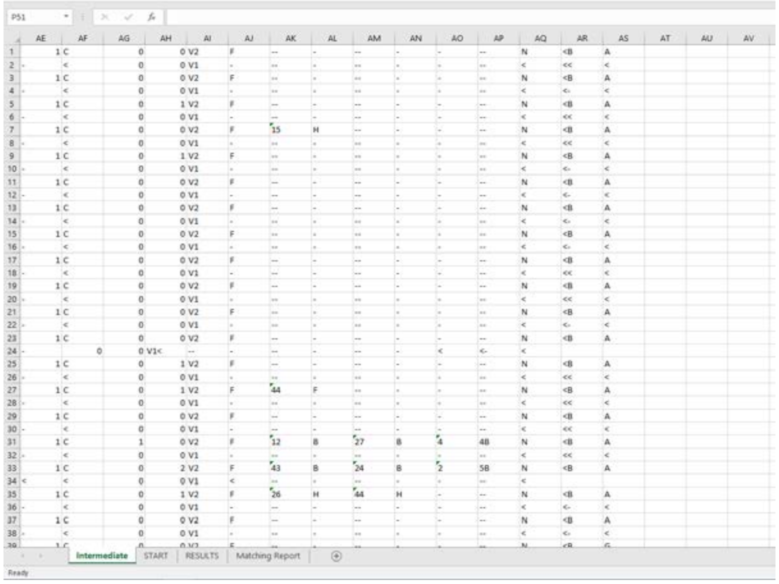
Microsoft Excel

You may correct the incorrectly formatted lines at this time. After you are done, please click the 'Resume' button on the DB sheet to continue importing the data.

OK

(a) PSMR (v1.3) detects an error in file

(b) A prompt indicates the need for user intervention

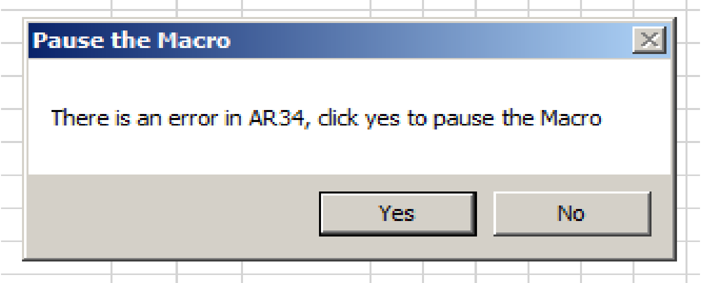


	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV
1	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
2	-	<	0	0	V1	-	--	-	--	-	-	--	<	<<	<			
3	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
4	-	<	0	0	V1	-	--	-	--	-	-	--	<	<-	<			
5	1	C	0	1	V2	F	--	-	--	-	-	--	N	<B	A			
6	-	<	0	0	V1	-	--	-	--	-	-	--	<	<<	<			
7	1	C	0	0	V2	F	15	H	--	-	-	--	N	<B	A			
8	-	<	0	0	V1	-	--	-	--	-	-	--	<	<<	<			
9	1	C	0	1	V2	F	--	-	--	-	-	--	N	<B	A			
10	-	<	0	0	V1	-	--	-	--	-	-	--	<	<-	<			
11	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
12	-	<	0	0	V1	-	--	-	--	-	-	--	<	<-	<			
13	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
14	-	<	0	0	V1	-	--	-	--	-	-	--	<	<-	<			
15	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
16	-	<	0	0	V1	-	--	-	--	-	-	--	<	<-	<			
17	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
18	-	<	0	0	V1	-	--	-	--	-	-	--	<	<<	<			
19	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
20	-	<	0	0	V1	-	--	-	--	-	-	--	<	<<	<			
21	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
22	-	<	0	0	V1	-	--	-	--	-	-	--	<	<-	<			
23	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
24	-	0	0	V1<	--	-	--	-	--	-	<	<-	<					
25	1	C	0	1	V2	F	--	-	--	-	-	--	N	<B	A			
26	-	<	0	0	V1	-	--	-	--	-	-	--	<	<<	<			
27	1	C	0	1	V2	F	44	F	--	-	-	--	N	<B	A			
28	-	<	0	0	V1	-	--	-	--	-	-	--	<	<<	<			
29	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
30	-	<	0	0	V1	-	--	-	--	-	-	--	<	<<	<			
31	1	C	1	0	V2	F	12	B	27	B	4	4B	N	<B	A			
32	-	<	0	0	V1	-	--	-	--	-	-	--	<	<<	<			
33	1	C	0	2	V2	F	43	B	24	B	2	5B	N	<B	A			
34	<	<	0	0	V1	<	--	-	--	-	-	--	<					
35	1	C	0	1	V2	F	26	H	44	H	-	--	N	<B	A			
36	-	<	0	0	V1	-	--	-	--	-	-	--	<	<-	<			
37	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			
38	-	<	0	0	V1	-	--	-	--	-	-	--	<	<-	<			
39	1	C	0	0	V2	F	--	-	--	-	-	--	N	<B	A			

(c) Non-responsive intermediate sheet

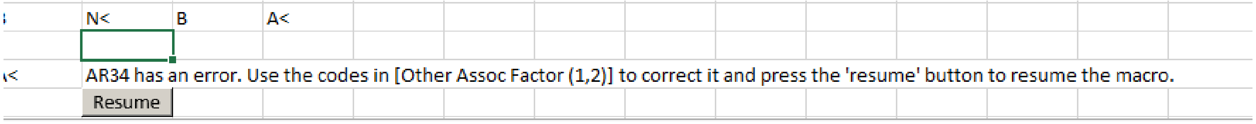
Figure 4-2. Error management in PSMR (v1.3)

We significantly improved the error handling in TSAR2XLS (v1.0) in three distinct ways, as shown in Figure 4-3. Firstly, when the tool finds an error, it specifically points which cell has an error (AR34), as opposed to the name of the row. Secondly, it provides specific instructions about the type of content that was missing in that cell so that the user can enter an appropriate code, after which the user can press a button to resume the import process. Finally, after the text files have been successfully imported, TSAR2XLS (v1.0) generates an error log to document what changes were made the TSAR files in the process of importing them into the tool. These modifications to the data import process make the error handling user-friendly and transparent.



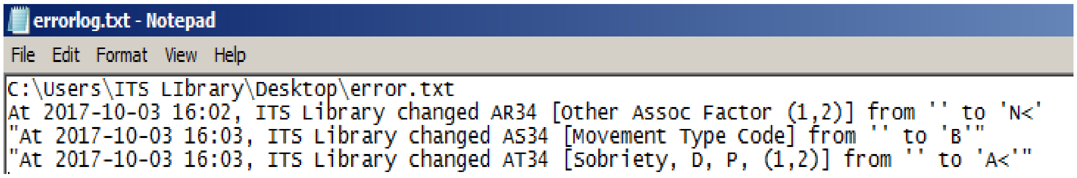
There is an error in AR34, click yes to pause the Macro

(a) TSAR2XLS detects an error in file



AR34 has an error. Use the codes in [Other Assoc Factor (1,2)] to correct it and press the 'resume' button to resume the macro.

(b) TSARSXLS provides specific instructions to enter missing data and resume the import process



```
C:\Users\ITS LIbrary\Desktop\error.txt
At 2017-10-03 16:02, ITS Library changed AR34 [Other Assoc Factor (1,2)] from '' to 'N<'
"At 2017-10-03 16:03, ITS Library changed AS34 [Movement Type Code] from '' to 'B'"
"At 2017-10-03 16:03, ITS Library changed AT34 [Sobriety, D, P, (1,2)] from '' to 'A<'"
```

(c) TSARSXLS generates an error log to record all changes made

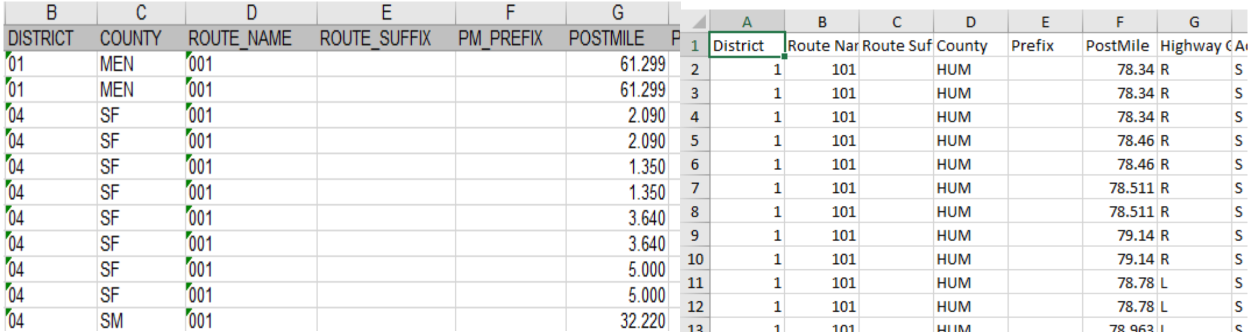
Figure 4-3. Error management in TSAR2XLS (v1.0)

Once the TSAR files are imported in the TSAR2XLS tool, it can be exported as an Excel file compatible with PSMR Tool (v2). TSAR2XLS also allows for SWITRS matching to be undertaken. However, that functionality is also available in PSMR (v2) in case the user does not select the SWITRS matching option during TSAR file import process.

4.3. Pedestrian Safety Monitoring Report Tool

4.3.1. Excel-based data import functionality

While PSMR (v1) used TSAR files as input crash data, PSMR (v2) uses only Excel-based files. Figure 4-4 shows the two types of Excel files that are expected to act as input crash files: (i) An Excel file obtained from TASAS directly, or (ii) an output of TSAR2XLS which converts TSAR files into an Excel worksheet. As Figure 4-4 indicates, the header row of the two files have different column numbers and formatting styles. However, the PSMR tool has been modeled to search for relevant columns by keywords, which provides greater flexibility when handling different types of Excel files. However, given the differences in the templates of the two Excel files, the tool does not allow files of different formats to be merged together. Thus, it is important to ensure the user chooses the input crash data from one data source (TASAS or TSAR).



B	C	D	E	F	G	P
DISTRICT	COUNTY	ROUTE_NAME	ROUTE_SUFFIX	PM_PREFIX	POSTMILE	
01	MEN	001			61.299	
01	MEN	001			61.299	
04	SF	001			2.090	
04	SF	001			2.090	
04	SF	001			1.350	
04	SF	001			1.350	
04	SF	001			3.640	
04	SF	001			3.640	
04	SF	001			5.000	
04	SF	001			5.000	
04	SM	001			32.220	

	A	B	C	D	E	F	G	
1	District	Route Nar	Route Suf	County	Prefix	PostMile	Highway (	A
2	1	101		HUM		78.34	R	S
3	1	101		HUM		78.34	R	S
4	1	101		HUM		78.34	R	S
5	1	101		HUM		78.46	R	S
6	1	101		HUM		78.46	R	S
7	1	101		HUM		78.511	R	S
8	1	101		HUM		78.511	R	S
9	1	101		HUM		79.14	R	S
10	1	101		HUM		79.14	R	S
11	1	101		HUM		78.78	L	S
12	1	101		HUM		78.78	L	S
13	1	101		HUM		78.963	L	S

(a) Excel file from TASAS

(b) Output of TSAR2XLS

Figure 4-4.Two types of Excel-based crash data files

4.3.2. Resolving road alignment issues during network screening

One of the issues identified in the network screening output in PSMR (v1.3) was that it combined crash data from road segments with different independent alignments. Figure 4-5 shows an example of a route which has the left and right independent alignments that are physically separate from each other. In this case, if the differences in alignment are not recognized, the network screening may combine crashes from the two segments and search for HCCLs based on the proximity of their postmiles.

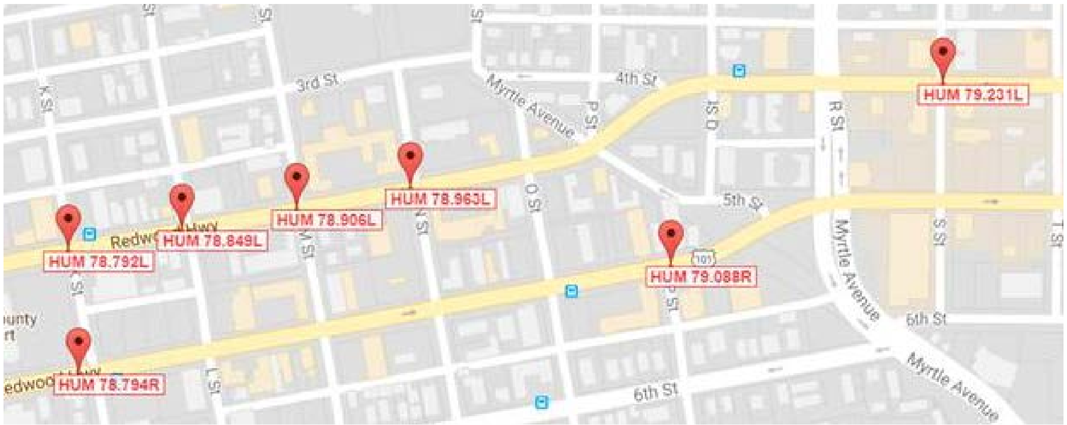
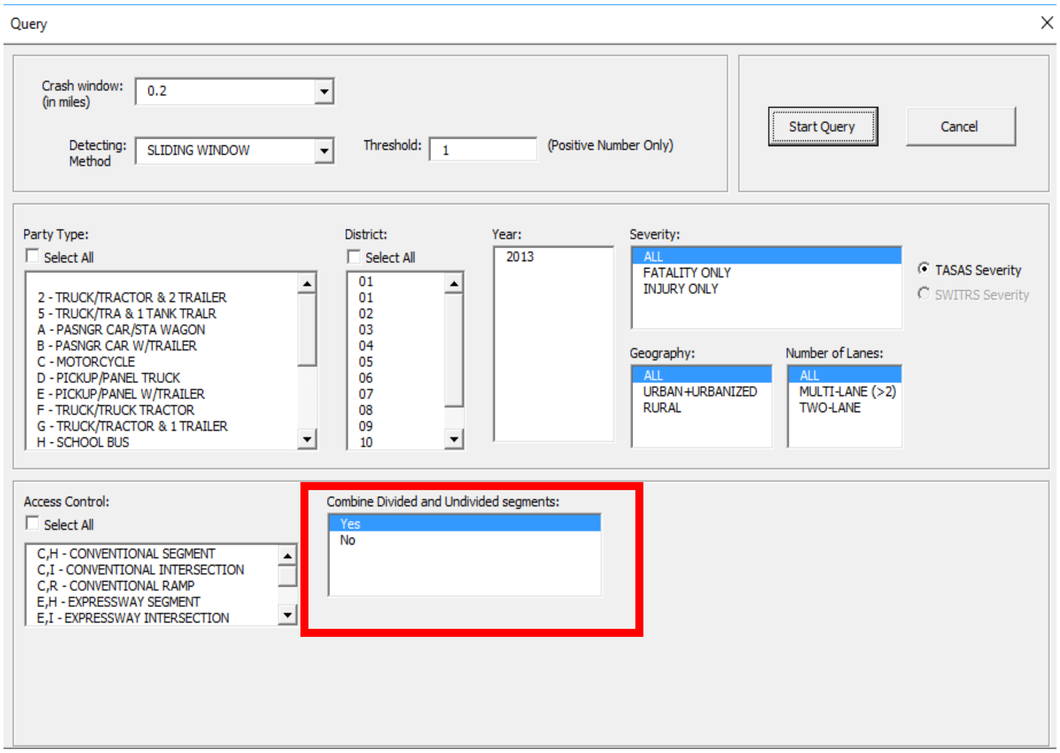


Figure 4-4. An example of a route with left and right independent alignments

To address this issue, we modified the network screening algorithm in PSMR (v2.0) to distinguish between left and right independent alignments, which can be identified using the column named “Highway Group” in the crash database. However, the highway group attribute also includes indicators corresponding to whether the underlying road segment is divided or undivided. Unlike left/right independent alignment, PSMR (v2) allows for the Caltrans expert user to choose whether divided and undivided segments can be combined for the purposes of HCCL identification (Figure 4-6).


Query

Crash window: (in miles) 0.2

Detecting: Method SLIDING WINDOW

Threshold: 1 (Positive Number Only)

Start Query
Cancel

Party Type:
□ Select All
2 - TRUCK/TRACTOR & 2 TRAILER
5 - TRUCK/TRA & 1 TANK TRALR
A - PASNGR CAR/STA WAGON
B - PASNGR CAR W/TRAILER
C - MOTORCYCLE
D - PICKUP/PANEL TRUCK
E - PICKUP/PANEL W/TRAILER
F - TRUCK/TRUCK TRACTOR
G - TRUCK/TRACTOR & 1 TRAILER
H - SCHOOL BUS

District:
□ Select All
01
01
02
03
04
05
06
07
08
09
10

Year:
2013

Severity:
ALL
FATALITY ONLY
INJURY ONLY

● TASAS Severity
○ SWITRS Severity

Geography:
ALL
URBAN+URBANIZED
RURAL

Number of Lanes:
ALL
MULTI-LANE (>2)
TWO-LANE

Access Control:
□ Select All
C,H - CONVENTIONAL SEGMENT
C,I - CONVENTIONAL INTERSECTION
C,R - CONVENTIONAL RAMP
E,H - EXPRESSWAY SEGMENT
E,I - EXPRESSWAY INTERSECTION

Combine Divided and Undivided segments:
Yes
No

Figure 4-5.Modified network screening query window in PSMR (v2.0)

4.4. Modifications to SWITRS Matching

In addition to the functionalities described above, both TSAR2XLS (v1.0) and PSMR (v2.0) include the same SWITRS matching functionality as found in PSMR (v1.3). While updating these tools, we also wanted to update the SWITRS matching functionality for two different reasons. Firstly, the matching feature in PSMR (v1.3) was slow. Secondly, it used postmile information as part of the attributes considered for finding a common crash in SWITRS and TASAS. However, recent changes in SWITRS documentation may result in postmile information not being available in SWITRS crash data.

Thus, to address these issues, the SWITRS matching algorithm has been updated in TSAR2XLS (v1.1) and PSMR (v2.1), which utilizes the following attributes to find a matching crash in SWITRS and TASAS:

- Date of crash (available as is in both SWITRS and TASAS)
- Time of crash (available as is in both SWITRS and TASAS)
- Concatenated string of JURIS and BADGE in SWITRS compared with “Common Accident Number” in TASAS

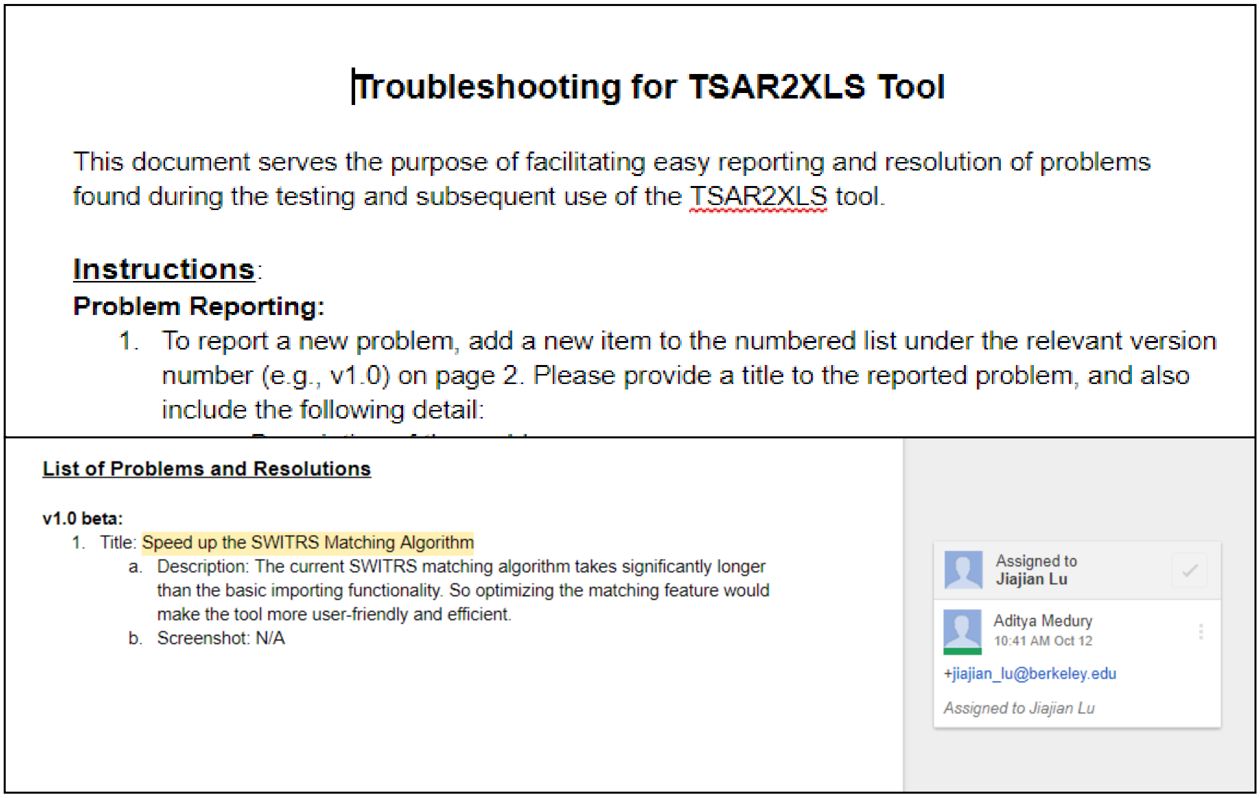
We tested the computational efficiency of the modified matching algorithm using test cases with varying number of records. We compare the running times of TSAR2XLS (v1.0), which uses the same matching algorithm as PSMR (v1.3), and TSAR2XLS (v1.1). Table 4-2 reveals that while the old matching algorithm is much faster when the test case included just 25 records, the new matching function scales much better as the file sizes increase. The reason for the inferior performance of new matching function for the first test case is that the revised matching algorithm uses a bigger SWITRS file which includes a larger set of attributes. In comparison, the previous version of the matching algorithm utilized SWITRS files that did not contain attributes such as JURIS and BADGE. Since there is a higher fixed cost (in computational time) for opening a bigger file, TSAR2XLS (v1.1) performs slightly worse than TSAR2XLS (v1.0) in the case of n=25.

Table 4-2. Running times for SWITRS matching function for different test cases (n corresponds to the number of records)

	n = 25	**n = 1247**	**n = 21467**
TSAR2XLS (v1.0)	3.59s	139.53s	2567.67s
TSAR2XLS (v1.1)	60s	125.72s	1408.45s

4.5. Troubleshooting for PSMR and TSAR2XLS

In addition to the tools discussed above, we have also initiated a web-based troubleshooting mechanism which allows for Caltrans expert users to report any issues that they face while using TSAR2XLS or PSMR tools back to the project team at SafeTREC. We developed the template for documenting these issues using Google Documents, which allows a user to explain a problem and assign it to one of the project team members. Once the issues is resolved, the problem can be marked as resolved, which will subsequently be conveyed to the person who reported the problem. Figure 4-7 shows snapshots of the troubleshooting template for TSAR2XLS.



Troubleshooting for TSAR2XLS Tool

This document serves the purpose of facilitating easy reporting and resolution of problems found during the testing and subsequent use of the TSAR2XLS tool.

Instructions:

Problem Reporting:

1. To report a new problem, add a new item to the numbered list under the relevant version number (e.g., v1.0) on page 2. Please provide a title to the reported problem, and also include the following detail:

List of Problems and Resolutions

v1.0 beta:

1. Title: Speed up the SWITRS Matching Algorithm
    a. Description: The current SWITRS matching algorithm takes significantly longer than the basic importing functionality. So optimizing the matching feature would make the tool more user-friendly and efficient.
    b. Screenshot: N/A

Assigned to
Jiajian Lu

Aditya Medury
10:41 AM Oct 12

+jiajian_lu@berkeley.edu

Assigned to Jiajian Lu

Figure 4-6.Troubleshooting template for TSAR2XLS

Chapter 5. Conclusions and Recommendations

This report aimed to create a comprehensive picture of pedestrian safety in California, as well as to continue and support the efforts for implementing a Pedestrian Safety Improvement Program in California. Each chapter in this report describes an activity that contributes to the overall strategy to enhance pedestrian safety in California. A concise summary, important insights, and some recommendations for each chapter are provided below:

Chapter 2 – Pedestrian Exposure Model described the process to develop a state-scale pedestrian exposure model for the California State Highway System (SHS). The report explains the scope of the model, the data that was collected, and the analytical and modeling assumptions that were used to produce the annual volume estimates.

Keyinsights:

- Local agencies have data that are beneficial for larger scale modeling projects. The project team was able to utilize such counts for the purpose of this project.
- The project team developed a framework to expand short term counts that are routinely conducted by Districts to annual pedestrian volumes. This framework allows Caltrans to convert short term counts of various durations to a common unit of observation of annual volumes.
- The team developed a direct demand model for estimation of annual pedestrian volumes on the state highway system. The model identifies the relationship between land-use and other variables about the surrounding environment and the expanded intersection counts.
- The research team applied the model to estimate annual pedestrian volumes at all applicable (count and non-count locations) on the California state highway system.
- Potential enhancements can include:
    - Improve classification of expansion factor groups for count locations. This can include improving the identification of outlier data in long term count data sets, and/or develop a tool to automate selection of factor groups and application of expansion factors.
    - Test alternative direct-demand model specifications and approaches to reducing model error and improve robustness.

Recommendations:

- Implement a repository to store data used to develop the pedestrian exposure model. An effort to propose the desired specification of such a repository is already being conducted.
- Recommend updating and re-estimating the pedestrian exposure model every 5 years. In this period, there should be additional Miovision count data from the PSMR-related site investigations, other District count activities, and counts from local jurisdictions. Data used for explanatory will also likely be updated in this period.

Chapter 3 – Contextualized Hotspot Clustering described a clustering approach to develop a pedestrian crash typology for the state highway system and evaluates the distribution of the proposed crash types within the crash population as well as within pedestrian HCCLs.

Key insights:

- Developed a crash typology to distinguish between different pedestrian crash dynamics occurring along the state highway system.
- The typology can be used to identify crash types that are recurring at specific locations.

Chapter 4 – Pedestrian Safety Toolkit described the enhancements made to the pedestrian safety monitoring report (PSMR) tool, along with the modifications made to the crash data import and SWITRS matching processes.

Key insight:

- Revised Pedestrian Safety Monitoring Report Tool to include better error handling capabilities and improve HCCL identification.

Recommendations:

- Develop a method to incorporate pedestrian crash typology into thePSMR toolto prioritize high collision concentration locations.
- Develop pedestrian intersection Safety Performance Functions for California for future incorporation into the PSMR tool.

References

1. Kuzmyak, J.R., J. Walters, M. Bradley, and K.M. Kockelman. Estimating Bicycling and Walking for Planning and Project Development: A Guidebook. National Cooperative Highway Research Program Report 770, Transportation Research Board, 2014.

2. Clifton, K.J., C.V. Burnier, S. Huang, M.W. Kang, and R. Schneider. "A Meso-Scale Model of Pedestrian Demand," Presented at the 4th Joint Meeting of the Association of Collegiate Schools of Planning and the Association of European Schools of Planning, Chicago, IL, July 6-11, 2008.

3. Clifton, K.J., P.A. Singleton, C.D. Muhs, and R.J. Schneider. Development of a Pedestrian Demand Estimation Tool, NITC-RR-677, National Institute for Transportation and Communities, Available online, http://ppms.otrec.us/media/project_files/NITC-RR-677_Final_Report.pdf, 2015.

4. Raford, N. and D. Ragland. "Space Syntax: Innovative Pedestrian Volume Modeling Tool for Pedestrian Safety," Transportation Research Record: Journal of the Transportation Research Board, Volume 1878, Washington D.C., pp. 66-74, 2004.

5. Raford, N. and D. Ragland. Pedestrian Volume Modeling for Traffic Safety and Exposure Analysis. University of California Traffic Safety Center white paper, Available online: http://repositories.cdlib.org/its/tsc/UCB-TSC-RR-2005-TRB2/. December 2005.

6. Pushkarev, B. and J. Zupan. "Pedestrian Travel Demand," Highway Research Record 355, Washington, D.C., 1971.

7. Benham, J. and B. G. Patel. "A Method for Estimating Pedestrian Volume in a Central Business District," Transportation Research Record 629, Transportation Research Board, Washington D.C., pp. 22-26, 1977.

8. Hankey, S., G. Lindsey, X. Wang, J. Borah, K. Hoff, B. Utecht, and Z. Xu. "Estimating Use of Non-Motorized Infrastructure: Models of Bicycle and Pedestrian Traffic in Minneapolis, MN," Landscape and Urban Planning, Volume 107, pp. 307-316, 2012.

9. Hankey, S. and G. Lindsey. "Facility-Demand Models of Peak-Period Pedestrian and Bicycle Traffic: A Comparison of Fully-Specified and Reduced-Form Models," Presented at Transportation Research Board Annual Meeting, Washington, DC, 2016.

10. Pulugurtha, S.S. and Repaka, S.R. "Assessment of Models to Measure Pedestrian Activity at Signalized Intersections," Transportation Research Record: Journal of the Transportation Research Board, Volume 2073, pp. 39-48, 2008.

11. Schneider R.J., L.S. Arnold, and D.R. Ragland. "A Pilot Model for Estimating Pedestrian Intersection Crossing Volumes," Transportation Research Record: Journal of the Transportation Research Board, Volume 2140, pp. 13-26, 2009.

12. Liu, X. and J. Griswold. "Pedestrian Volume Modeling: A Case Study of San Francisco," Association of Pacific Coast Geographers Yearbook, Volume 71, 2009.

13. Haynes, M. and S. Andrzejewski. GIS Based Bicycle & Pedestrian Demand Forecasting Techniques, Presentation for US Department of Transportation, Travel Model Improvement Program, Fehr & Peers Transportation Consultants, April 29, 2010.

14. Jones, M.G., S. Ryan, J. Donlan, L. Ledbetter, L. Arnold, and D. Ragland. Seamless Travel: Measuring Bicycle and Pedestrian Activity in San Diego County and its Relationship to Land Use, Transportation, Safety, and Facility Type, Prepared by Alta Planning & Design and UC Berkeley Safe Transportation Research & Education Center, California Department of Transportation Task Order 6117, 2010.

15. Miranda-Moreno, L.F. and D. Fernandes. "Pedestrian Activity Modelling at Signalized Intersections: Land Use, Urban Form, Weather, and Spatio-Temporal Patterns," Transportation Research Record: Journal of the Transportation Research Board, Forthcoming, 2011.

16. Schneider, R.J., T. Henry, M.F. Mitman, L. Stonehill, and J. Koehler. "Development and Application of the San Francisco Pedestrian Intersection Volume Model," Transportation Research Record: Journal of the Transportation Research Board, Volume 2299, pp. 65-78, 2012.

17. Grembek, O., C. Bosman, J.M. Bigham, S. Fine, J.B. Griswold, A. Medury, R.L. Sanders, R.J. Schneider, A. Yavari, Y. Zhang, and D.R. Ragland. Pedestrian Safety Improvement Program: Final Technical Report, Prepared by the UC Berkeley Safe Transportation Research and Education Center for the California Department of Transportation, March 31, 2014.

18. Schneider, R.J., L.S. Arnold, and D.R. Ragland. "A Methodology for Counting Pedestrians at Intersections: Using Automated Counters to Extrapolate Weekly Volumes from Short Manual Counts," Transportation Research Record 2140, pp. 1-12, 2009.

19. Nordback, K., W.E. Marshall, B.N. Janson, and E. Stolz. "Estimating Annual Average Daily Bicyclists: Error and Accuracy," Transportation Research Record: Journal of the Transportation Research Board, Volume 2339, pp. 90-97, 2013.

20. Nordback, K., W.E. Marshall, and B.N. Janson. "Development of Estimation Methodology for Bicycle and Pedestrian Volumes Based on Existing Counts," Colorado Department of Transportation, Available online, http://www.coloradodot.info/programs/research/pdfs/2013/bikecounts.pdf/view, October 2013.

21. Nosal, T., L.F. Miranda-Moreno, and Z. Krstulic. "Incorporating Weather: A Comparative Analysis of Average Annual Daily Bicyclist Estimation Methods," Transportation Research Record: Journal of the Transportation Research Board, Volume 2468, pp. 100-110, 2014.

22. Ryus, P., E. Ferguson, K.M. Laustsen, R.J. Schneider, F.R. Proulx, T. Hull, and L. Miranda-Moreno. Methods and Technologies for Collecting Pedestrian and Bicycle Volume Data: Guidebook on Pedestrian and Bicycle Volume Data Collection, National Cooperative Highway Research Program Report 797, Available online, http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_797.pdf, 2014.

23. Liggett, R., H. Huff, R. Taylor-Gratzer, N. Wong, D. Benitez, T. Douglas, J. Howe, J. Cooper, J. Griswold, D. Amos, and F. Proulx. Bicycle Crash Risk: How Does It Vary, and Why? California Department of Transportation, Caltrans Task No. 2801, Available online, http://www.lewis.ucla.edu/wp-content/uploads/sites/2/2016/08/Final-Report-to-Caltrans-Bicycle-Crash-v3.pdf, 2016.

24. Griswold, J. B., A. Medury, R.J. Schneider, and O. Grembek. "Comparison of Pedestrian Count Expansion Methods: Land Use Groups Versus Empirical Clusters", Accepted for publication in the Transportation Research Record (No. 18-05579), 2018.

25. Depaire, B., G. Wets, and K. Vanhoof. Traffic accident segmentation by means of latent class clustering. *Accident Analysis & Prevention*, *40*(4), pp.1257-1266, 2008.

26. Ng, A. The EM Algorithm. CS229 Lecture Notes, Stanford University, 2009. URL: http://cs229.stanford.edu/notes/cs229-notes8.pdf

Appendix A – Scatter plots of model variables

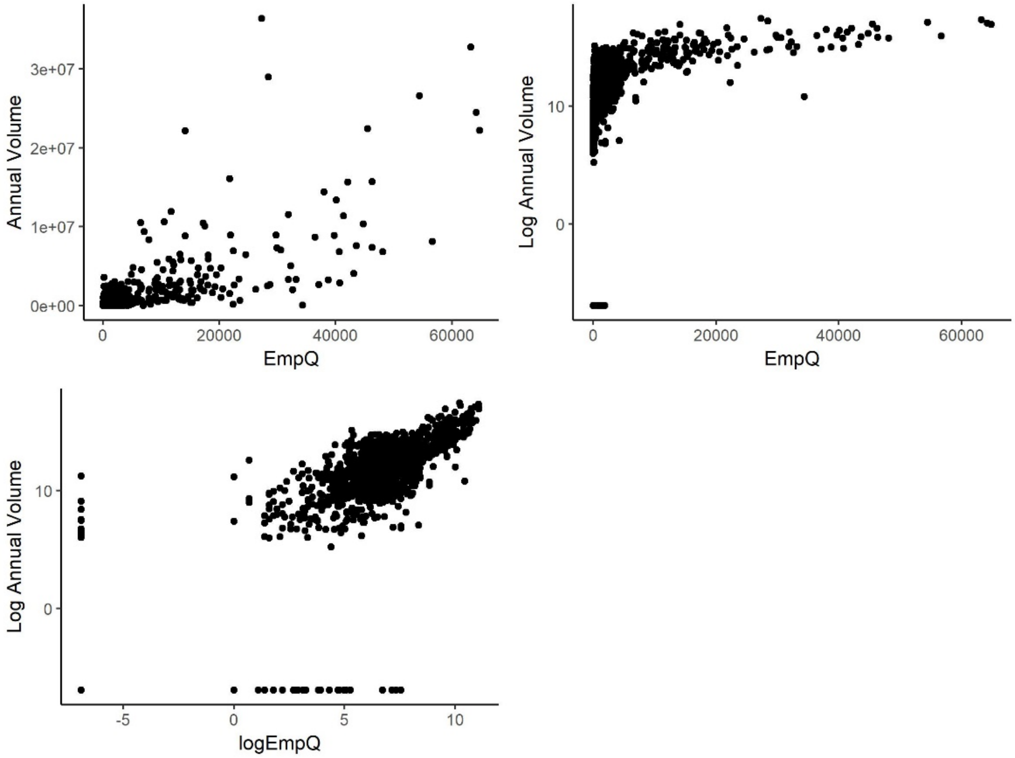


Figure 0-1. Scatter plots for number of employees within a quarter mile

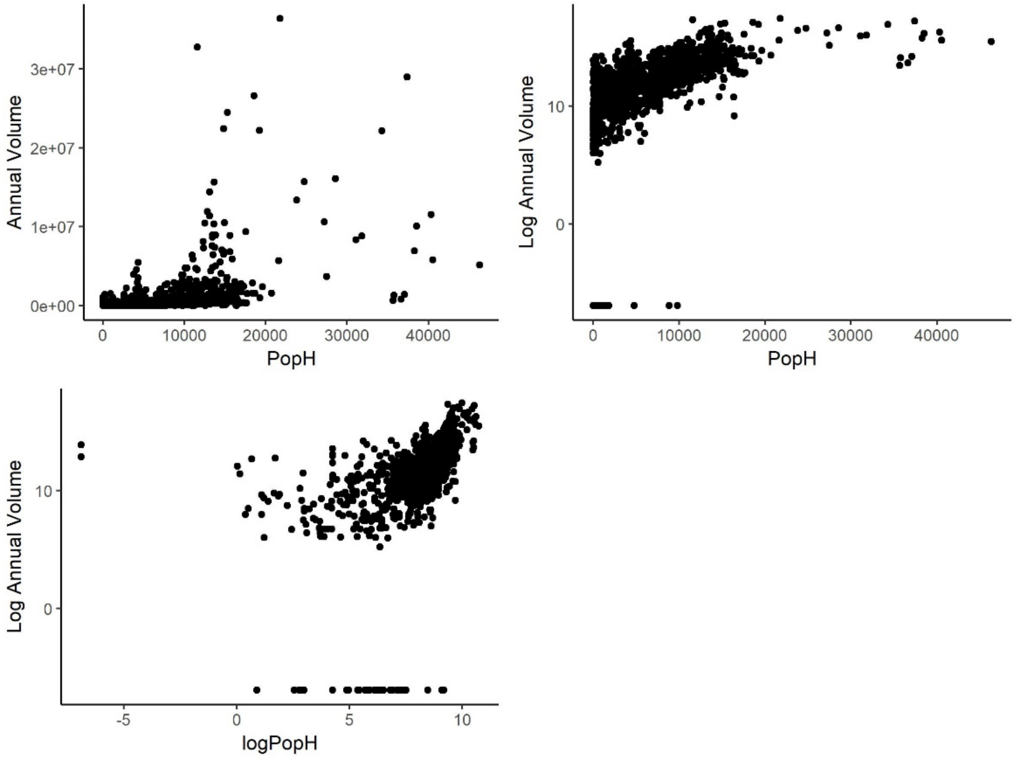


Figure 0-2. Scatter plots for population within a half mile

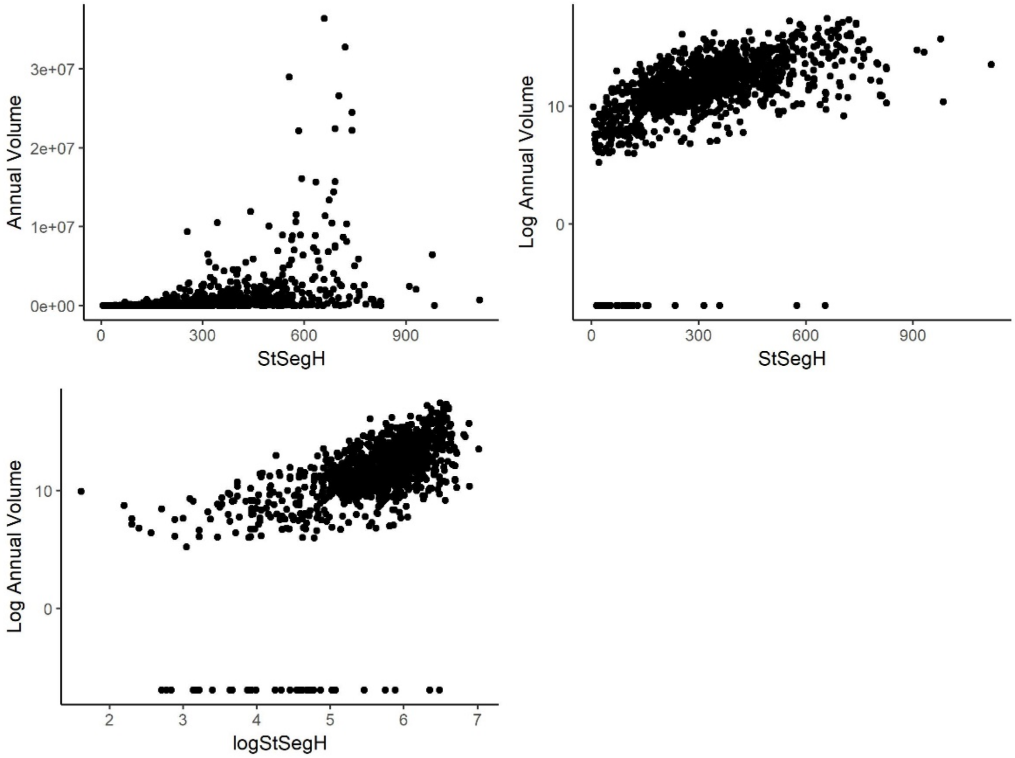


Figure 0-3. Scatter plots for number of street segments within a half mile

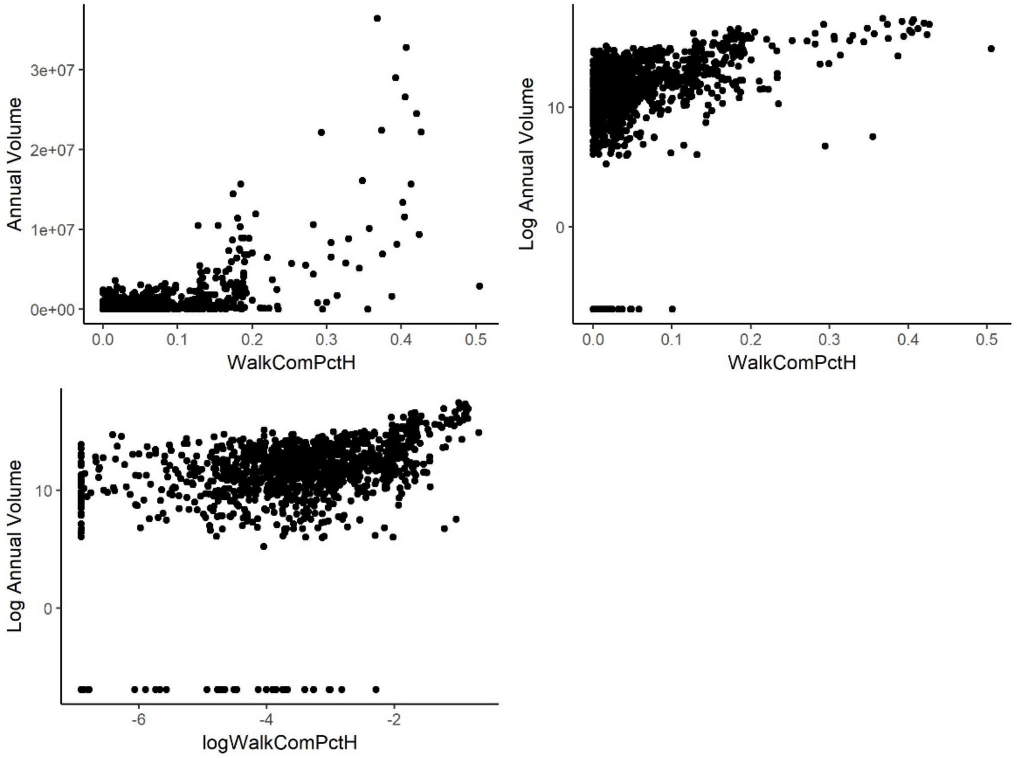


Figure 0-4. Scatter plots for walk commute mode share within a half mile

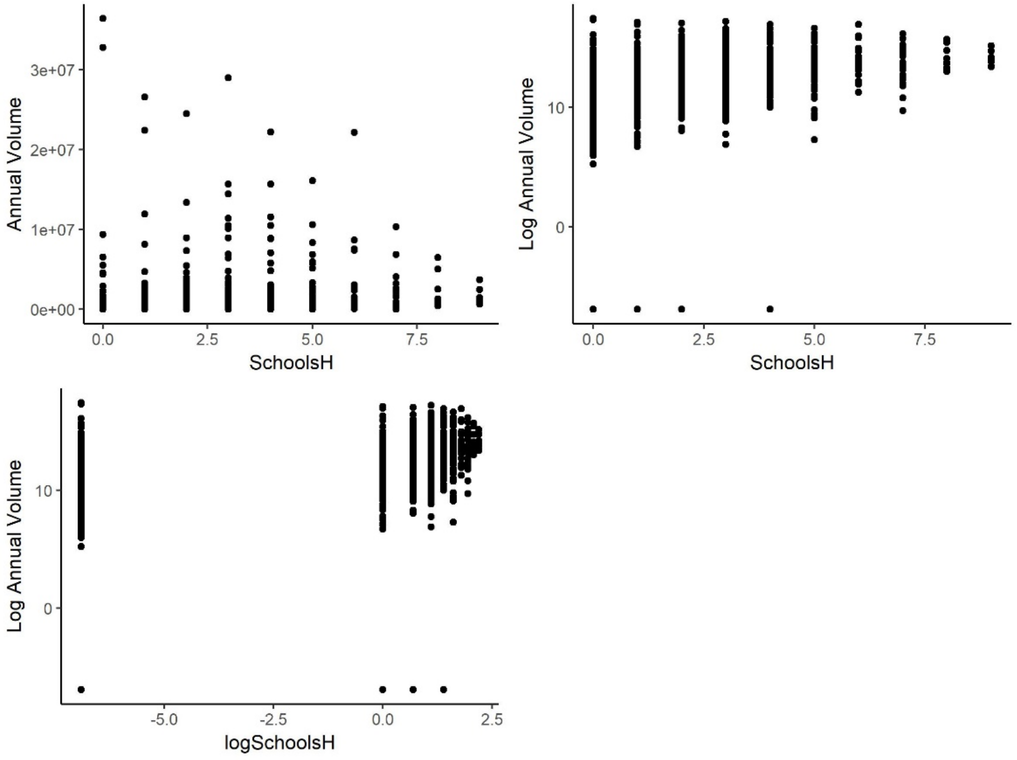


Figure 0-5. Scatter plots for number of schools within a half mile