RESEARCH

UDOT

Keeping Utah Moving

Utah Department of Transportation - Research Division
4501 South 2700 West - P.O. Box 148410 - SLC,UT 84114-8410

Report No. UT-20.17

UTILIZING ARCHIVED TRAFFIC SIGNAL PERFORMANCE MEASURES FOR PEDESTRIAN PLANNING AND ANALYSIS

Prepared For:

Utah Department of Transportation
Research & Innovation Division

Final Report
August 2020

DISCLAIMER

The authors alone are responsible for the preparation and accuracy of the information, data, analysis, discussions, recommendations, and conclusions presented herein. The contents do not necessarily reflect the views, opinions, endorsements, or policies of the Utah Department of Transportation or the U.S. Department of Transportation. The Utah Department of Transportation makes no representation or warranty of any kind, and assumes no liability therefore.

ACKNOWLEDGMENTS

The authors acknowledge the Utah Department of Transportation (UDOT) for funding this research, and the following individuals on the Technical Advisory Committee for helping to guide the research:

- Ryan Bailey, UDOT
- Dan Bergenthal, Salt Lake City
- Travis Evans, UDOT
- Heidi Goedhart, UDOT
- Jamie Mackey, UDOT
- Kevin Nichol, UDOT
- Angelo Papastamos, UDOT
- Matt Seipold, UDOT
- Mark Taylor, UDOT
- Stephanie Tomlin, UDOT

The authors also thank the following students from Utah State University for their assistance with data collection: Skyler Allred, Colby Bench, Allie Boyer, Sadie Boyer, Kevin Brown, Maren Chadwick, Jordan Duncan, Emily Fica, Matthew Harris, Tyler Kendall, Jacob Leatham, Riley Manwaring, Luke Martineau, Niranjan Poudel, Nichole Rogers, and Michael Ruiz-Leon.

TECHNICAL REPORT ABSTRACT

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
UT-20.17	N/A	N/A

4. Title and Subtitle	5. Report Date
UTILIZING ARCHIVED TRAFFIC SIGNAL PERFORMANCE MEASURES FOR PEDESTRIAN PLANNING AND ANALYSIS	August 2020
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Patrick A. Singleton, Ferdousy Runa, Prasanna Humagain	

9. Performing Organization Name and Address	10. Work Unit No.
Utah State University Department of Civil and Environmental Engineering 4110 Old Main Hill Logan, UT 84322-4110	5H08462H
	11. Contract or Grant No. 19-8237

12. Sponsoring Agency Name and Address	13. Type of Report & Period Covered
Utah Department of Transportation 4501 South 2700 West P.O. Box 148410 Salt Lake City, UT 84114-8410	Final Report Sep 2018 to Jul 2020
	14. Sponsoring Agency Code PIC No. UT18.602

15. Supplementary Notes

Prepared in cooperation with the Utah Department of Transportation and the U.S. Department of Transportation, Federal Highway Administration

16. Abstract

The overall goal of this project was to explore the use of continuous pedestrian traffic signal data from the Automated Traffic Signal Performance Measures (ATSPM) system to estimate pedestrian volumes at signalized intersections. Our objectives were to identify patterns of pedestrian activity at signalized intersections, develop methods to estimate pedestrian crossing volumes from signal data, and create a prototype visualization. Using one year of data from 1,522 Utah traffic signals, we applied time series clustering across two new "pedestrian activity metrics" and identified seven distinct patterns of hourly and weekly pedestrian activity. These seven typologies varied by the magnitude (high, medium, and low) and the number (one or two) of weekday peak hours. Based on these typologies, we randomly selected 90 Utah signals, used UDOT traffic cameras to record over 10,000 hours of video, and manually counted almost 175,000 pedestrians crossing at the intersections. Using processed hourly pedestrian actuations and detections from ATSPM data, we estimated five non-linear regression models (segmented by pedestrian activity, cycle length, and pedestrian recall) using pedestrian signal data to predict hourly pedestrian crossing volumes. Overall, our estimates were strongly correlated with observed volumes (0.84) and had a low error (+/- 3.0 on average). These results—which use orders of magnitude more data than had previously been assembled on this topic—demonstrate the validity of using pedestrian data from traffic signals to estimate levels of pedestrian activity. We also developed a prototype dashboard to interactively visualize pedestrian signal data.

17. Key Words	18. Distribution Statement	23. Registrant's Seal
Actuated traffic signal controllers, Cluster analysis, Pedestrians, Pedestrian actuated controllers, Pedestrian counts, Pedestrian detectors, Traffic signals, Traffic signal controllers, Walking	Not restricted. Available through: UDOT Research Division 4501 South 2700 West P.O. Box 148410 Salt Lake City, UT 84114-8410 www.udot.utah.gov/go/research	N/A

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	87	N/A

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

UNIT CONVERSION FACTORS

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. (Adapted from FHWA report template, Revised March 2003)

LIST OF ACRONYMS

AADP	Annual Average Daily Pedestrians
ATSPM	Automated Traffic Signal Performance Measures
CORT	Temporal Correlation
COVID-19	Novel Coronavirus Pandemic
CSV	Comma-Separated Values
DOT	Department of Transportation
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
FHWA	Federal Highway Administration
EC	Empirical Clustering
EU	Euclidian Distance
HAWK	High-Intensity Activated Crosswalk Beacon
LU	Land Use
MAE	Mean Absolute Error
PAM	Pedestrian Activity Metric
PNG	Portable Network Graphics
RMSE	Root Mean Square Error
TOC	Traffic Operations Center
UDOT	Utah Department of Transportation
USU	Utah State University

EXECUTIVE SUMMARY

Data on pedestrian activity is useful for multimodal transportation planning, traffic safety analyses, and health impact assessments, but traditional data collection methods are not able to efficiently capture data on walking continuously for many locations. One promising and ubiquitous data source is information on pedestrian detections and actuations recorded in high-resolution traffic-signal-controller data logs. A couple of studies have investigated these datasets as proxies for pedestrian activity, but they have only done this for single locations. Every time a pedestrian push button is pressed in the state of Utah, this activity is recorded, and UDOT archives these traffic signal pedestrian actuation data for use in its Automated Traffic Signal Performance Measures (ATSPM) system. The overall goal of this research project was to explore the use of traffic signal data to develop estimates of pedestrian activity at signalized intersections. To achieve this goal, this project had three objectives: to identify patterns of pedestrian activity, develop methods to estimate pedestrian crossing volumes from signal data, and create a prototype visualization.

First, we obtained one year (July 2017 through June 2018) of data from 1,522 Utah traffic signals and calculated six "pedestrian activity metrics" (PAMs) for each hour of an average week. Then, we applied time series clustering to two different PAMs, and through cross-classification identified seven distinct patterns of hourly and weekly pedestrian activity. These seven typologies varied by the magnitude of their peak hours (high, medium, and low) as well as the shape and number (one or two) of weekday peak hours. The realism of the typologies was validated against built-environment and locational characteristics.

Second, we used these typologies to randomly select 90 Utah signals for in-depth study and data collection. Specifically, we utilized UDOT traffic cameras to record at least 24 hours of video of each crossing at every studied signal. Then, we watched the videos and recorded the timestamp and details of pedestrian crossing events, to allow us to match the observed crossings to traffic signal data from the ATSPM system. In total, we recorded more than 10,000 hours of video of 320 crosswalks, and manually counted almost 175,000 pedestrians crossing at intersections. Next, we estimated five (quadratic or piecewise linear) regression models—segmented by pedestrian activity, cycle length, and pedestrian recall—to predict hourly

pedestrian crossing volumes from processed measures of pedestrian calls and pedestrian detections. Overall, our model results predicted pedestrian crossing volumes remarkably well, with strong correlations (0.84) and small error (+/- 3.0 on average), which is notable given the large sample size (orders of magnitude more data than had previously been assembled on this topic), variety of locations studied, and the few predictors used.

Third, we developed a prototype online dashboard to interactively visualize pedestrian signal data (raw and estimated volumes) in map, figure, and table formats.

Overall, our results demonstrate the validity of using pedestrian data from traffic signals to estimate levels of pedestrian activity. As a result, using traffic signal data and our models, estimates of pedestrian volumes can be generated for various time periods/intervals and in hundreds if not thousands of locations throughout Utah, and also in other states that use a similar ATSPM system. This novel source of pedestrian big data can now be used for a variety of important transportation activities, including as measures of exposure in pedestrian safety analyses, to help prioritize pedestrian infrastructure investments, and to relate walking levels with weather, air quality, and the built environment, among other tasks. We also offer recommendations for future research and implementation.

1.0 INTRODUCTION

1.1 Problem Statement

Multimodal transportation planning, traffic safety analyses, and health impact assessments require information on how many people are walking in various locations throughout the day. However, traditional data collection methods for levels of pedestrian activity are insufficient for these purposes. Manual intersection or street segment counts are time consuming and often infeasible to conduct over long periods of time. Instruments such as infrared counters can record continuous data on trail users, but they are costly to deploy across multiple locations.

One potential data source that is relatively ubiquitous in both time and space (available 24/7 at many intersections) is the high-resolution data logs from traffic signal controllers. Every time a pedestrian push button is pressed in the state of Utah, this activity is recorded, and UDOT archives these traffic signal pedestrian actuation data for use in its Automated Traffic Signal Performance Measures (ATSPM) system. The use of pedestrian signal data is a potentially rich source of information about levels of pedestrian activity.

Nevertheless, obstacles must be overcome before pedestrian actuations can be successfully used as proxies for intersection pedestrian volumes. These include: analysis of pedestrian actuation patterns for different types of signalized intersections, validation of pedestrian actuation data against observed pedestrian counts, and development of conversion factors to translate actuations into levels of pedestrian activity (volumes). This research project tackles these obstacles to develop methods that can transform pedestrian traffic signal data into valuable information on walking activity levels, which will be useful for pedestrian planning as well as health and safety analyses.

1.2 Objectives

This project explores the use of continuous pedestrian actuation data from the Automated Traffic Signal Performance Measures (ATSPM) system to develop estimates of pedestrian activity. Towards this overall aim, this project has three specific objectives:

1. Identify patterns of pedestrian activity at traffic signals.
2. Develop methods to estimate pedestrian volumes from signal data.
3. Create a prototype to visualize pedestrian signal activity.

1.3 Scope

This project accomplishes these objectives through the following major tasks:

1. Assemble pedestrian data from traffic signals throughout Utah.
2. Analyze assembled data and identify traffic signal typologies based on pedestrian activity patterns and traffic signal settings.
3. Collect multi-day data on observed pedestrian counts at a sample of intersections using UDOT's overhead video cameras.
4. Compare counts to actuations, and use regression models to develop factoring methods that estimate pedestrian intersection crossing volumes.
5. Create a prototype online tool and graphical interface that visualizes estimated levels of pedestrian activity.

1.4 Report Outline

Section 1.0 introduces the project and its motivations, objectives, and major tasks. Section 2.0 provides background material on the research topic and methods, including existing work using traffic signal data for pedestrian analysis, cluster analysis methods, and information about linear regression. Section 3.0 describes the data collection process: downloading traffic-signal-controller log data, creating pedestrian activity metrics, processing those data, recording video data, counting pedestrian crossings from the videos, and assembling all of the data together. Section 4.0 reports the data evaluation/analysis, including the cluster analysis results, the development of factoring methods to estimate pedestrian crossing volumes from pedestrian signal data, and the creation of a prototype visualization. Section 5.0 summarizes the report by highlighting the major findings, noting limitations, and outlining potential steps for future work. Section 6.0 provides recommendations for implementation of the research findings.

2.0 RESEARCH METHODS

2.1 Overview

This chapter first presents a brief literature review on the use of traffic signal data for pedestrian activity analysis. A couple of studies have tested using pedestrian traffic signal actuations as a proxy for pedestrian volumes, with success but only for individual sites. In order for these rich data to be used for this purpose, we must replicate these studies in a larger number and variety of locations. Before doing so, it is prudent to classify the large number of signals into smaller groups or typologies, based on pedestrian activity patterns, to ensure our findings are widely generalizable. Thus, this chapter describes the details of one machine learning method used for determining similar groups of elements: time series clustering. Finally, we summarize the methods of linear regression that are used to find the best factors to convert pedestrian signal data to estimated pedestrian volumes.

2.2 Traffic Signals and Pedestrians

Pedestrian data is essential for many transportation engineering, planning, and research efforts. Pedestrian volume is a necessary measure of exposure in pedestrian safety analyses and could be a useful measure of physical activity in health impact assessments. Quantifying walking levels would be useful for prioritizing investments, designing infrastructure, and completing other important multimodal transportation planning and operational activities. All of these tasks require information on how many people are walking in various locations throughout the day.

Unfortunately, traditional data collection methods for levels of walking activity are insufficient for efficiently collecting pedestrian data in many locations over long time periods. Manual counts can be done in-person in real time, or after the fact if video cameras have been used. Although accurate and feasible across many sites, manual counts are time consuming and costly (mostly due to many person-hours of effort), so they are most appropriate for relatively short durations (FHWA, 2016). Alternatively, various technology-based automated counting methods exist, including passive infrared, active infrared, radar, seismic, and pressure sensors. Automated methods are best over long durations and for identifying systematic variations and

the impacts of weather and special events, but they often involve larger up-front costs (for equipment, etc.) thus limiting the number of locations where they can be deployed; they also benefit from periodic validation using manual count methods (FHWA, 2016; Ryus et al., 2014). Newer and emerging methods—such as video image processing and the use of crowdsourced data—are promising (StreetLight, 2019), but they still have limitations: video-based methods require a network of cameras and extensive software processing; and mobile app-based data come from a (potentially unrepresentative) segment of the population and require expansion factors. In summary, existing ways of monitoring pedestrian travel can usually collect data for either (but not both of) many locations or long time periods.

One potential data source that is relatively ubiquitous in both time and space (available 24/7 at many intersections) is the high-definition data logs from traffic signal controllers. Many (but not all) traffic signals require people walking who want to cross an approach to press a pedestrian push button to request the walk indication. The push-button is thus an active sensor that (barring any equipment malfunction) confirms that a person was indeed present at that location at a specific time. Although it is certainly not perfect—one person may press the button multiple times, or a group of people may only press the button once—pedestrian data from traffic signal controller logs could be a useful and ubiquitous automated source of pedestrian data that already exists at minimal additional cost.

Until recently, this rich set of signal event data was not being systematically logged. Smaglik et al. (2007) developed a general method and module for automatically logging time-stamped event data from traffic signal controllers. These high-resolution data loggers record many types of traffic signal events, including active phase changes, phase control and overlap events, and vehicle and pedestrian detection events. Each record includes a timestamp, an event code, and an event parameter representing a phase or overlap number, detector channel, or other information (Sturdevant et al., 2012). Several pedestrian-relevant events are commonly logged:

- *Event code 0, Phase On*: This event occurs with the activation of the phase on, such as the start of green or the start of the walk interval.
- *Event code 21, Pedestrian Begin Walk*: This event occurs with the activation of the walk indication for a particular phase.

- *Event code 22, Pedestrian Begin Clearance*: This event occurs with the activation of the flashing don't walk indication for a particular phase.
- *Event code 23, Pedestrian Begin Solid Don't Walk*: This event occurs when the don't walk indication becomes solid, with the termination of the pedestrian clearance interval.
- *Event code 45, Pedestrian Call Registered*: This event occurs when a call to service for a particular phase is registered from pedestrian demand. Note that this event may not occur if pedestrian recall is set for the phase.
- *Event codes 89 and 90, PedDetector Off and PedDetector On*: These events occur when the signal from the pedestrian push-button is deactivated or activated, after any delay or extension is processed, for a particular pedestrian detector channel. Multiple pedestrian detection events may occur for a single pedestrian call registered.

Traffic signal data—especially records of pedestrian detections and pedestrian phase calls—may provide valuable information about pedestrian activity levels over time at a location, as long as the signalized intersection has phases with walk indications and crossings with pedestrian detectors (usually pedestrian push-buttons). Nevertheless, obstacles must be overcome before pedestrian signal data can be successfully used as a proxy for intersection pedestrian volumes. These efforts include: validation of pedestrian actuation data against observed pedestrian counts, and development of conversion factors to translate actuations into levels of pedestrian activity (volumes).

2.2.1 Previous Work

To our knowledge, only three studies have investigated the use of pedestrian data from traffic signal controller logs to estimate walking activity at signalized intersections.

Day et al. (2011) analyzed data on traffic-signal pedestrian phase actuations per hour at one signalized intersection in Indiana over 18 months. They identified patterns of pedestrian signal activity as a function of time of day, day of week, weather, other seasonal effects, special events, and a change in the pedestrian phase configuration. The authors also demonstrated that it was feasible to record pedestrian actuations over a long period of time with minimal additional cost, but they did not compare actuations with observed pedestrian counts.

Blanc et al. (2015) conducted a 24-hour pilot study of pedestrian activity at one signalized intersection in Oregon that had actuated pedestrian crossings (using push-button detection) on all four crossings. The authors used video data to manually count 596 pedestrians, which they compared to 482 pedestrian phases from the traffic signal controller logs. They developed adjustment factors for each phase and for the intersection overall. They also compared pedestrian counts to pedestrian actuations for each crosswalk and found correlations of 0.83 or greater, demonstrating the potential for traffic-signal pedestrian data to adequately approximate pedestrian crossing volumes at a signalized intersection. Finally, the authors demonstrated the potential to apply their adjustment factors to pedestrian phase data and calculate estimates of daily and annual average daily pedestrian counts.

Kothuri et al. (2017) returned to the same Oregon intersection two years later to replicate the previous study's findings. During daylight hours over nearly three days, the authors used video data to manually count 818 pedestrians, and signal controller log data to record 723 pedestrian phases. Adjustment factors (pedestrians per phase) were roughly the same magnitude as before (0.9 to 1.2), and correlations were nearly as good in most cases (around 0.80, although one crossing was about 0.67).

Overall, these studies demonstrate that it appears to be possible to use pedestrian signal data in order to estimate pedestrian crossing volumes. However, research is very limited and has studied only one or two intersections and a few dozen hours at a time. There is a need for much larger-scale research to see if these relationships hold (and if the validity of pedestrian signal data is maintained) in many different kinds of locations and under different conditions (including pedestrian recall, when the walk indication comes on without having to press the push-button). Thus, the ultimate purpose of this study is to examine and validate the use of pedestrian data from traffic-signal controller logs against observed pedestrian counts, in order to develop methods that use pedestrian signal data to estimate pedestrian crossing volumes.

2.3 Time Series Clustering

Machine learning and pattern recognition techniques have been used in multidisciplinary research in areas such as finance, geography, and electronics (Aghabozorgi et al., 2015), and

their application is increasing in the field of transportation and traffic engineering. Depending on the types of input data, one of two data mining techniques are typically used: supervised or unsupervised learning (Sathya & Abraham, 2013). Supervised learning requires training data based on a standard and known output, whereas unsupervised learning deals with unlabeled data or raw data. Unsupervised learning is more applicable to exploratory situations in which the specific numbers and types of desired outcomes are unknown, such as the classification of traffic signals according to their patterns of pedestrian activity.

Clustering—a form of unsupervised learning—is a technique used to place similar objects into the same group without prior knowledge of a group's definitions. These groups or clusters are formed by maximizing the similarity between objects in a particular group and minimizing the similarity with objects from other groups (Aghabozorgi et al., 2015). Clustering for temporally static, cross-sectional big data is straightforward to implement, but it is more complex for time series data. Since the input data set for this study involve time series of pedestrian events at traffic signals, the more specific method of time series clustering is required.

The basic theory behind time series clustering is to convert time series data in the form of static data so that clustering algorithms developed for static data can be easily applied (Liao, 2005). There are three common approaches to time series clustering: shape-based, feature-based, and model-based approach (Aghabozorgi et al., 2015). The shape-based approach compares the profile or shape of the time series by analyzing the peaks and trends of time series (Montero & Vilar, 2014). In feature-based approaches, some characteristics of the raw time series data are extracted, and clustering algorithms are applied to the features. For model-based methods, the original time series data is converted into suitable model parameters and then clustering algorithms are applied to those parameters (Aghabozorgi et al., 2015).

Time series clustering has been widely used in the field of transportation. Côme and Oukhellou (2014) analyzed the usage (arrival/departure counts) of a bicycle sharing system in Paris and grouped the stations from high-volume to low-volume clusters by using clustering algorithms. Artificial neural networks in combination with clustering have been used to classify intersections based on vehicle trajectories or accident severity incidents (Akoz & Karsligil, 2014). Similarly, clustering methods have been applied in traffic-flow prediction models (Smith

& Demetsky, 1994), traffic signal planning (Wang et al., 2005), and improving efficiency of traffic signals (Datesh et al., 2011).

Critical steps in the cluster analysis process—selecting a (dis)similarity measure, choosing a clustering algorithm, and determining an optimal number of clusters—are discussed below. A step-wise overview of the entire clustering methodology adopted can be shown in Figure 2.1 below the following subsections.

2.3.1 (Dis)similarity Measures

In cluster analysis, (dis)similarity measures can be classified into two broad categories: shape based and structure based (Montero & Vilar, 2014). Shape-based (dis)similarity measures compare time series based on the direct proximity between their values, allowing us to compare the absolute magnitude of pedestrian activity across intersections. Two shape-based (dis)similarity measures are Euclidian distance (EU) and dynamic time warping (DTW). Given two time series F_i and F_j of length T, the following distance equations apply:

$$d_{EU}(F_i, F_j) = \left(\sum_{t=1}^{T} \left(F_{it} - F_{jt}\right)^2 \right)^2 \qquad (Eq.\ 2.1)$$

$$d_{DTW}(F_i, F_j) = \left(\sum_{t=1}^{T} \left| F_{it} - F_{jt} \right| \right) \qquad (Eq.\ 2.2)$$

Structure-based (dis)similarity measures compare time series based on the simultaneity of their (increasing/decreasing) patterns, allowing us to compare the relative trajectories of pedestrian activity across intersections. Two structure-based (dis)similarity measures are discrete wavelet transform (DWT) and temporal correlation (CORT) (Chouakria & Nagabhusan, 2007). DWT transforms time series into their wavelet approximations and then finds the dissimilarity between those wavelet approximations. CORT, which measures the proximity of temporal variation between time series, can be calculated using the following distance equation:

$$d_{CORT}(F_i, F_j) = \frac{\sum_{t=1}^{T-1} \left(F_{i(t+1)} - F_{it}\right)\left(F_{j(t+1)} - F_{jt}\right)}{\sqrt{\sum_{t=1}^{T-1} \left(F_{i(t+1)} - F_{it}\right)^2} \sqrt{\sum_{t=1}^{T-1} \left(F_{j(t+1)} - F_{jt}\right)^2}} \qquad (Eq.\ 2.3)$$

Optimal (dis)similarity measures are mostly data- and use-specific. Hence, we experimented with different functions and selected ones based on visualization and statistical goodness-of-fit (see section 4.2 for more details).

2.3.2 Clustering Algorithms

Two common algorithms to perform time series clustering are: k-means and hierarchical clustering (Liao, 2005; Kassambra, 2017). The algorithms use dissimilarity measures to group observations into clusters, in which observations are similar to others within the same cluster and different from observations in other clusters. K-means iteratively assigns observations to groups, calculates the group center point, and then reassigns observations to whichever group has the closest center. Hierarchical clustering starts with each observation in its own group, then iteratively combines pairs of similar groups until all groups have been combined. The k-means algorithm is found to be computationally efficient but tends to produce poor results for datasets whose shape of clusters are non-spherical (not bounded by closed shape) and data points are closer to each other, when compared to hierarchical clustering. Therefore, k-means is suitable to use with shape-based (dis)similarity measures, whereas hierarchical clustering is better for structure-based measures. Another difference is that k-means requires a given (k) number of clusters, whereas hierarchical clustering produces results for any number of clusters. In either case, one must specify or select the number of clusters for the final result.

2.3.3 Determining the Optimal Number of Clusters

The final step in cluster analysis is to determine the optimal number of clusters (between one and the number of observations) that adequately represent patterns within a dataset. In k-means clustering, increasing the number of clusters decreases the total within sum of squares (the objective function), but with diminishing returns. Common tools to assist in the selection of the number of clusters include an elbow curve (Kodinariya & Makwana, 2013) and a silhouette curve (Kassambra, 2017). For hierarchical algorithm, one can visualize the dendogram (tree structure) to determine a suitable cut-off point. The gap statistic curve (Tibshirani et al., 2001) can also be used for both algorithms.

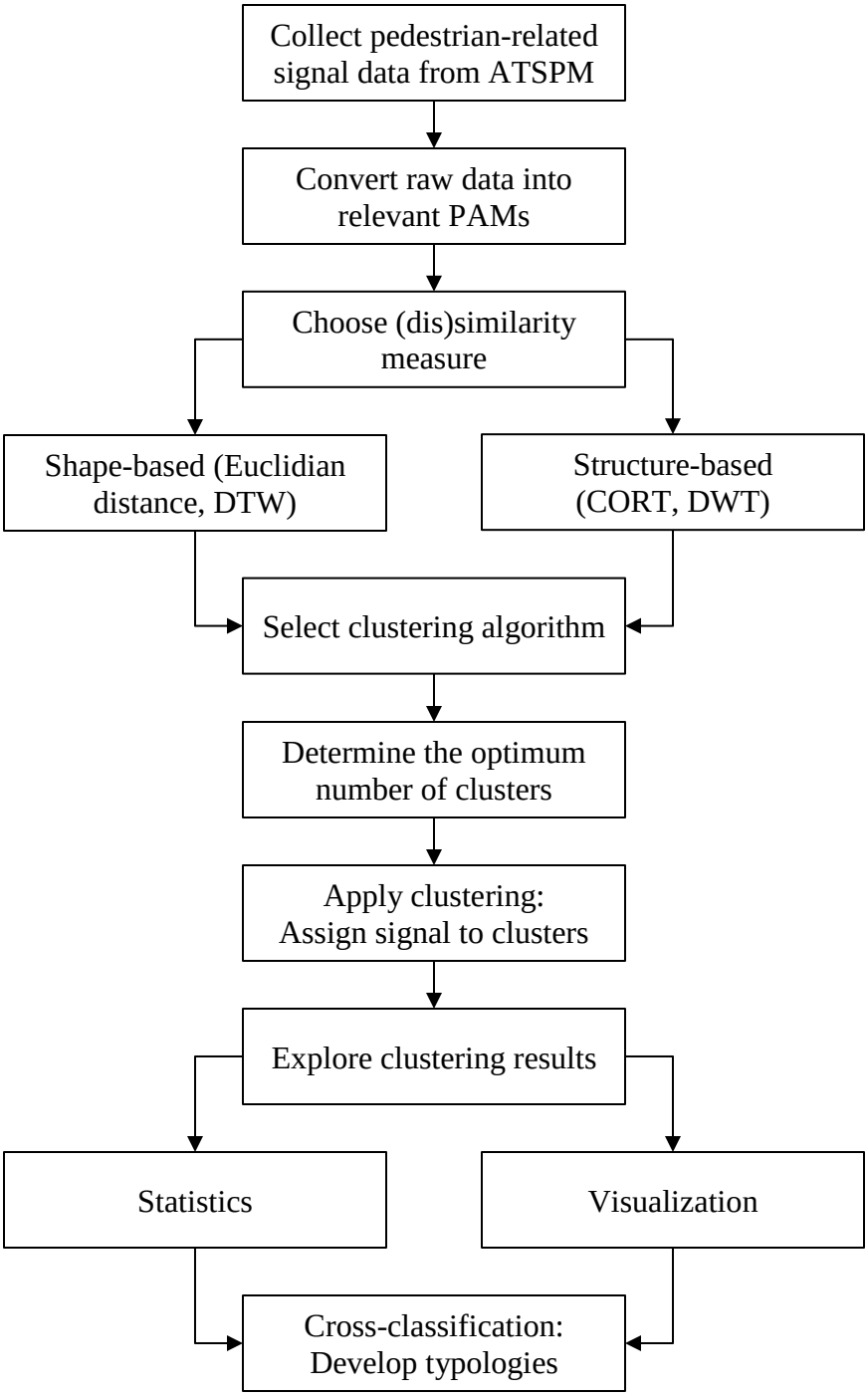


Figure 2.1 Overview of cluster analysis methodology

2.4 Linear Regression

Linear regression is a statistical data analysis technique that is used to determine the relationship between one particular variable of interest (Y) and one or many other variables (X). The X variables are called the independent, explanatory, or predictor variables. The dependent or outcome variable is denoted as Y. There are two types of linear regression: simple and multiple.

In simple linear regression (see Eq. 2.4), a single independent variable (X) is used to predict the value of a dependent variable (Y). In multiple linear regression, two or more independent variables (Xs) are used to predict the value of a dependent variable (Y) (see Eq. 2.5). The only difference between simple and multiple linear regression is the number of independent variables. Independent variables can be transformations of themselves, such as a quadratic or parabolic function (see Eq. 2.6).

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (Eq. 2.4)$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \varepsilon_i \qquad (Eq. 2.5)$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 {X_i}^2 + \varepsilon_i \qquad (Eq. 2.6)$$

where, for each observation *i*:

- Y is the dependent variable
- X is an explanatory variable
- β_0 is the intercept (the value of y when x = 0)
- $\beta_{1.....j}$ is the slope or parameter estimates for each X variable
- ε is the residuals (the deviations from the fitted line to the observed values)

There are several ways to measure the fit of the model to the original data:

- The coefficient of determination (R^2) can measure how close the data are to the fitted regression line. R^2 is the proportion of total variance explained or accounted for by the model (both the independent variables X, and the parameters β). A value of R^2 close to 1 means that most of the variance is explained by the model.

- Root Mean Square Error (RMSE) is another measure of model fit. Lower values of RMSE mean that the differences between the observed values and the model's predicted values are small, indicating a better fit. RMSE is a good measure of how accurately the model predicts the response, and it is a particularly important criterion for fit if the main purpose of the model is the prediction.
- Mean Absolute Error (MAE) is the simplest measure of model fit. The MAE is the average of the absolute values of the differences between the forecasted values and the actual values. MAE tells us how big of an error (positive or negative) we can expect from the forecast on average.
- A good-fit model equally distributes residuals around zero (when looking at a plot of residuals vs. observed values). There are no systematic over- or under-predictions.

2.5 Summary

In this chapter, we first summarized the results of a couple of studies that have tested using pedestrian traffic signal actuations as a proxy for pedestrian volumes, with success but only in single sites. We utilize a similar approach as was used in these three studies: collect automated traffic-signal pedestrian data, compare to manual counts (from video), and develop factoring methods to estimate pedestrian volumes. However, we do this for multiple intersections and multiple days, thus yielding more robust and generalizable results. The methods of linear regression we use to develop the factoring methods were also summarized in this chapter. The results of the factoring methods/models are presented later in Section 4.3.

In order to ensure our results are applicable to a variety of locations and situations, we must first classify traffic signals into different typologies based on pedestrian activity patterns. In this chapter, we also detailed the machine learning technique (time series cluster analysis) that we use for determining the groups of signals from which we must sample. The results of the cluster analysis are presented later in Section 4.2.

3.0 DATA COLLECTION

3.1 Overview

This chapter summarizes the steps involved with data collection, assembly, and processing. First, we introduce the raw pedestrian data we can obtain from the traffic-signal controller logs through the ATSPM system. Next, we describe the data assembly process to support time series clustering: calculating different pedestrian activity metrics, processing ATSPM data to generate the cluster analysis inputs, and assembling geographic data used to characterize the clusters. Finally, we detail the steps involved with collecting and processing video and ATSPM data to support the analysis to develop pedestrian volume estimation methods: sampling intersections for video data collection, collecting observed pedestrian data from video recordings, assembling ATSPM data, and combining the two data sources.

3.2 Traffic Signal Pedestrian Data

3.2.1 High-Resolution Traffic-Signal Controller Logs

Traffic signal controllers manage the safe operation of signalized intersections and their signal control infrastructure, such as vehicle and pedestrian indications/displays. This role includes interpreting and responding to external information about user demand through vehicle and pedestrian detectors (Urbanik et al., 2015). As a result, controllers deal with up to hundreds of events per minute, from phase changes to detector events. Such second-by-second event information—which can be as fine-grained as specific cycles and individual approaches—is useful for traffic-signal-operations management and for calculating signal performance measures. As previously mentioned (see section 2.2), these high-resolution traffic-signal controller data can now be systematically logged (Smaglik et al., 2007), including timestamped information about the specific event and associated phase or detector (Sturdevant et al., 2012). These logs may include valuable information about pedestrian user demand, if the signalized intersection has walk indications and pedestrian detectors (usually pedestrian push-buttons). In fact, such information has been used in a handful of studies to measure pedestrian activity levels (Blanc et al., 2015; Day et al., 2011; Kothuri et al., 2017). Earlier (in section 2.2), we noted that

several pedestrian-relevant events are commonly logged, including: pedestrian detection events or push-button presses (events 90 and 89), pedestrian calls registered (event 45), phase on (event 0), and the start of the walk, flashing don't walk, and solid don't-walk intervals (events 21, 22, and 23).

Table 3.1 shows a typical example of a how a traffic-signal controller log represents an instance of pedestrian user demand. This is for Signal ID 5306, located at the intersection of Main St. (US-89/US-91) and 400 N (US-89) in Logan, UT. Phase 8 (recorded as the event parameter) is associated with the northern crosswalk across Main St., and the pedestrian walk indication is not on recall. Approximately 32 seconds after noon on January 1, 2019, a person walking arrived at the intersection and pressed the pedestrian push-button twice in quick succession. The controller received this information through the detector card and noted the pedestrian detector on (90) at 32.6 seconds and off (89) at 32.9 seconds. Since this was the first pedestrian detection event for phase 8 during this cycle, the controller also registered a pedestrian call (45) at 32.6 seconds. The controller also noted the pedestrian detector on at 33.1 seconds and off at 33.5 seconds, but no pedestrian call needed to be registered for this second detection event. At 55.0 seconds, phase 8 was served (0) and the walk indication turned on (21). Five seconds later, the flashing don't-walk indication (22) started. At 1 minute, 22.0 seconds the solid don't-walk indication (23) appeared, signaling the end of the walk phase.

Table 3.1 Example Traffic-Signal Controller Log

Signal Id	Timestamp	Event Code	Event Parameter
5306	01/01/2019 12:00:32.600	90	8
5306	01/01/2019 12:00:32.600	45	8
5306	01/01/2019 12:00:32.900	89	8
5306	01/01/2019 12:00:33.100	90	8
5306	01/01/2019 12:00:33.500	89	8
5306	01/01/2019 12:00:55.000	0	8
5306	01/01/2019 12:00:55.000	21	8
5306	01/01/2019 12:01:00.000	22	8
5306	01/01/2019 12:01:22.000	23	8

3.2.2 Automated Traffic Signal Performance Measures (ATSPM) System

To harness the power of high-resolution traffic-signal controller log data for signal systems operation and management, the Automated Traffic Signal Performance Measures

(ATSPM) system has been developed to convert raw data into useful performance measures (ATKINS, 2016; Day et al., 2014; Day et al., 2016). UDOT is a national leader in developing and deploying the ATSPM system for real-time management and archived performance assessment of traffic signals throughout the state. The data are archived and the performance measures are made available to the public (https://udottraffic.utah.gov/ATSPM/). This is made possible in part by a centralized system involving partnerships between UDOT and local agencies and an extensive network of connectivity to most signals in Utah. As of fall 2018, UDOT had connected 97% of the 1,237 state-owned signals and 82% of the 874 city/county-owned signals to the system (Taylor & Mackey, 2018).

Currently, there is one pedestrian-related performance measure calculated by the ATSPM system: pedestrian delay. This metric measures the time difference between a pedestrian call registered (45) and the subsequent walk indication (21) as an approximation of the crossing delay experienced by a waiting pedestrian. See Figure 3.1 for an example pedestrian delay graphic for the same intersection, phase, and day as the example of Table 3.1.

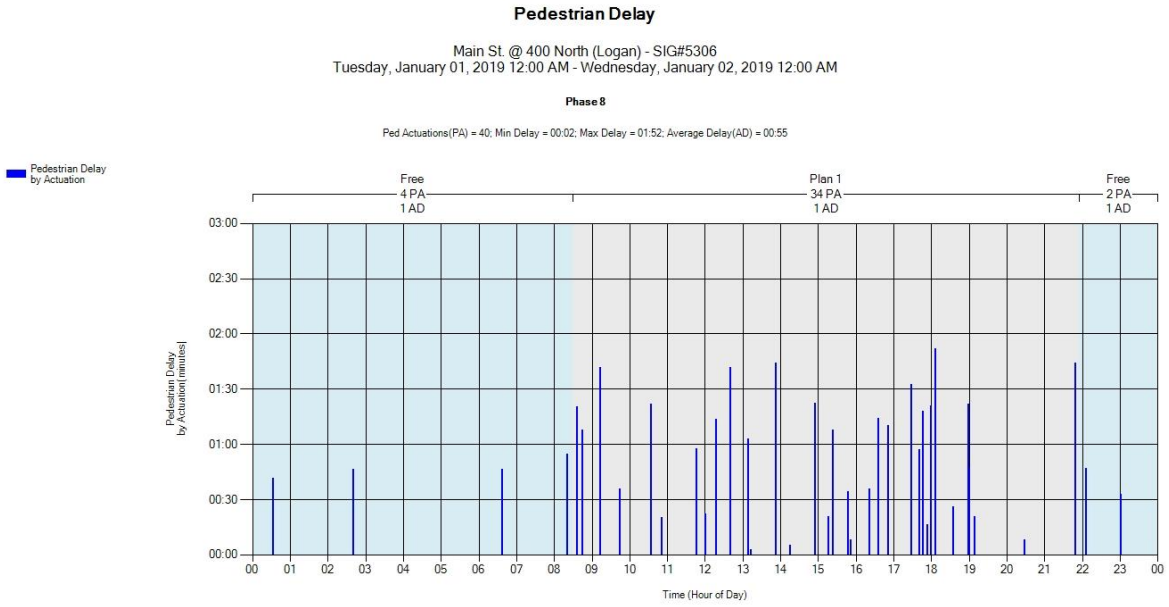


Figure 3.1 Example pedestrian delay performance measure

UDOT recently developed a way to download the raw traffic-signal controller log data on the ATSPM website, which can be accessed by select personnel with log-in credentials. Currently, downloads are limited to a single signal and just over one-million records at a time, but the logs can be filtered by event code, event parameter, and timestamp.

3.3 Data for Typologies and Clustering

The time-series clustering methods used to create typologies of pedestrian activity patterns require data over a long time period (e.g., a year) for all signals in the study area. In the following subsections, we define the different pedestrian activity metrics we considered, describe the data processing steps, and note the locational data we assembled to conceptually validate the typologies generated.

3.3.1 Pedestrian Activity Metrics

Before we can do the cluster analysis, we must select a metric to use when calculating the time series. With input from the technical advisory committee, we developed several different metrics for measuring intersection-level "pedestrian activity" from traffic-signal controller log data. Specifically, we created six different pedestrian activity metrics (PAMs) that could be useful for measuring patterns in intersection pedestrian activity using signal data. These PAMs are highlighted in Table 3.2.

Table 3.2 Pedestrian Activity Metrics (PAMs) Considered

PAM	Equation	Definition
1	$\#90_t$	# pedestrian detections per unit time t
2	$\#45_t$	# pedestrian calls registered per unit time t
3	$\#90_t \div \#0_t$	# pedestrian detections per phase
4	$\#45_t \div \#0_t$	# pedestrian calls registered per phase
5	$100\% \times \#90_t \div \sum_1^T \#90_t$	# pedestrian detections at time t, as a percentage of the weekly total
6	$100\% \times \#45_t \div \sum_1^T \#45_t$	# pedestrian calls detected at time t, as a percentage of the weekly total

As shown in Table 3.3, each metric can be calculated for a particular intersection (overall for all phases, or for each phase separately) and for a single time unit within a given time period,

and it may be averaged over any larger time period. For this study, we calculated all metrics for each intersection (overall) and for each hour of the week, averaged over one year. The objective behind averaging across the whole year was to nullify some of the effects of temporal variation caused by special events, festivals, and other unusual activities.

Table 3.3 Information about Traffic-Signal Pedestrian Activity Metrics

Parameter	Options	Selected
Intersection scope	All phases, single phase	All phases
Time unit	Year, quarter/season, month, week, day, hour, 15 minutes, cycle, phase	Hour
Time period	Year, quarter/season, month, week, day, hour, 15 minutes, cycle	Week
Averaged	None, year, quarter/season, month, week, day	Year

The most basic metrics are counts of the number of pedestrian detections (event code 90) and pedestrian calls registered (event code 45) within a unit of time *t*. These simple calculations are easy to perform and interpret, and they are comparable across sites as an initial magnitude of the level of pedestrian activity.

$$PAM_{1t} = \#90_t \qquad (Eq. 3.1)$$

$$PAM_{2t} = \#45_t \qquad (Eq. 3.2)$$

The second set of metrics are the number of pedestrian detections and pedestrian calls registered per phase. By dividing by the number of phase starts (event code 0) within a unit of time *t*, these metrics can account for the fact that signals may have different and varying cycle lengths, thus different numbers of opportunities for pedestrian crossings. By normalizing by the number of phases, these metrics are more comparable across signals with short and long cycle lengths. They also retain the ability to assess the magnitudes of the original metrics, so sites with more pedestrian activity can be differentiated from sites with less pedestrian activity.

$$PAM_{3t} = \#90_t \div \#0_t \qquad (Eq. 3.3)$$

$$PAM_{4t} = \#45_t \div \#0_t \qquad (Eq. 3.4)$$

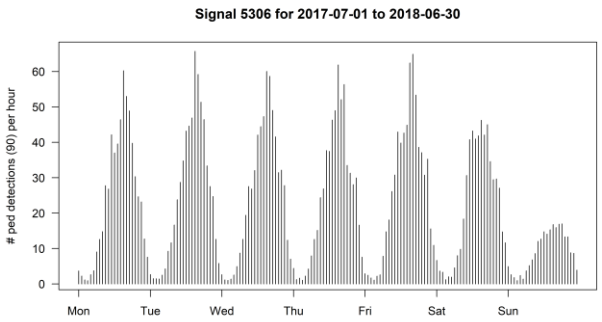
The last set of metrics are the number of pedestrian detections and pedestrian calls registered in a time unit *t* as a proportion of the total number within a longer time period *T*. By normalizing by some longer-time metrics of pedestrian activity, these metrics are also able to somewhat account for differences and variations in cycle lengths or number of cycles. Furthermore, by normalizing by a measure of the magnitude at a site, these metrics result in dimensionless values (expressed as a percentage) that do not depend on the overall pedestrian activity level at a site. Thus, these metrics are good measures of the shape of pedestrian activity patterns within a time period.

$$PAM_{5t} = 100\% \times \#90_t \div \sum_1^T \#90_t \qquad (Eq. 3.5)$$

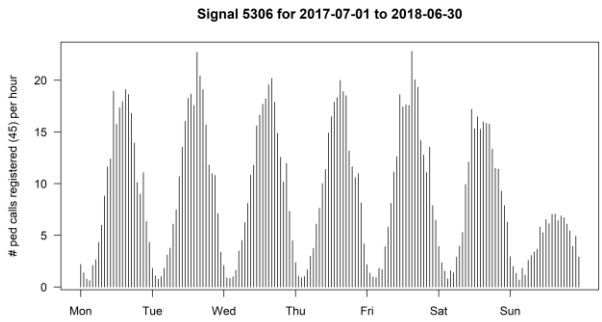
$$PAM_{6t} = 100\% \times \#45_t \div \sum_1^T \#45_t \qquad (Eq. 3.6)$$

Other considerations involving these metrics is the use of pedestrian detections versus pedestrian calls registered. For the first and second sets of metrics, pedestrian detections will be larger than pedestrian calls registered due to the possibility of multiple push-button presses prior to the walk phase being served. For the second set of metrics, a pedestrian call can only be registered once per phase, so the metric using pedestrian calls registered per phase is effectively the proportion of phases in which there were any pedestrian detections. For the third set of metrics, both metrics should be approximately equal, with the pedestrian detection metric perhaps slightly more peaked.

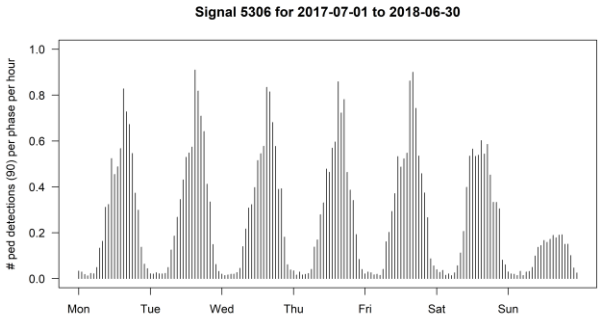
Figure 3.2 displays the annual average hourly and weekly pedestrian activity metrics for an example signal.



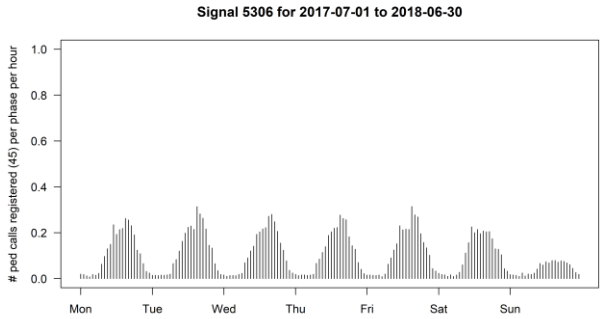
PAM_{1t}: Average pedestrian detections per hour



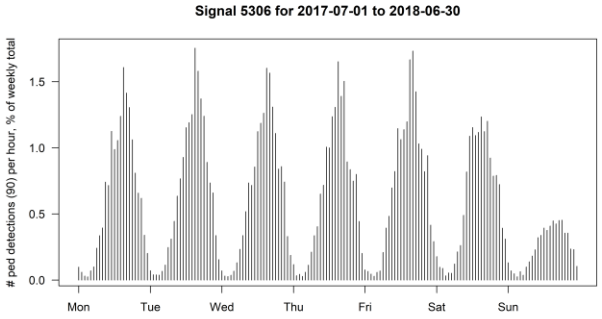
PAM_{2t}: Average pedestrian calls registered per hour



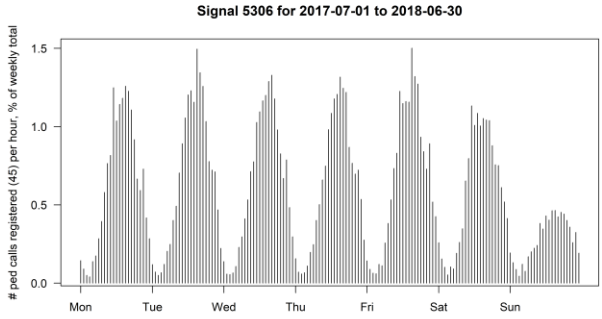
PAM3t: Average pedestrian detections per phase per hour



PAM4t: Average pedestrian calls registered per phase per hour



PAM5t: Average pedestrian detections per hour, percentage of weekly total



PAM6t: Average pedestrian calls registered per hour, percentage of weekly total

Figure 3.2 Example plots of pedestrian activity metrics for Signal 5306

For the purposes of developing typologies of pedestrian activity at traffic signals, we selected as the pedestrian activity metric: the number of pedestrian detections (PAM_{1t}) per hour in a week, averaged over the course of a full year. This metric captures both magnitude and shape. Preliminary analyses found very similar clusters when using pedestrian calls registered (PAM_{2t}) or pedestrian detections per phase (PAM_{3t}) instead.

3.3.2 Data Processing

The processing of the pedestrian-signal pedestrian data proceeded as follows. All data processing was conducted using custom scripts in R.

First, raw controller log data (as CSV files) were downloaded for every available traffic signal from the ATSPM website. Due to the roughly million-record limit on downloads, only pedestrian-relevant event codes (0, 21, 22, 23, 45, 89, and 90) were selected for all phases. Furthermore, only three months (or less) of data were downloaded at a time, as the number of these events was less than 1,000,000 in three months for most signals. Some signals had too many pedestrian events, so the downloads were split into shorter time periods and merged back into three-month files afterwards. Data were downloaded for the following four quarters:

- July 1, 2017 through September 30, 2017
- October 1, 2017 through December 31, 2017
- January 1, 2018 through March 31, 2018
- April 1, 2018 through June 30, 2018

Second, once all traffic signal data were downloaded, the quarterly logs for each signal were merged and sorted by timestamp to create a year-long raw data file spanning July 2017 through June 2018. Third, for each signal, the number of phases (#0s), pedestrian calls registered (#45s), and pedestrian detections (#90s) were tabulated for each hour in the entire year. Fourth, these tabulations were then averaged across the entire year for each of the 168 hours in a typical week. This generated the first two pedestrian activity metrics: the number of pedestrian detections and pedestrian calls registered per hour (PAM_{1t}, PAM_{2t}). Fifth, additional traffic-signal pedestrian activity metrics were calculated. These included the number of pedestrian detections and pedestrian calls registered per hour per phase (PAM_{3t}, PAM_{4t}), as well as the number of pedestrian detections and pedestrian calls registered per hour as a proportion of the total number within a week (PAM_{5t}, PAM_{6t}). Sixth, tables and figures summarizing the pedestrian activity metrics for each hour in an average week mid-2017 through mid-2018 were saved as CSV and PNG files.

Table 3.4 and Figure 3.2 show examples of the final processed data tables and figures for all six pedestrian activity metrics at an example signal.

Table 3.4 Example Processed Data Table

Signal	Start	End	#0s	PAM_1 (#45s)	PAM_2 (#90s)	PAM_3 (#45/0)	PAM_4 (#90/0)	PAM_5 (%45s)	PAM_6 (%90s)
5306	Mon 6AM	Mon 7AM	185.17	4.33	9.08	0.02	0.05	0.28	0.24
5306	Mon 7AM	Mon 8AM	94.29	6.00	12.60	0.06	0.13	0.40	0.34
5306	Mon 8AM	Mon 9AM	90.60	8.81	14.81	0.10	0.16	0.58	0.40
5306	Mon 9AM	Mon 10AM	89.04	11.62	27.77	0.13	0.31	0.76	0.74
5306	Mon 10AM	Mon 11AM	82.81	12.40	26.83	0.15	0.32	0.82	0.72
5306	Mon 11AM	Mon 12PM	80.52	18.96	42.17	0.24	0.52	1.25	1.13
5306	Mon 12PM	Mon 1PM	81.40	15.75	37.00	0.19	0.45	1.04	0.99
5306	Mon 1PM	Mon 2PM	81.08	17.35	39.56	0.21	0.49	1.14	1.06
5306	Mon 2PM	Mon 3PM	81.79	17.94	46.40	0.22	0.57	1.18	1.24

3.3.3 Traffic Signal Geographic Data

To characterize the different types of pedestrian activity patterns resulting from the cluster analysis (see section 4.2), built environment data were collected at the Census block group level from the Smart Location Database (Ramsey & Bell, 2014). The specific measures used were household density (number of households per area), population density (number of people per area), and non-automobile employment accessibility (number of jobs within 30 minutes by walk and transit). These built environment measures were calculated as the area-weighted average of all Census block groups located within a 0.5-mile circular buffer from each signalized intersection.

3.4 Data for Factoring Methods and Regression Models

3.4.1 Study Locations

The next step in this research project involved recording videos of traffic at signalized intersections throughout Utah and counting the number of pedestrian crossings (Task 3). These observed pedestrian counts were then compared to different traffic-signal pedestrian activity metrics to develop factors and factoring methods for estimating pedestrian volumes from traffic signal data (Task 4). The purpose of developing clusters of signals (typologies of traffic-signal

pedestrian activity patterns) [Task 2] was to ensure that we collect video data from a comprehensive sample of traffic signals so that our results are generalizable to different situations and conditions. In order to improve upon previous research and make our findings more generalizable and transferrable, our goal was to study a wide variety of locations that had different urban forms (urban to rural), were in different regions, and saw different levels of pedestrian activity (high to low). In this subsection, we describe the stratified random sampling of locations to study.

Our sampling procedure began by filtering the list of signalized intersections to just those with a UDOT or local traffic camera that we could use to view (and record) live traffic camera feeds from around the state. We identified 521 signals with a traffic camera, but only 430 of those signals with cameras were assigned to a typology (see section 4.2). Given the available time and budget, we decided to sample up to 100 (23%) of these signals with cameras.

Our stratified random sampling targets had two goals in mind. First, we wanted to include a sufficient number of signals from each of the typologies. Due to the small number of signals with cameras in some typologies, we first included a minimum of 6 signals (or all if less than 6) in each typology. This assigned 39 signals; for the remaining 61, we allocated them roughly proportionally to the number of signals in each typology. We also reserved 4 slots for HAWK signals (pedestrian hybrid beacons) which had not been assigned a typology. The total number of selected study intersections in each typology is shown in Table 3.5. Second, we wanted to include signals from throughout the state of Utah in different contexts and on different types of roadways. Therefore, we segmented our possible list of signals with cameras by UDOT Region and signal owner (state vs. local). Subject to the limits of what was available (for example, only four of Salt Lake City's signals had traffic cameras), we attempted to select a number of signals in each region by owner based on the proportion of total signals (not the number with cameras). The total number of selected study intersections by UDOT Region is shown in Table 3.6.

Although we planned to study up to 100 signals, we ended up only collecting data at 90 signals, mostly due to connectivity issues with the video cameras. The location of the 90 studied signalized intersections with cameras for video data collection is shown in Figure 3.3 and listed in Appendix A.

Table 3.5 Sampling Targets by Typology

Typology		# signals	# with cameras	# selected	# studied
I	High, single peak	33	7	6	3
II	Medium, single peak	164	49	15	14
III	Medium, double peak (a)	12	3	3	3
IV	Medium, double peak (b)	12	6	6	5
V	Low, single peak	864	260	40	41
VI	Low, double peak (a)	252	55	13	9
VII	Low, double peak (b)	185	50	13	11
	Total	1,522	430	96	86

Table 3.6 Sampling Targets by UDOT Region

Region / Owner		Signal IDs	# signals	# with cameras	# selected	# studied
1	UDOT	5000s–5400s	343	66	16	19
	Local	5500s–5900s	138	23	6	3
2	UDOT	7000s–7900s	589	153	24	24
	Salt Lake County	4000s–4900s	352	24	14	12
	Salt Lake City	1000s–1900s	282	10	7	7
3	UDOT	6000s–6400s	317	136	14	12
	Local	6500s–6900s	149	42	6	0
4	UDOT	8000s–8500s	129	35	8	8
	Local	8600s–8900s	73	32	5	5
		Total	2,372	521	100	90

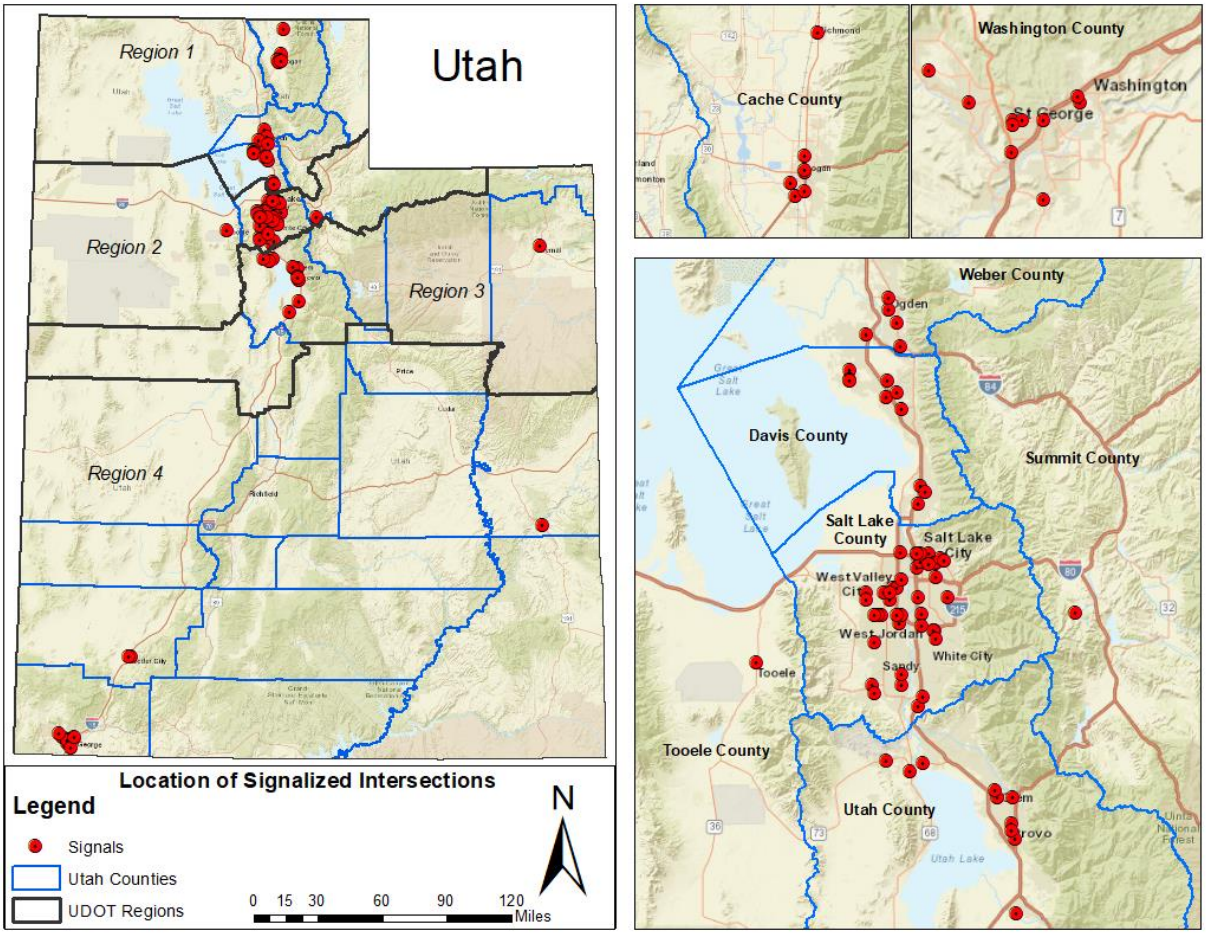


Figure 3.3 Data collection sites at Utah signalized intersections

3.4.2 Video Data Collection

3.4.2.1 Recording Videos

UDOT's overhead traffic cameras were used to record video data of various intersections throughout Utah at different times of day and for various seasons during 2019. We recorded almost 10,900 hours of multiday videos, including views of 320 crosswalks at our 90 intersections, resulting in around 24,000 crossing-hours of video. The list of intersections is shown in Appendix A and represents signals in each cluster and from all parts of Utah.

The USU Time Lab has a fiber optic connection to UDOT's Traffic Operation Center and can view (and record) live traffic camera feeds from around the state. We developed procedures

for recording live feeds, adding timestamps, and saving the resulting video file. The procedure of video recording is presented below:

- Contact the UDOT TOC to request any slight adjustment to the camera field of view, or obtain access to camera pan/tilt/zoom and adjust the camera field of view ourselves, and then place a lock on the view for the duration of recording.
- Use VLC Media Player to record and add a timestamp to the live stream of the camera feed, in raw h264 format, for a time period (usually 48–60 hours).
- Convert the video file to MP4 and split into 6-hour sections (~2.5 GB each).
- Save the videos on a password-protected cloud-based storage system.

We recorded most intersections twice, since cameras could usually only cover two crosswalks at a time. We attempted to capture at least (and often more than) 24 hours of videos of each crosswalk, at least once during the year. Some locations were recorded two or more times throughout the year to observe seasonal differences (winter/spring, summer, and fall). We recorded videos from approximately three to five signals each week. Due to bandwidth limits, we could usually record only three or four videos at once. Moreover, we recorded on weekdays and weekends to observe day-of-week differences. Figure 3.4 shows the distribution of the total number of crossing-hours of video recorded by month. Similar distributions for weekdays and hours varied much less: between 3,313 and 3,670 crossing-hours per weekday, and between 932 and 1,042 crossing-hours per hour.

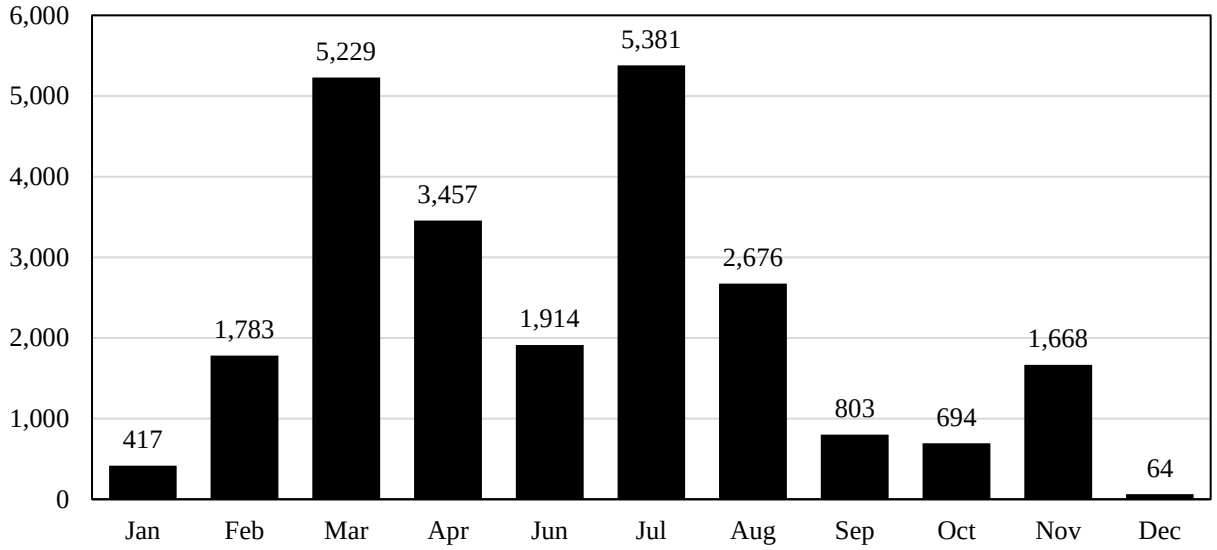


Figure 3.4 Total crossing-hours of video recorded by month of 2019

3.4.2.2 Counting Pedestrians

After recording videos, we manually watched the videos and transcribed timestamped pedestrian crossing events in order to obtain pedestrian crossing volumes to compare against pedestrian signal data.

We developed standardized data collection procedures, including training materials and a graphical user interface (built in R using Shiny) for recording pedestrian crossing events and saving data in a standardized format. Basically, every time a person (or a group of people) was observed using the crosswalk, we recorded this as a pedestrian crossing event that included a timestamp and the number of pedestrians using each crosswalk. (We did not record direction of travel. Also, to make it easier to count groups, we only required that pedestrians be somewhere in the crosswalk at the recorded time.) In addition to counting people walking under their own power (e.g., not being carried or pushed in a stroller), we recorded whether crosswalk users were using other modes, including people skateboarding, using scooters, on bicycles, in wheelchairs, and other vehicles (such as golf carts or sidewalk snowplows). We also collected additional information: whether a pedestrian turned the corner without crossing a street, whether a pedestrian had already crossed another leg of the intersection, whether a pedestrian was crossing the street far outside of the crosswalk, and other special circumstances (via notes).

Figure 3.5 shows an example of the information included in the Video tab of the interface. Figure 3.6 shows two example crossing events recorded from UDOT traffic cameras. In the left image, one person is walking in the east crosswalk (the camera faces south). In the right image, there are two people walking in the west crosswalk (the camera faces west). Figure 3.7 shows how this event (in the left video) would be recorded in the Add Event tab of the interface. Note the event time and the one pedestrian listed for Crosswalk 2 (East).

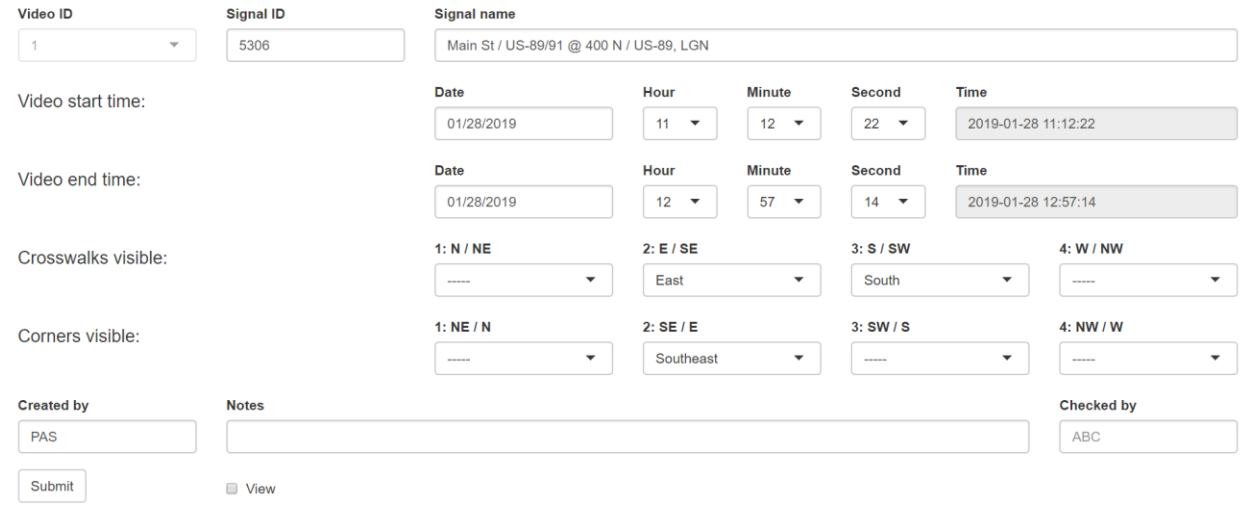


Figure 3.5 Example of the Video tab of the interface

Figure 3.6 Example of pedestrian crossing events recorded on video

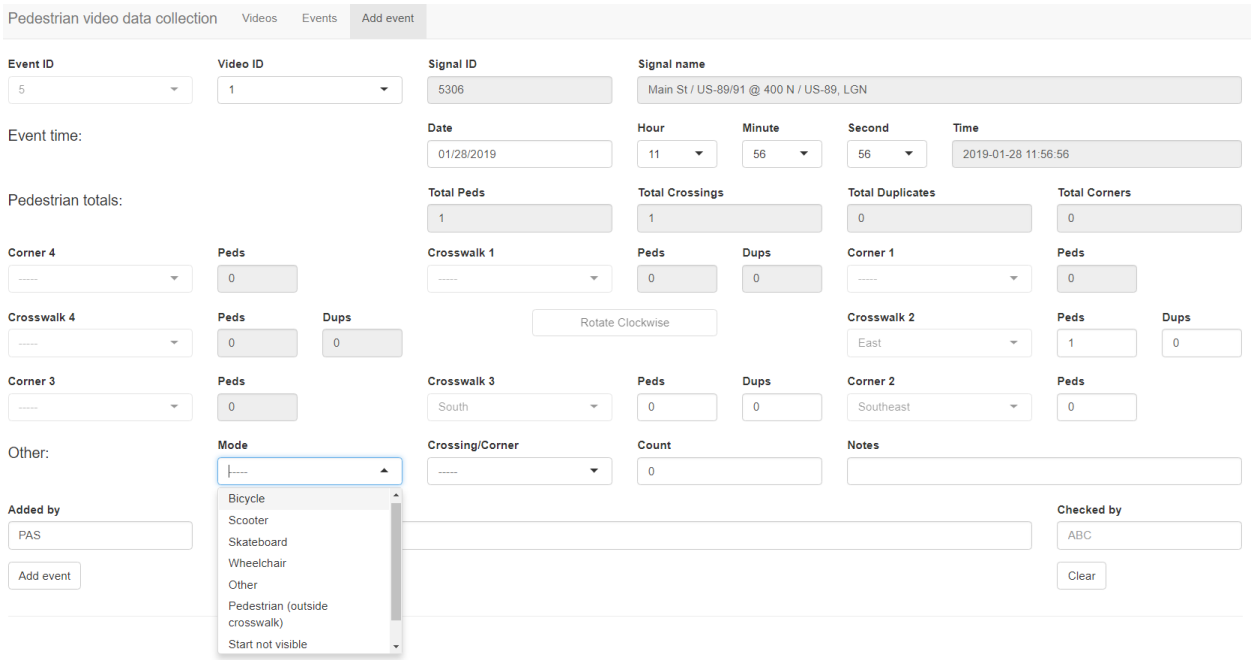


Figure 3.7 Example of the Add Event tab of the interface

The data entry interface resulted in two data files being written for each set of videos: one with information about the video recording, and one with information about the events (which was continuously added to by the Event tab of the interface).

Pedestrian crossing events were recorded by a team of up to 15 trained undergraduate students, who collectively spent around 2,600 hours watching video and counting pedestrians. (Over the 10,900 hours of video, this reflects an ability to watch videos at an average of 4x speed.) As shown in Table 3.7, our team counted around 175,000 people walking, about 12,600 people bicycling, and fewer numbers of other crosswalk users.

Table 3.7 Total Counts of People Walking and Other Activities

Activity	Count (#)
People walking	174,923
People bicycling	12,628
People using (e-)scooters	2,453
People skateboarding	897
People in wheelchairs	537
Other sidewalk users	221
Pedestrians crossing outside of crosswalk	1,151
Pedestrians turning corner	9,350

Before finalizing the pedestrian crossing data collected from the videos, we performed several quality checking procedures (including mass checks for missing or potentially erroneous information, and spot checks against the videos) involving multiple trained personnel. Overall, we corrected about 3% of all event records, mostly for minor and easily fixable issues such as unnecessary notes or timestamp errors. Some data entry errors may remain, but we are confident that we identified and fixed the majority and the most serious issues.

3.4.3 ATSPM Data Assembly

As previously mentioned (see section 3.3.2), we obtained the high-resolution traffic-signal controller log data from UDOT, via a raw-data export-download page on UDOT's ATSPM system website (https://udottraffic.utah.gov/ATSPM/). For each video recording, we downloaded pedestrian-relevant data for that signal for the same time period as the video.

We also performed a timestamp check of the video data against the signal data, to ensure that events lined up. We had added the timestamp to the video feed on the computer that was recording the videos, so there could have been a time lag between when the event occurred (and was recorded by the controller) and when the event appeared in the video (and was recorded using the computer's clock). To check this, we picked a couple of traffic signal events that were visible from the video (e.g., pedestrian begin clearance / flashing walk) for a particular crossing, and recorded the timestamp on the video. We then matched the relevant record from the ATSPM data for that event code (e.g., 22) and phase. (See Table 3.8 for example checks.) In nearly all cases, the time difference was only a couple of seconds, and always less than six seconds. Since this time discrepancy was much smaller than our desired temporal unit of analysis (one hour), we did not make any timestamp adjustments.

Table 3.8 Example Timestamp Checks

Signal ID	ATSPM timestamp	Event code	Event parameter	Video timestamp	Time difference
7184	2/13/2019 18:02:55	22	4	2/13/2019 18:02:57	2 sec
5108	3/17/2019 19:31:07	22	8	3/17/2019 19:31:10	3 sec
1021	3/4/2019 18:06:33	22	6	3/4/2019 18:06:33	0 sec
8634	3/4/2019 18:39:07	22	2	3/4/2019 18:39:08	1 sec

3.4.4 Processing and Merging Data

Once all pedestrian crossing event data from videos and traffic signal data from ATSPM were collected and checked, we proceeded with assembling, processing, and merging the data into one final aggregated dataset for analysis. We selected the hour (e.g., 9:00AM to 10:00AM) as the temporal unit of analysis, and the individual crossing (e.g., Signal 5306, crosswalk associated with phase 8) as the spatial unit of analysis.

For the pedestrian crossing event data, we aggregated data by crossing and hour, summing up (separately) the number of people walking, bicycling, etc. We also calculated what would become our new outcome variable of hourly pedestrian crossing volume, using the sum of the people walking, skateboarding, and in wheelchairs (not including people bicycling or using scooters). Thus, we obtained hourly volumes of pedestrians and other crosswalk users for each signal and crossing.

For the pedestrian traffic signal data, we performed a similar process of aggregating pedestrian-relevant events by crossing (using the associated phase number) and hour. Thus, we obtained hourly counts of pedestrian signal activity—phase on (event 0), pedestrian begin walk (event 21), pedestrian call registered (event 45), and pedestrian detector on (event 90)—for each signal and phase.

In addition, we constructed several new measures of pedestrian signal activity, for a couple of reasons. First, while processing the data, we noticed that some traffic signal controllers did not record a pedestrian call registered (event 45)—such as when the phase was in pedestrian recall (since the walk indication came on automatically)—or recorded multiple pedestrian calls registered if the pedestrian push-button was pressed while the walk indication was on (between events 21 and 22). Second, we wanted to account for the fact that people may press the push-button multiple times in quick succession, which may reduce the ability of pedestrian signal data to predict pedestrian crossing volumes. The new pedestrian-signal activity metrics we constructed (again, aggregating by crossing-hour) were:

- *New 45s*: imputed pedestrian calls registered, with some variations
    - *45A*: In a sequence of events with just {0, 21, 22, 90}, the number of 90 events

immediately preceded by a 0 or 22 event.

- *45B*: In a sequence of events with just {0, 21, 90}, the number of 90 events immediately preceded by a 0 or 21 event.
- *45C*: In a sequence of events with just {0, 90}, the number of 90 events immediately preceded by a 0 event.

- *New 90s*: unique (or time-filtered) pedestrian detections, with some variations
  - *90A*: In a sequence of events with just {90}, the number of events with a time difference ≥ 5 seconds from the previous 90 event.
  - *90B*: In a sequence of events with just {90}, the number of events with a time difference ≥ 10 seconds from the previous 90 event.
  - *90C*: In a sequence of events with just {90}, the number of events with a time difference ≥ 15 seconds from the previous 90 event.

All of these data were then merged together into one long pedestrian video count and signal-activity data file for use in the modeling and factoring processes described in Section 4.3. See Table 3.9 for an example of this final data file.

Table 3.9 Example Final Data Table

SIGNAL	TIME1	TIME2	P	PED	BIKE	SCOOT	SKATE	WHEEL	OTHER	PEDOUT
7086	3/4/2019 13:00	3/4/2019 14:00	2	16	2	0	0	0	0	0
7086	3/4/2019 14:00	3/4/2019 15:00	2	21	3	0	0	0	0	0
5030	7/11/2019 8:00	7/11/2019 9:00	2	10	1	0	0	0	0	0
5030	7/11/2019 9:00	7/11/2019 10:00	2	7	4	0	0	0	0	0
7382	8/4/2019 20:00	8/4/2019 21:00	8	8	2	2	4	0	0	0
7382	8/4/2019 21:00	8/4/2019 22:00	8	14	0	0	0	0	0	0

SIGNAL	TIME1	TIME2	A00	A21	A45	A90	A45A	A45B	A45C	A90A	A90B	A90C
7086	3/4/2019 13:00	3/4/2019 14:00	31	8	7	12	7	8	8	11	11	11
7086	3/4/2019 14:00	3/4/2019 15:00	30	12	12	35	12	12	12	21	18	15
5030	7/11/2019 8:00	7/11/2019 9:00	32	9	9	24	9	9	9	13	12	12
5030	7/11/2019 9:00	7/11/2019 10:00	33	8	8	18	8	8	8	12	12	10
7382	8/4/2019 20:00	8/4/2019 21:00	30	10	14	33	10	12	12	14	12	12
7382	8/4/2019 21:00	8/4/2019 22:00	32	10	10	53	10	10	10	18	13	12

Additionally, we enhanced our aggregated pedestrian signal dataset with other variables reflecting traffic signal operations, also at the crossing and hourly level:

- *Pedestrian recall*: whether or not the phase was likely set to pedestrian recall for some or all of the hour. This was determined using a complex series of conditional statements

(involving relative counts of events 0, 21, 90, and new 45s) which classified 98% of situations, followed by a predictive model to classify the remaining 2%. Records were checked and only 0.3% had to be manually corrected.

- *Cycle length*: approximate average cycle length. This was calculated by dividing the number of minutes (with observed data) in the hour by the number of phase on events (0).

Prior to estimating any models, we slightly filtered our data in order to obtain more reliable results. First, we removed any records with missing pedestrian signal data. These were usually either: (1) "crossings" where a legal crossing (and therefore, pedestrian signals and push-buttons) didn't exist; or (2) hours where ATSPM data was missing entirely. Second, we removed around 1,000 observations with less than 45 minutes of video coverage within the hour. These were usually partial hours when a video started or ended. We worried that including short hours would bias our models towards zero, leading to slightly inaccurate factors and predictions. Third, we removed four hours of outliers on signal data at one signal that we could tell were due to a malfunctioning push-button, and 15 hours of outliers on pedestrian counts at a second signal that occurred due to major concerts and sporting events at an adjacent arena. After this filtering process, 22,630 crossing-hour observations remained.

3.5 Summary

In this chapter, we summarized the steps involved with data collection, assembly, and processing. To prepare data for using cluster analysis to identify typologies of pedestrian activity patterns, we first downloaded raw traffic-signal controller log entries of pedestrian-relevant events (pedestrian indications, calls, and detections) for every signal in Utah using the ATSPM system from July 2017 through June 2018, and then we calculated six pedestrian activity metrics—pedestrian detections and pedestrian calls registered per hour, averaged over the year for each hour in a week, also divided by the number of phases and normalized into a percentage of the weekly total—that we created to quantify pedestrian activity levels from traffic signal data. To collect and assemble data for using regression models to develop factoring methods to estimate pedestrian volumes from signal data, we first selected study locations using a strategic random sampling procedure. Next, we recorded over 10,000 hours of multiple days' video from January 2019 to December 2019 at 90 signalized intersections throughout Utah. Using recorded

videos, we then manually counted pedestrian events and entered this information into a database using a user interface. Finally, we combined our count data with signal data from ATSPM and produced the dataset used for model estimation in the next chapter.

4.0 DATA EVALUATION

4.1 Overview

This chapter reports the detailed results of our analysis. First, we present results from the time-series cluster analysis, including the pedestrian activity metrics and clustering algorithms selected, the cross-classification of signals into typologies, descriptions of pedestrian activity patterns for signals in each typology by shape and magnitude, and characterizations by location. Second, we present results from our linear regression models that developed factoring methods, including our evaluation criteria, results from preliminary testing, final model results (including visual plots), overall results, and an example application of the models to predict pedestrian volumes. Third, we document the prototype visualizations we have developed to display pedestrian signal data and estimated pedestrian volumes at signalized intersections. We end the chapter with a summary of key takeaways.

4.2 Typologies of Pedestrian Activity Patterns (Time Series Clustering)

Time-series cluster analysis (see section 2.3) requires data that are continuous, uniform, and of equal length. First, we assembled one year of ATSPM data and calculated six pedestrian activity metrics for most signals in Utah (as described in section 3.3). Overall, 850 (of 2,372) intersections with no (or incomplete) PAMs were removed prior to clustering, leaving a final analysis dataset of 1,522 signalized intersections. (The number of intersections used was slightly less for PAM_{3t} and slightly higher for PAM_{4t}.) Each time series contained 168 hourly observations, one for each hour/weekday combination (the annual average hourly/weekday value of the PAM).

We performed a total of twelve different time-series clustering procedures on these data, using two different (dis)similarity measures and clustering algorithms on each of the six PAMs. Optimal (dis)similarity measures are mostly data- and use-specific, and clustering algorithms can work better with certain measures. Since PAM_{1t} and PAM_{2t} essentially measure the absolute magnitude of pedestrian activity, shape-based measures (EU and DTW) and a k-means clustering algorithm were used. Alternatively, structure-based measures (CORT and DTW) and hierarchical

clustering algorithms were appropriate for comparing the dimensionless patterns of PAM_{5t} and PAM_{6t}. For PAM_{3t} and PAM_{4t}, which are based on magnitude but normalized somewhat. Both types of measures and algorithms (EU and k-means, CORT and hierarchical) were tested.

When computing the optimum number of clusters for each PAM, the recommended methods (elbow curve, silhouette curve, gap statistic) suggested slightly different numbers of clusters. In order to compare results across PAMs, (dis)similarity measures, and clustering algorithms, we decided to use the same number of clusters in each case: Three clusters provided a reasonable selection across selection criteria. Therefore, every analysis used three clusters for comparative purposes. The TSCLUST package in R (Montero & Vilar, 2014) was used to perform the clustering.

4.2.1 Cluster Analysis Results

Cluster analysis results are summarized in Table 4.1.

Table 4.1 Cluster Analysis Results

			Cluster sizes						
PAM	**(Dis)similarity measure**	**Algorithm**	*1st*	*2nd*	*3rd*	**total within cluster sum of squares**	**average distance within cluster**	**average silhouette width**	**Dunn index**
PAM_{1t}	EU	k-means	1,140	311	71	33,170,045	106.157	0.448	0.424
PAM_{1t}	DTW	k-means	1,128	329	65	1,430,250,297	726.113	0.385	0.338
PAM_{2t}	EU	k-means	1,301	188	33	5,330,000	39.040	0.457	0.240
PAM_{2t}	DTW	k-means	1,303	188	31	306,900,000	285.900	0.434	0.190
PAM_{3t}	EU	k-means	1,350	133	13	6,005	1.175	0.512	0.099
PAM_{3t}	CORT	hierarchical	1,485	9	2	12,458	1.130	0.540	0.078
PAM_{4t}	EU	k-means	1,389	131	15	542	0.330	0.393	0.152
PAM_{4t}	CORT	hierarchical	1,194	328	13	730	0.390	0.551	0.283
PAM_{5t}	CORT	hierarchical	1,093	277	152	22,151	3.973	0.162	0.830
PAM_{5t}	DWT	hierarchical	1,149	316	57	22,269	4.427	0.071	0.684
PAM_{6t}	CORT	hierarchical	1,061	264	197	16,406	3.478	0.242	0.800
PAM_{6t}	DWT	hierarchical	1,027	368	127	17,590	3.820	0.181	0.512

Clustering of PAM_{1t} and PAM_{2t} resulted in three distinct clusters of signalized intersections, easily distinguished by their magnitude. One high activity (but small) group had weekday peak hours of ~140 pedestrian detections (~85 pedestrian calls registered); a medium activity group was more in the range of 75 detections (30 calls registered); and the lowest activity group (but the largest, constituting about 75% of all signals) averaged less than 25 detections (about 10 calls registered) in the highest weekday hour. Results using the measures

EU and DTW were nearly identical—almost 95% of intersections in clusters of PAM_{1t} (EU) were included in same clusters of PAM_{1t} (DTW)—indicating that either measure could be used for classification.

Cluster results for PAM_{5t} and PAM_{6t} were also similar no matter which measure (CORT and DWT) was used: The overlap between clusters of PAMs was in the range of 60–70%. Interestingly, there were fewer visual distinctions between cluster shapes. Differences were mostly apparent in the number and magnitude of the daily peaks: Two groups had two daily peak hours (one with higher peaks and lower troughs), while one group had only an afternoon peak. The one-peak group also had slightly higher relative weekend activity than the two-peak groups.

Conversely, consistent results could not be obtained from the clustering of PAM_{3t} and PAM_{4t}. Almost 95% of intersections were assigned to the lowest magnitude clusters. Observations in the smallest cluster for PAM_{3t} could be considered outliers, as they show (unexpectedly) high average pedestrian activity at night. Similarly, several observations have PAM_{4t} values greater than one (indicating more pedestrian calls than phases), which is also unexpected. (There were also issues with infinite values resulting from dividing by zero.) These irregularities require further investigation and suggest that PAM_{3t} and PAM_{4t} may not be the best metrics for creating pedestrian activity typologies.

Internal validation of clustering results considered two aspects: compactness (the proximity of objects within same cluster) and separation (the distinctiveness of objects in one cluster from those in others) [Kassambara, 2017]. Lower values of compactness measures—total within cluster sum of squares and average distance within cluster—reflect well-formed clusters. Clusters with higher average silhouette width (a separation measure) are more distinctive: One represents perfectly formed clusters, whereas negative values reflect placement of objects in the wrong cluster. Higher values on the Dunn index (a combined measure) implies compact and well-separated clusters.

As shown in Table 4.1, overall, all the clusters performed satisfactorily along compactness and separation lines, although there were differences between (dis)similarity measures. Clusters formed using EU were more compact and better separated than those formed

by DTW (for PAM_{1t} and PAM_{2t}); CORT performed better than DWT for clusters of PAM_{5t} and PAM_{6t}.

4.2.2 Typologies (Pedestrian Activity Patterns)

The aim of our study is to develop typologies (or factor groups) of pedestrian activity patterns at signalized intersections that can differentiate between higher/lower volumes and daily/weekly temporal variations. Ideally, this would include measures of both the absolute magnitude as well as the relative shape of pedestrian activity across hours in a week. Luckily, we had both in the form of PAM_{1t} and PAM_{2t} for magnitude and PAM_{5t} and PAM_{6t} for relative shape. (Recall, also, the limitations with using PAM_{3t} and PAM_{4t}.)

Given the substantial overlap between each set of four cluster results, we selected just one from each set. Given their superior performance, we chose EU for PAM_{1t} and PAM_{2t} and CORT for PAM_{5t} and PAM_{6t}. Based on preliminary results of ongoing research by the authors—suggesting that calls registered (#45s) is a slightly more accurate predictor of pedestrian volumes than detections (#90s), which can overestimate in the face of multiple push-button presses—we selected PAM_{2t} and PAM_{6t}. We acknowledge that the selection of other PAMs or cluster methods to define typologies could have been similarly justified.

We used the results from the time series clustering of PAM_{2t} (EU) and PAM_{6t} (CORT) to cross-classify signals into seven typologies or factor groups of pedestrian activity patterns at signalized intersections. Table 4.2 shows the frequencies of each typology and how they relate to the clustering results. Figure 4.1 plots the average values of the PAMs across all signals in each typology. Note that each figure has one time series line but two axes—the number of pedestrian calls registered per hour (left) and same thing expressed as a percentage of the weekly total (right)—because PAM_{6t} is constructed as a percentage depiction of PAM_{2t}. As discussed in more detail below, each typology can be most clearly characterized by its weekday peak magnitude and the number of daily peaks.

Table 4.2 Typologies Based on Cross-Classification

Magnitude: PAM_{2t} (EU) (# calls registered)	**Relative shape: PAM_{6t} (CORT) (% weekly total)**		
	Single peak	**Double peak (a)**	**Double peak (b)**
High	Type I: 33	—	—
Medium	Type II: 164	Type III: 12	Type IV: 12
Low	Type V: 864	Type VI: 252	Type VII: 185

Overall, all typologies had some similarities in observed average weekly pedestrian activity patterns. Unsurprisingly, pedestrian activity was highest during daytime and evening hours, with most intersections recording little-to-no activity overnight. Peak pedestrian hours were more common in the afternoon and early evening than in the morning. Weekend pedestrian activity (especially on Sundays) was lower than on weekdays, but often without a clear single peak hour. Tuesdays often had the largest peak hour of pedestrian activity, while Mondays and Fridays tended to have slightly lower peaks than other weekdays.

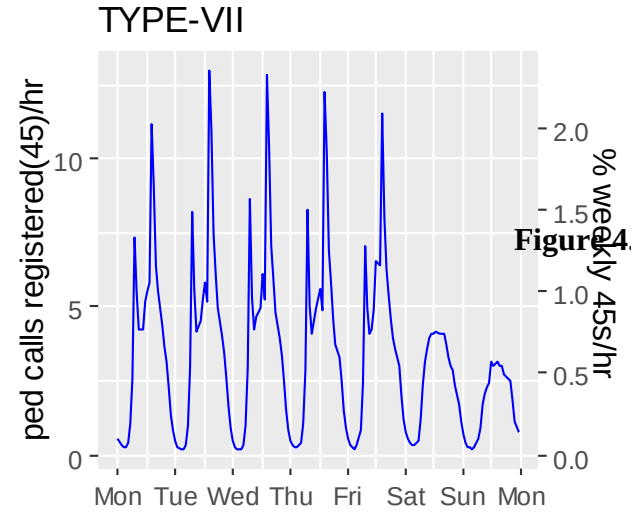


Figure 4.1 Plots of mean values of typologies

Type I: High, Single Peak

Only 33 intersections (2.2%) belonged to this category, which were characterized by consistently high midday pedestrian activity (>75 pedestrian calls registered per hour) and only slightly less weekend activity. Peak weekday hours constituted only about 1% of total weekly activity. These intersections have high pedestrian activity during the day on weekdays and weekends

Type II: Medium, Single Peak

164 intersections (10.8%) fell into this category. These intersections had similar pedestrian activity patterns as Type I, except with slightly lower magnitudes: Peak midday hours saw >30 pedestrian calls registered. Again, the peak hours on weekdays represented around 1% of total weekly pedestrian calls. These intersections have medium levels of daytime pedestrian activity throughout the week.

Type III: Medium, Double Peak (a)

Just 12 intersections (0.8%) belonged to this type, which had clear AM peaks and larger PM peaks (>30 and >50 pedestrian calls registered per hour, respectively), with substantially less activity (~10 pedestrian calls per hour) in between and on weekends. Consistent with this discontinuous pattern, PM peak hours had nearly 3% of weekly activity. These intersections have high weekday pedestrian activity during a couple of AM and PM hours and much lower activity at other times and on weekends.

Type IV: Medium, Double Peak (b)

Another 12 intersections (0.8%) fell into this type, which had two peaks and similar patterns as Type III. Some differences include: lower peak magnitudes (>20 and >40 pedestrian calls registered per hour), lower PM peak proportions (nearly 2% of weekly total), and more midday activity (20–30 pedestrian calls per hour). These intersections have high pedestrian activity during one or two PM hours and medium activity at other times and on weekends.

Type V: Low, Single Peak

These intersections were numerous, representing more than half (864, 56.8%) of all signals. These intersections averaged <10 pedestrian calls registered per hour, with a single PM

peak hour. As with Types I and II, peak hours constituted a little more than 1% of weekly totals. These intersections tend to have low pedestrian activity all the time.

Type VI: Low, Double Peak (a)

These intersections were also common (252, 16.6%), and displayed similar shapes but lower magnitudes than Type III. They had clear AM and larger PM peak hours (>10 and >5 pedestrian calls registered), and PM peaks were typically 2–3% of total weekly activity. These intersections have low-to-medium pedestrian activity during one or two PM hours and much lower activity otherwise.

Type VII: Low, Double Peak (b)

185 intersections (12.2%) belonged to this type, which was quite similar to Type VI. Minor differences included: slightly lower peak hour proportions (~2% of weekly totals) and slightly higher midday pedestrian activity. These intersections also have low pedestrian activity outside one or two PM hours of low-to-medium activity.

4.2.3 Built Environment Characteristics

Pedestrian demand at intersections is influenced by land use characteristics, employment generators, residential density, access to transit, and proximity to schools (Hankey et al., 2012). To further characterize and validate our typologies, we calculated various built environment attributes for the Census block groups within 0.5 miles of each signalized intersection. Results—averaged across all signals in each typology—are shown in Table 4.3.

Table 4.3 Typologies by Built Environment Characteristics

Typology	Population density[a]	Household density[b]	Employment accessibility[c]	# Transit stops[d]	# Schools[d]
Type I	4,684	2,018	221,556	34.5	1.76
Type II	5,975	2,650	328,670	28.8	2.26
Type III	3,342	1,178	37,379	9.2	1.58
Type IV	5,911	2,245	228,773	22.2	2.33
Type V	3,489	1,342	98,200	12.4	1.13
Type VI	3,794	1,399	94,874	11.3	1.56
Type VII	4,216	1,492	91,921	12.1	2.08

Units: [a] households/mi^2; [b] population/mi^2; [c] jobs within 30 minutes, [d] within 0.5 mi

Results were relatively consistent with expectations about relationships between the built environment and pedestrian activity. In general, signals in the high and medium typologies (Types I–IV) were in locations with greater residential density and employment accessibility and more transit stops than signals in the low typologies (Types V–VII). Type III intersections were somewhat anomalous to these trends, but they were a small group (just 12). There were no clear trends with proximity to schools.

When visually inspecting maps of signalized intersections by typology (not shown), some trends emerged. High activity (Type I) intersections were concentrated in downtown Salt Lake City and at large universities (Utah State University, University of Utah, and Brigham Young University). Intersections with medium pedestrian activity (Types II–IV) were mostly found in downtowns, major suburban commercial areas, adjacent to universities, and near transit, parks, and other amenities. The lowest activity (Types V–VII) intersections were dispersed everywhere in Utah.

4.3 Pedestrian Volume Estimation Methods (Regression Modeling)

Since our ultimate goal was to develop an easy way to use pedestrian signal data to estimate pedestrian crossing volumes, we estimated regression models that in their simplest form reflect a factoring approach: Multiply some pedestrian-signal activity metric by a number (or plug it into a simple equation), and get an estimate of pedestrian crossing volume. (We fixed the intercept to be zero.) Overall, this is an adaptation and expansion of correlative methods used in previous research (Blanc et al., 2015; Kothuri et al., 2017).

There were a variety of ways in which to develop and specify our models, using: different independent variables (pedestrian-signal activity measures), different model forms (linear or various non-linear specifications), different phasing (separating out by phase numbers or pedestrian recall), different segmentations (ways to split the dataset into different groups of signals with similar characteristics), and other additional variables to consider. Therefore, we chose both quantitative and qualitative criteria for assessing our models:

- *Root mean square error (RMSE)*: a common measure of the forecasting accuracy of a statistical model. Specifically, this is calculated as the square root of the mean of the

sum of the squared errors/residuals (difference between the actual and predicted values). Lower values (closer to 0) indicate greater accuracy and are desired. Larger errors have greater weight than smaller errors.

- *Mean absolute error (MAE)*: another common measure of forecasting accuracy. Specifically, this is the mean of the absolute value of the errors/residuals. Lower values (closer to 0) indicate greater accuracy and are desired. MAE is easier to interpret than RMSE.
- *Correlation*: a measure of the association between pairs of values. Specifically, this is the Pearson correlation between the observed and predicted values. Larger values (closer to 1) indicate greater similarity and are desired.
- *Plot of observed versus predicted values*: Specifically, this is a plot of actual values (x-axis) versus predicted values (y-axis), with a 1:1 "perfect prediction" line drawn. Points below the line are underestimated and points above the line are overestimated by the model. Ideally, points would fall close to the line with no large groups consistently above or below the line.
- *Plot of residuals versus observed values*: Specifically, this is a plot of the errors/residuals (x-axis) versus observed values (y-axis), with a horizontal "perfect prediction" line through zero drawn. Points below the line are overestimated and points above the line are underestimated by the model. Ideally, points would fall close to the line with no large groups consistently above or below the line.
- *Model parameters*: Specifically, these are the estimated coefficients of the model which would be multiplied by the independent variables to predict values of the dependent variable. Contingent on the model specification, these values should be not too small but also not too large, and should be noticeably different for each segment (otherwise it would be useless to split the data into different groups).
- *Simplicity*: The method should be simple, easy to understand, and easy to apply. There shouldn't be too many calculations, inputs, or segments needed.

4.3.1 Preliminary Testing

In order to quickly narrow in on the best options from among the large number of possible alternative models, we first estimated 150 different models corresponding to the unique combinations (6 x 5 x 5) of the following options:

- *Independent variables*: 45A, 45B, 45C, 90A, 90B, or 90C.
    - Preliminary analysis revealed that all of these yielded better fitting models than the original counts of events 45 and 90.
- *Model forms*: linear, piecewise linear (with one breakpoint), quadratic, exponential, and power.
- *Phasing*: all phases, phases 2/6 only, phases 4/8 only, phases and hours with pedestrian recall, phases and hours without pedestrian recall.
    - At most UDOT signals, phases 2/6 are crossings of the side street (parallel to the main street), while phases 4/8 are crossings of the main street (parallel to the side street). When in coordination during the day, phases 2/6 are often set to pedestrian recall, while phases 4/8 are usually never set to pedestrian recall.

Next, we assessed these 150 models using the three quantitative criteria (RMSE, MAE, and correlation). This effort revealed several suggestions for which options to continue considering and which to discard as inferior prior to the final models:

- *Independent variables*: 45B had the best overall performance (lowest RMSE and MAE, second highest correlation)—although 45A and 45C were within 1–2% of 45B in terms of model accuracy—while 90C was best in some situations (with 90A and 90B also not that much worse). Therefore, we proceeded with testing either 45B or 90C.
- *Model forms*: All of the non-linear specifications (except for exponential, which had generally poor fits) performed better than a simple linear (one multiplicative factor) model. Yet, piecewise linear and quadratic specifications both offered the best improvements over a linear model: 6% lower RMSE, 16% lower MAE, and 5-6% higher correlation. Therefore, we proceeded with testing both piecewise linear and quadratic models.

- *Phasing*: Splitting the data by phases/hours with/without pedestrian recall fit the data better than splitting the data by phase number (2/6 vs. 4/8); this also accommodated signals with non-standard pedestrian phases. Therefore, we proceeded with segmenting the data using pedestrian recall.

Before estimating the final models, we also tested whether segmentations by other variables improved the models:

- *Day of week*: We tested whether there were different relationships on Mondays through Friday than on Saturdays and Sundays. Although there were some statistically significant differences (not surprising, given the large sample sizes), segmenting by day of week did not offer enough meaningful improvement in model fit to warrant the additional complexity.
- *Time of day*: We also tested whether there were different relationships during the day than at night (defined as between 9pm and 6am). Again, the models with statistically significant differences did not substantially improve model fit enough to justify this segmentation.
- *Pedestrian activity*: We examined if relationships were different at signals with different levels of pedestrian activity. First, we conducted a new time-series cluster analysis (using k-means) on patterns of pedestrian signal activity, similar to what we did when selecting locations, but using the new 45B measure. Results (available upon request) suggested splitting signals into two groups: 135 signals with high pedestrian activity, and 1,476 signals with low pedestrian activity. In order to apply this clustering at other locations and in the future (when pedestrian activity levels might change), we deterministically split signals using a breakpoint of greater than 350 annual average daily pedestrian signal activity (45B) to assign signals to the high-activity group. (Only 0.4% of signals were reclassified due to this deterministic grouping.) Then, for the regression models, splitting data into high- vs. low-pedestrian activity locations offered substantial improvements in some cases, so we proceeded with segmenting by pedestrian activity.

To summarize, when developing the final models, we considered: using 45B or 90C for independent variables, having piecewise linear or quadratic model forms, and segmenting by pedestrian recall, pedestrian activity, and cycle length.

4.3.2 Final Model Results

Our testing of various models and specifications yielded five models in which the data were segmented along several different dimensions. Below, each final model is described in its own section, followed by a summary of all model results and an example application.

4.3.2.1 HAWK Signal Crossings

This model applies to all crossings at HAWK signals. There were 234 observations in the dataset. The best model was a quadratic specification using the A90C metric. The correlation was very high at 0.915, and the mean absolute error was 14.6, which is fairly large due to the high volumes observed at HAWK signals. The observations and model-predicted values are shown in Figure 4.2.

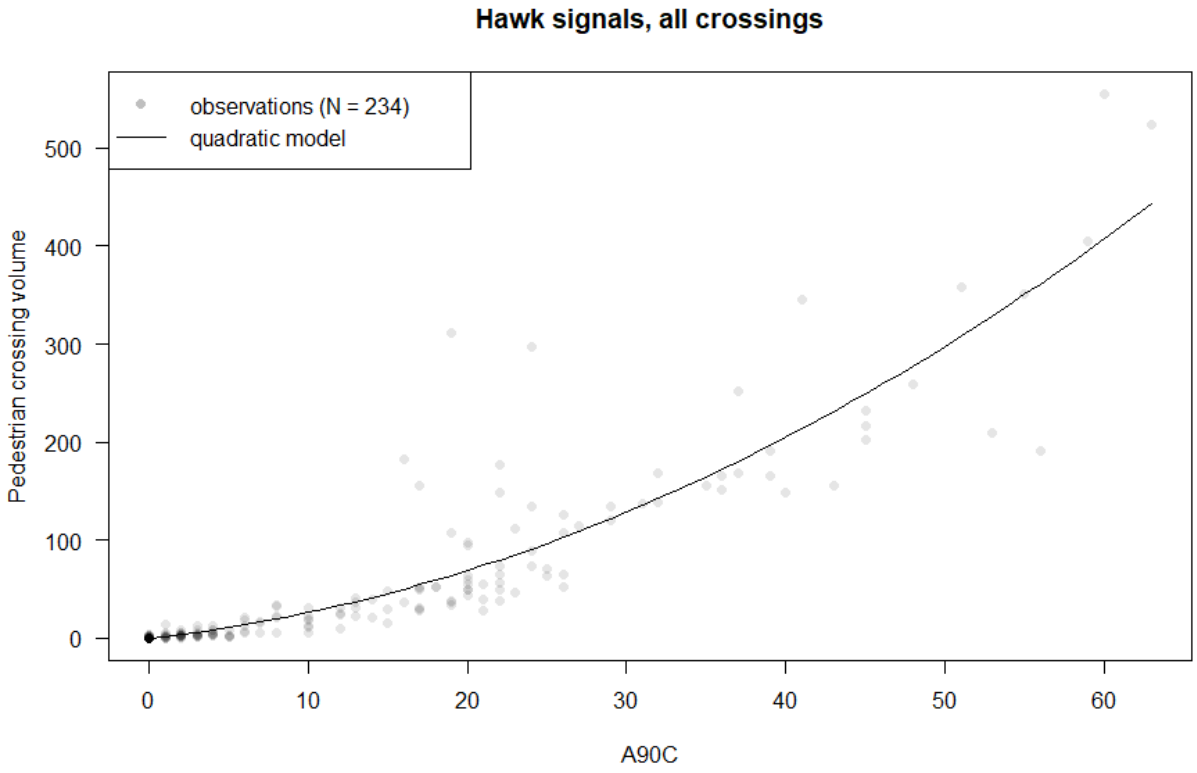


Figure 4.2 Model for all crossing at HAWK signals

4.3.2.2 Crossings with Pedestrian Recall at High-Activity Signals

This model applies to all crossings at high-activity signals when they operate with pedestrian recall. There were 1,440 observations in the dataset. The best model was a quadratic specification using the A45B metric. This model was the worst-performing model, but still had a correlation of 0.649. The somewhat large mean absolute error of 17.8 is not surprising due to the high pedestrian volumes experienced at these signals. The observations and model-predicted values are shown in Figure 4.3.

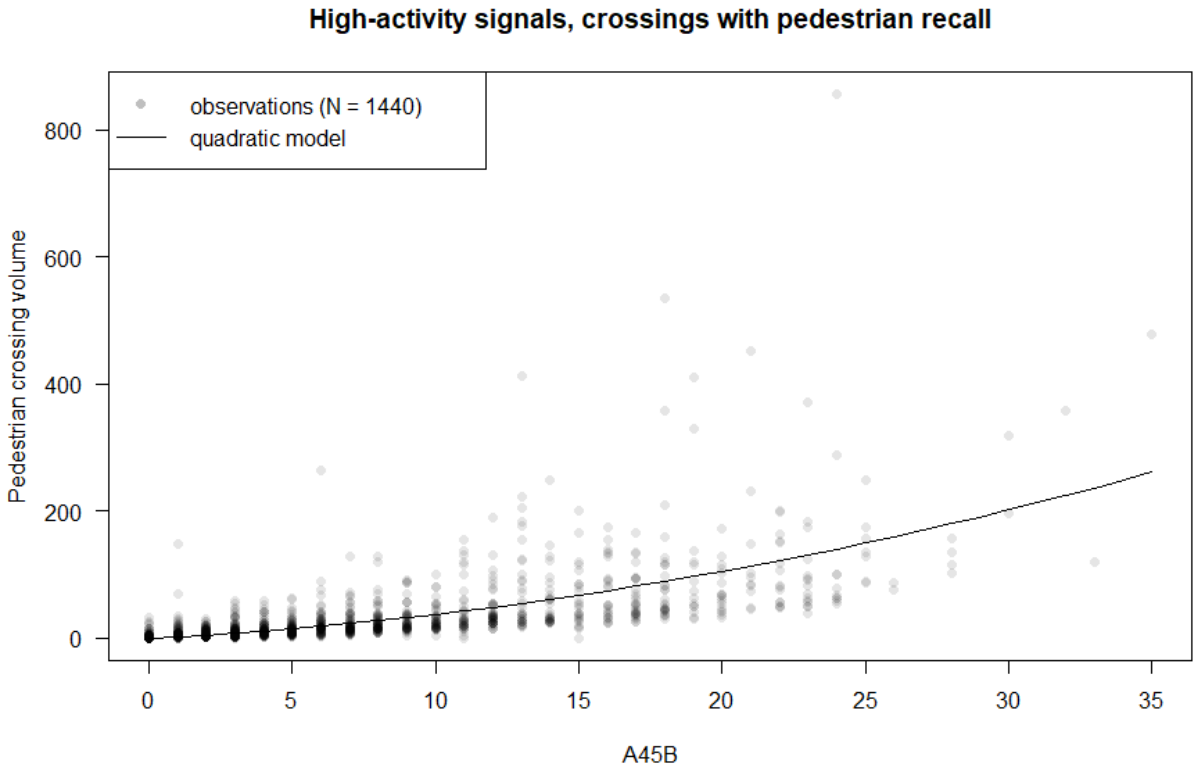


Figure 4.3 Model for crossings with pedestrian recall at high-activity signals

4.3.2.3 Crossings with Pedestrian Recall at Low-Activity Signals

This model applies to all crossings at low-activity signals when they operate with pedestrian recall. There were 3,644 observations in the dataset. The best model was a quadratic specification using the A45B metric. This model had a good correlation of 0.804 and a low mean absolute error of 2.0. The observations and model-predicted values are shown in Figure 4.4.

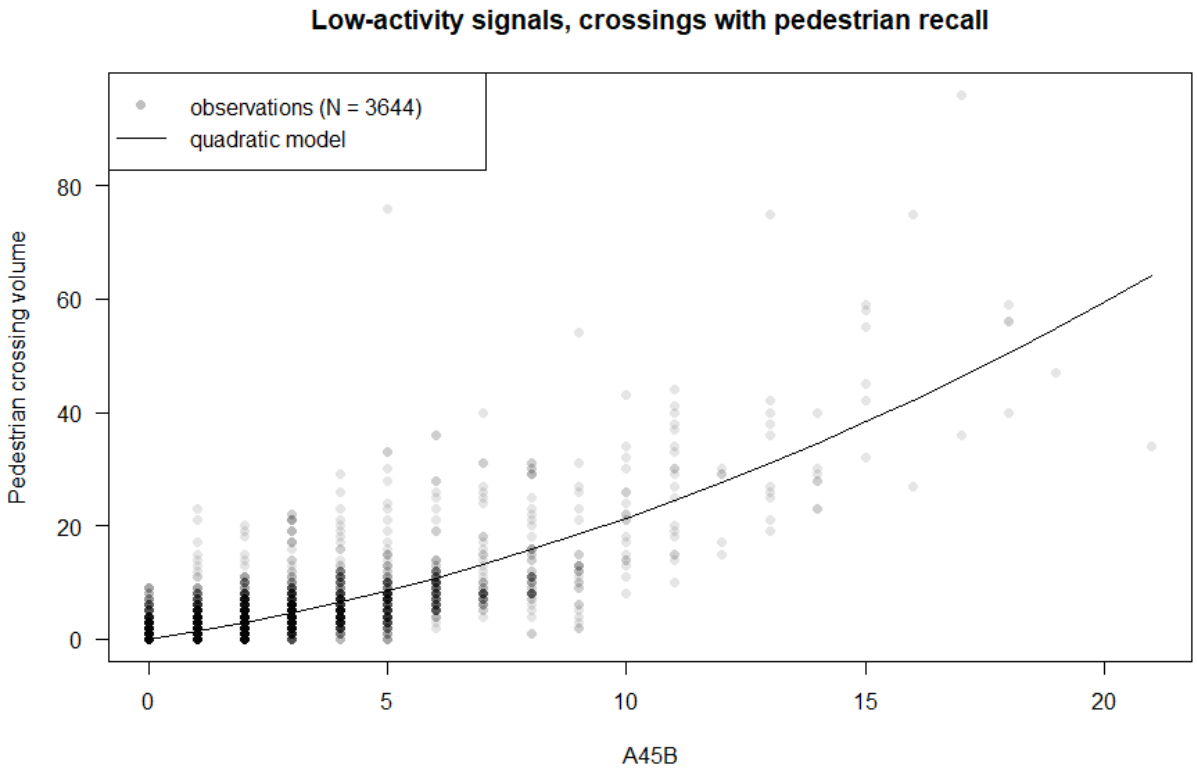


Figure 4.4 Model for crossings with pedestrian recall at low-activity signals

4.3.2.4 Crossings without Pedestrian Recall Having Short Cycle Lengths (average < 1.5 minutes)

This model applies to crossings at all signals that operate not on pedestrian recall and have shorter average cycle lengths less than 90 seconds. There were 5,874 observations in the dataset. The best model was a piecewise linear specification using the A90C metric (with a break in the slope at 28). This model was the best-performing individual model, with a correlation of 0.946 and mean absolute error of 1.6. The observations and model-predicted values are shown in Figure 4.5.

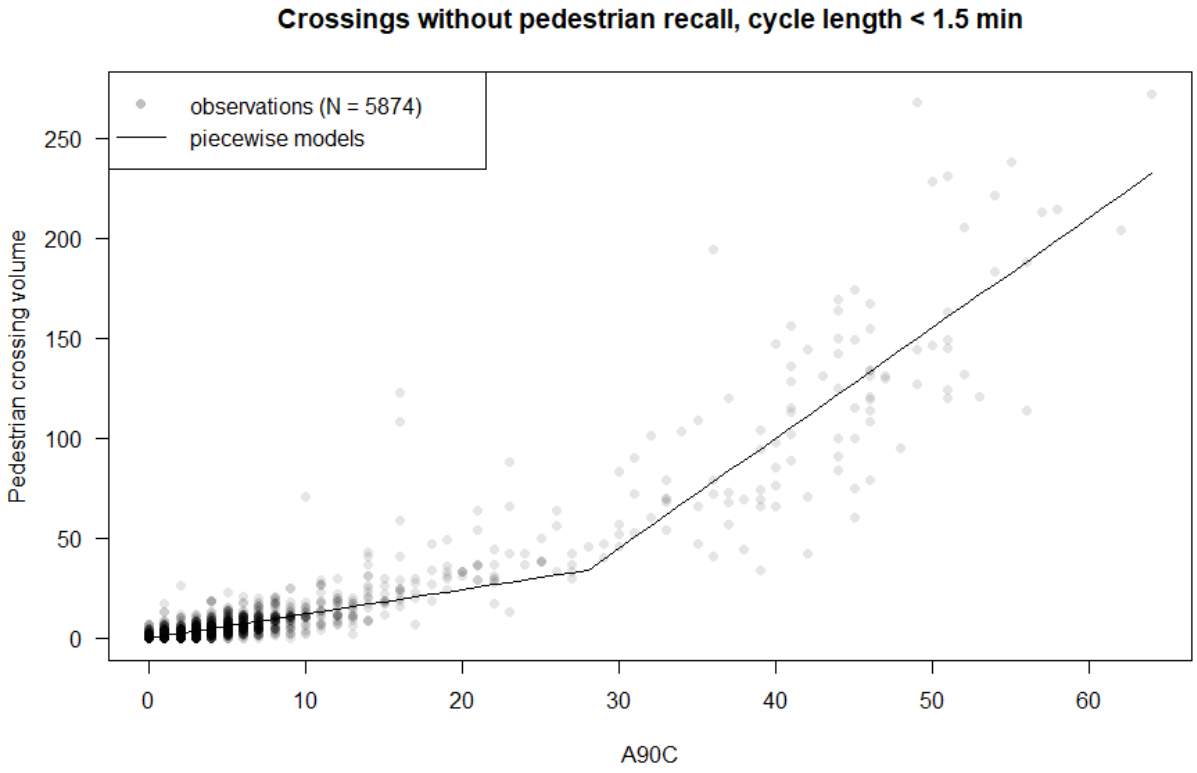


Figure 4.5 Model for crossings without pedestrian recall and with shorter cycle lengths

4.3.2.5 Crossings without Pedestrian Recall Having Long Cycle Lengths (average ≥ 1.5 minutes)

This model applies to crossings at all signals that operate not on pedestrian recall and have longer average cycle lengths equal to or greater than 90 seconds. There were 11,438 observations in the dataset (which represents about half of the sample). The best model was a piecewise linear specification using the A90C metric (with a break in the slope at 28). Note that this model and the previous one were estimated so that the first slope was equal, and that they diverged with different slopes only after the breakpoint. This model also performed strongly, with a correlation of 0.894 and a mean absolute error of 1.9. The observations and model-predicted values are shown in Figure 4.6.

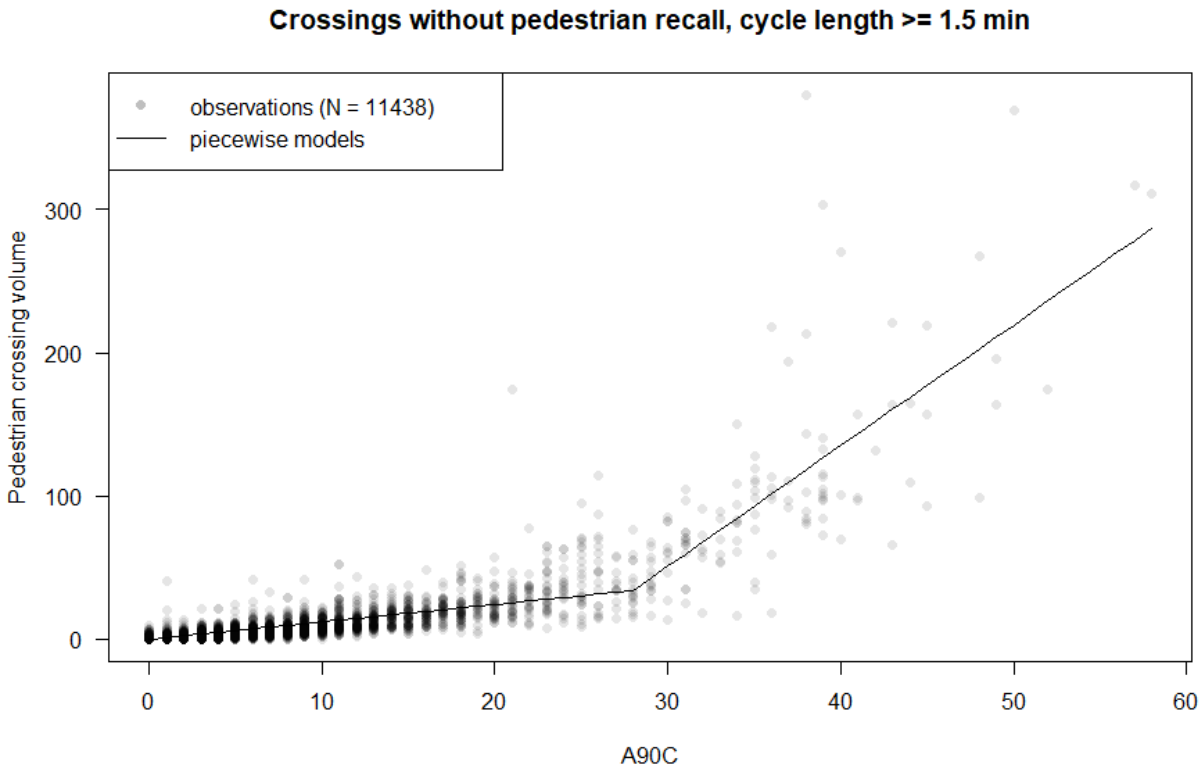


Figure 4.6 Model for crossings without pedestrian recall and with longer cycle lengths

4.3.3 Overall Results

Table 4.4 shows the performance of each of the individual models and the overall performance of the combined models.

Table 4.4 Model Goodness-of-Fit Statistics

Model	N	R^2	RMSE	MAE	Correlation
HAWK signal crossings	234	0.873	35.304	14.623	0.915
Crossings with pedestrian recall at high-activity signals	1,440	0.561	40.836	17.841	0.649
Crossings with pedestrian recall at low-activity signals	3,644	0.723	3.943	1.965	0.804
Crossings without pedestrian recall having short cycle lengths (average < 1.5 minutes)	5,874	0.898	5.625	1.562	0.946
Crossings without pedestrian recall having long cycle lengths (average ≥ 1.5 minutes)	11,438	0.819	6.332	1.885	0.894
Overall	22,630	0.724	12.247	2.961	0.839

Table 4.5 presents the model estimation results for the five models (estimated in one pooled model with many segmentations). All estimated coefficients are statistically significantly different from zero. People looking to apply our models can use these coefficients to estimate pedestrian crossing volumes from pedestrian signal data.

Table 4.5 Model Estimation Results

Model	Variable	B	SE	t	p
HAWK signal crossings	A90C	1.790	0.118	15.147	0.000
	$A90C^2$	0.083	0.003	29.144	0.000
Crossings with pedestrian recall at high-volume signals	A45B	2.304	0.091	25.313	0.000
	$A45B^2$	0.148	0.005	29.380	0.000
Crossings with pedestrian recall at low-volume signals	A45B	1.310	0.132	9.930	0.000
	$A45B^2$	0.083	0.014	6.026	0.000
Crossings without pedestrian recall having short cycle lengths (average < 1.5 minutes)	A90C	1.215	0.018	68.970	0.000
	A90C – 28 (if A90C > 28)	4.292	0.086	50.132	0.000
Crossings without pedestrian recall having long cycle lengths (average ≥ 1.5 minutes)	A90C	1.215	0.018	68.970	0.000
	A90C – 28 (if A90C > 28)	7.214	0.126	57.288	0.000

4.3.4 Example Application: Annual Average Daily Pedestrians

In order to demonstrate the utility of our models, we applied them to a year's worth of data (July 2017 through June 2018) from all traffic signals in Utah to estimate annual average daily pedestrian (AADP) crossing volumes. For each hour at each crossing of each signal, we calculated A45B and A90C, determined whether the phase was likely on pedestrian recall, and calculated the approximate average cycle length, and then we applied the respective models. (Signal data were cleaned for missing data before applying the models.) Finally, we summed our estimates for each signal and day, and averaged daily totals across the year.

We list the top ten highest (estimated) pedestrian volume signalized intersections in Table 4. (There may actually be higher-volume pedestrian intersections in Utah, but many downtown Salt Lake City intersections always operate on pedestrian recall and have no push-buttons and thus no pedestrian activity data.) The high-volume locations make intuitive sense. Most of these signals are located in a small area of downtown Salt Lake City characterized by large centers of employment, shopping, and culture, as well as frequent transit service. For example, Signal 7244 is located adjacent to the Salt Lake City Public Library, the Salt Lake City and County Building, and a light rail station. Two other signals (5807 and 6631) are located at the edge of large university campuses (Utah State University and Brigham Young University). The remaining two high pedestrian volume signals are in downtown Moab, a city in eastern Utah that sees high tourist activity due to its location adjacent to Arches and Canyonlands National Park.

Table 4.6 Signals in Utah with the Highest Estimated Average Pedestrian Volumes

Rank	*Signal*	*Location*	*Estimated AADP*
1	7138	S Temple & State St, Salt Lake City	6,737
2	7244	400 S & 200 E, Salt Lake City	4,868
3	7139	100 S & State St, Salt Lake City	4,519
4	7248	400 S & 600 E, Salt Lake City	4,450
5	5807	700 N & 800 E, Logan	4,446
6	8303	100 S & Main St, Moab	4,307
7	7243	400 S & Main St, Salt Lake City	4,009
8	7142	400 S & State St, Salt Lake City	3,909
9	8302	Center St & Main St, Moab	3,544
10	6631	1230 N & Canyon Rd, Provo	3,476

4.4 Developing a Prototype Visualization/Tool

We developed a prototype online tool and graphical interface that can visualize both raw (or processed) pedestrian traffic signal data as well as estimated pedestrian crossing volumes, the results of applying the factoring methods and models described in the previous section. The proposed visualization contains two views:

- *Map* (Figure 4.7): This view contains a map (and associated table) showing results for one time period across many signals. The user can select the start and end dates and times, and decide which kind of data (signal data, or estimated ped volumes) should be displayed. There is also a toggle to show Total data or Averaged data. If Averaged data is selected, then the user can also specify a time unit (year, month, weekday, hour) as well as the specific time unit to show. For example, you can show the daily values for an average Monday during the spring of 2018, or the specific values for 4pm on the second Monday in May. The symbols on the map change automatically and present the value as shown in both size and color. Data shown in the table can be downloaded as a CSV file.
- *Figure* (Figure 4.8): This view contains a figure (and associated table) showing results for one signal. The user can select the start and end dates and times, indicate whether all or just a select few phases should be displayed, and decide which kind of data (signal data, or estimated ped volumes) should be displayed. There is also the option to specify a time unit (year, month, weekday, hour) for aggregating the data. There is also a toggle to show Total data or Averaged data. If Total data is selected, the figure presents a time series between the specified start/end dates and for the specified time unit. If Averaged data is selected, the figure presents an averaged time series (averaged between the start/end dates) for all versions of the time unit. For example, you can show the average daily values for each weekday during the spring of 2018, or the specific values for each hour on the second Monday in May. The map is clickable to show the values. Data shown in the table can be downloaded as a CSV file.

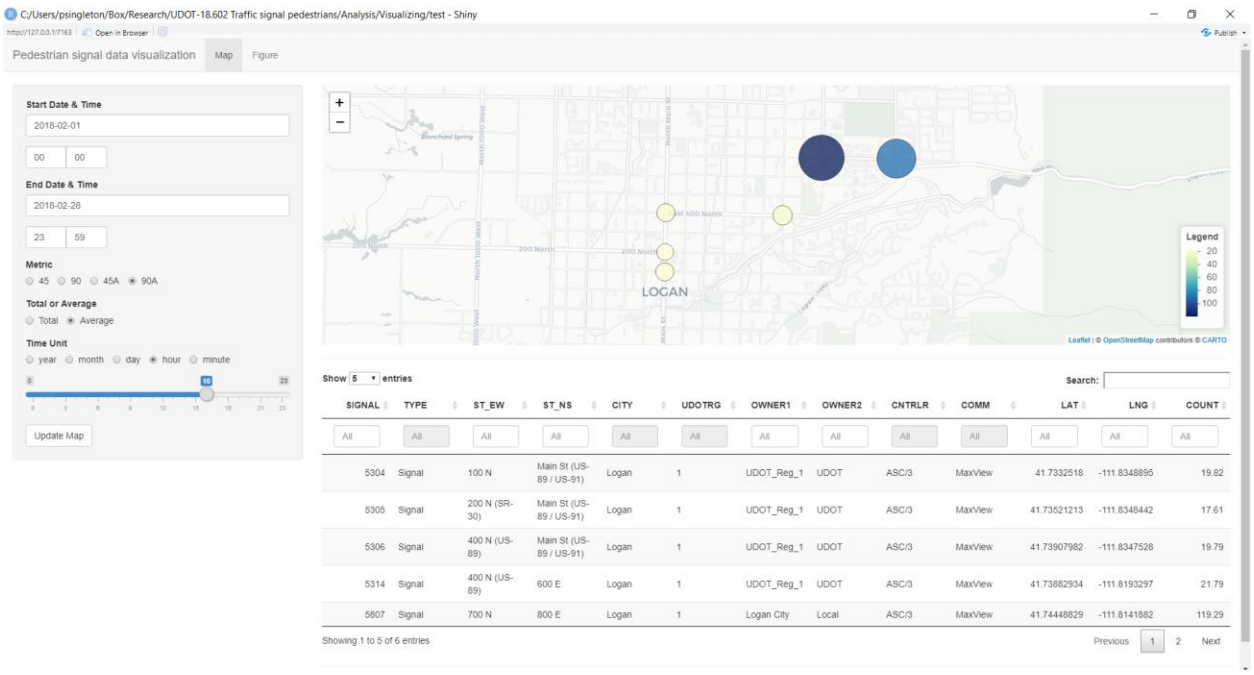


Figure 4.7 Map view of the prototype pedestrian signal activity visualization

Figure 4.8 Figure view of the prototype pedestrian signal activity visualization

The interface was developed as an R Shiny app in R using Shiny and Leaflet. It uses pre-processed data obtained from the ATSPM signal database and is processed using custom R scripts. The app takes this data and visualizes it through various user interfaces.

We hope that this prototype visualization is useful for UDOT when deciding how to proceed with future work visualizing and making sense of pedestrian signal data.

4.4.1 Dashboard Monitoring Pedestrian Activity and COVID-19

Additionally, due to the recent coronavirus pandemic, we have created a similar working online tool and graphical interface to visualize and quantify daily pedestrian signal activity (A90B) in 2020 and compared to the same day in 2019. This dashboard is available at https://singletonpa.shinyapps.io/ped-covid19/ and data are updated every few weeks. The interface has three components:

- *Information* (Figure 4.9): This page contains general information about the dashboard, how to use it, and how it was developed, as well as showing some aggregate trends.
- *Map of many signals* (Figure 4.10): This page displays a map showing data for all signals for a given date. The circles display both the 2020 pedestrian activity (size) as well as the % of 2019 values (color). One can zoom or pan to particular locations, and adjust the date or watch things change over time. There is also a table showing all values that can be downloaded as a CSV file.
- *Figure of one signal* (Figure 4.11): This page displays a figure and table showing 2020 and 2019 data for one signal for a given time period. One can select different date ranges, copy the figure, or download the table as a CSV file.

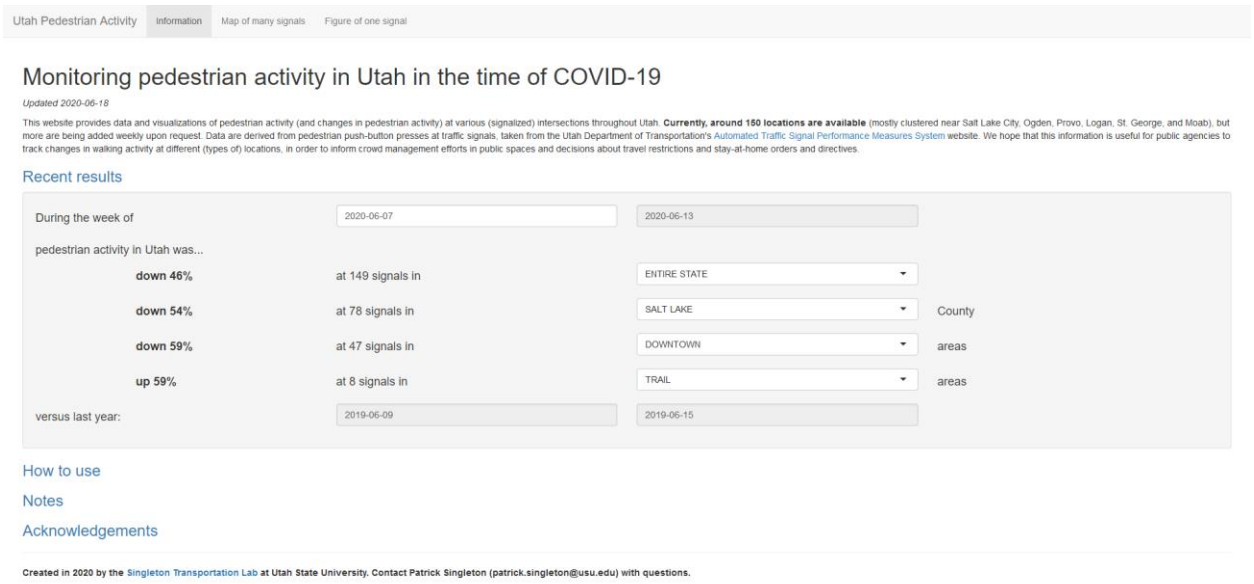


Figure 4.9 Information page of the pedestrian activity COVID-19 dashboard

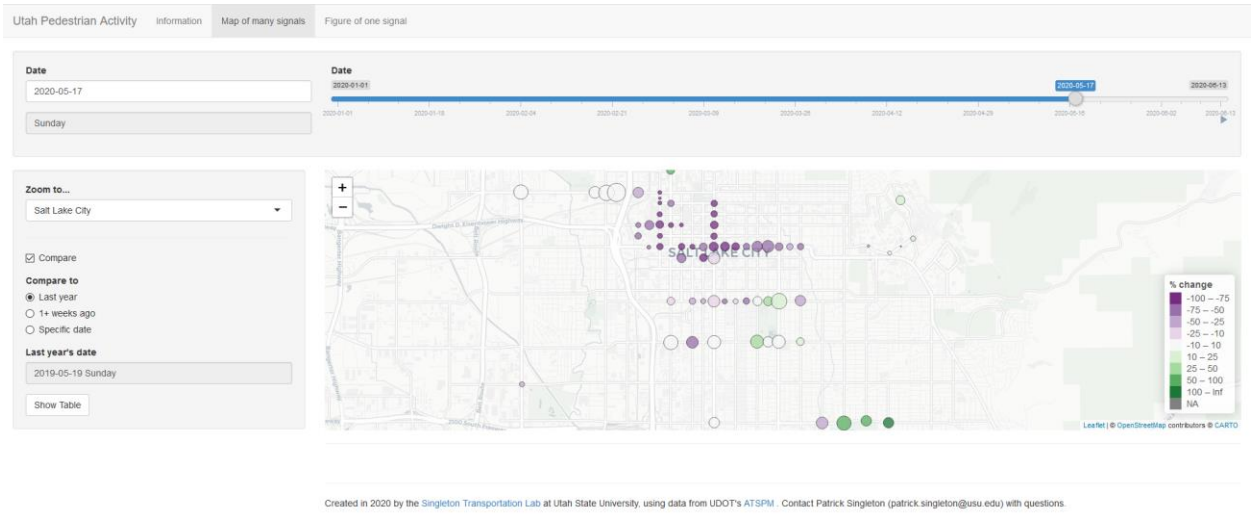


Figure 4.10 Map of many signals page of the pedestrian activity COVID-19 dashboard

Figure 4.11 Figure of one signal page of the pedestrian activity COVID-19 dashboard

The dashboard was developed as an R Shiny app in R using Shiny and Leaflet. Every week, data from the previous week are queried using SQL from the ATSPM signal database. The data are then processed using custom R scripts to calculate the daily totals of signal activity (A90B) at each signal. The app takes this data and visualizes it through various user interfaces.

We hope that this dashboard is useful for UDOT and other agencies for tracking pedestrian activity and making decisions about signal timing, open streets, and other operational activities.

4.5 Summary

In this chapter, we reported detailed results of our two primary analyses of pedestrian traffic signal data. First, we used time-series cluster analysis on one year of average hourly/weekday observations at 1,522 signals in Utah to develop seven typologies of pedestrian activity patterns. The typologies were the result of a cross-classification of clusters on both magnitude and relative shape. There were few high-volume intersections, some medium-volume locations, and many low-volume signals. Each of the medium- and low-volume signals were also split by whether they had a single midday peak or two daily peaks. We also characterized the signal typologies by built environment characteristics, finding expected positive relationships

between higher volume signals and greater residential density and employment accessibility. The higher volume typology was dominated (although not exclusively) by signals in UDOT Region 2, while the lower volume typologies existed in all parts of the state.

Second, we used various simple non-linear regression models to relate observed pedestrian crossing volumes with pedestrian push-button-based measures of pedestrian signal activity. We developed five different models, based on segmentations of signals: HAWK signals, crossings with pedestrian recall at high- or low-activity signals, and crossings without pedestrian recall with short or long cycle lengths. Overall, our models were able to predict pedestrian crossing volumes with good accuracy: a correlation of 0.84 and a mean absolute error of 3.0. We also applied our models to a year's worth of traffic signal data to identify the estimated pedestrian volumes at the highest-volume signalized intersections in Utah.

Third, we also document the prototype visualizations that we have created to display pedestrian signal data and estimated pedestrian volumes. These user interfaces and dashboards have map, figure, and table views, with the ability to download data as well. We hope that these prototypes are useful when visualizing pedestrian signal data and as a starting point for UDOT to develop working implementations.

5.0 CONCLUSIONS

5.1 Summary

The overall objective of this project was to explore the use of continuous pedestrian traffic signal data to develop estimates of pedestrian activity. Towards this aim, this project had three primary analysis objectives:

1. To identify patterns of pedestrian activity at traffic signals.
2. To develop methods to estimate pedestrian volumes from signal data.
3. To create a prototype to visualize pedestrian signal activity.

Section 1.0 introduced the project, while section 2.0 provided background material on the research topic and analysis methods, including time series clustering and linear regression. Section 3.0 described the data collection process for both objectives, including obtaining and processing high-resolution traffic-signal controller log data, recording videos, and counting pedestrian crossings. Section 4.0 reported the data evaluation/analysis, including the seven identified pedestrian activity patterns (typologies) and the five developed regression models (factoring methods) to estimate pedestrian volumes, and described the prototype visualizations. Section 6.0 will provide recommendations for implementation of the research findings. In this section, we conclude by highlighting the major findings from our data collection and analyses, and noting limitations and challenges.

5.2 Findings

5.2.1 Typologies of Pedestrian Signal Activity Patterns

In this project, we investigated the use of a novel source of pedestrian big data (pedestrian push-button information from traffic signal controller logs) and a machine learning technique (time series clustering) to identify typologies of pedestrian activity patterns across signalized intersections in one US state (Utah). In the process, we collected one year's worth of data from more than 1,500 signals, constructed six different pedestrian activity metrics (PAMs)

for each hour in an average week, and tested various clustering techniques, including four different (dis)similarity measures and two algorithms. After statistically and visually analyzing our clustering results, we developed typologies using a cross-classification of two clustering results: (1) the number of pedestrian calls registered (event code 45) per hour, using EU; and (2) the percentage of weekly pedestrian calls registered, using CORT. Our analysis resulted in seven pedestrian activity typologies of intersections, distinguished mostly by the magnitude and number of the daily peaks.

Pedestrian signal data have the potential to overcome a major obstacle: the lack of continuous data on pedestrian activity at multiple locations and for prolonged time periods. Our analysis of average weekly pedestrian activity patterns at signalized intersections across Utah highlights some similarities as well as differences between various typologies. Expected daily and weekly traffic patterns—lower activity overnight and on weekends—appear in all typologies. The magnitude of pedestrian activity across locations is skewed, with many low, some medium, and just a few intersections having very high levels. More interesting are the daily differences: Most intersections have somewhat uniformly high/medium/low levels of midday pedestrian activity (~1% of weekly totals per hour), while some others clearly exhibit an AM and a larger PM peak hour (2–3% of weekly totals) on weekdays. To the extent that these typologies are applicable to other states, this information is useful for understanding detailed spatiotemporal patterns in walking behavior.

A novel contribution of our study is using both *magnitude (counts)* and *weekly patterns* in identifying typologies by combining results from two clustering results. Previous studies related to factor groups assume same factors for locations having similar weekly patterns. In our study, we find that although some intersections have similar weekly patterns (such as Type I, Type II, Type V), there are considerable differences in magnitude of counts. Hence, future studies might note the methodological extension of the study and capture similarities in both magnitude as well as weekly patterns when employing clustering approaches. Additionally, we developed different PAMs for our analysis, and pedestrian calls registered (#45s) was found to be most efficient in producing clusters that are distinct and compact, which could be used as a guide in practitioners pursuing further research on pedestrian push-buttons.

Two approaches, empirical clustering (EC) and land-use-based (LU) methods discussed in relevant literature, have been applied to a very low number of count locations/intersections (<500). The trend in pedestrian data is shifting towards "big data" collected through pedestrian push-buttons (as in our case) or any form of crowd-sourced data. Hence, it is important to acknowledge the viability of both of these classification approaches. In the case of big data, the LU classification approach, which requires knowledge of attributes around the count sites, might be infeasible, and time-consuming. Besides, it is difficult to classify locations with mixed use characteristics (i.e., locations not typically conforming to existing categories such as CBD, school, commercial, etc.) Our study shows that the EC approach could be used for group intersections, and secondly it can also capture the surrounding land use variations simultaneously.

5.2.2 Methods to Estimate Pedestrian Volumes from Traffic Signal Data

In this project, we examined and validated the use of pedestrian data from high-resolution traffic-signal controller logs against observed pedestrian counts, and developed methods that use pedestrian signal data to estimate pedestrian crossing volumes. Specifically, we estimated five simple non-linear regression models that compared hourly pedestrian-signal activity metrics derived from push-button presses against observed pedestrian counts from over 22,000 hours of videos recorded in 2019 for 320 crosswalks at 90 signalized intersections in Utah. The model estimates were strongly correlated with observed pedestrian crossing volumes (0.84) and had a mean absolute error of only 3.0, which is notable given the large sample size and variety of locations studied and the few predictors used.

Model results match expectations considering the interaction between pedestrian behavior and signal operations. The (comparably) poorer fits of the models for crossings with pedestrian recall results from pedestrians not having to press the push-button to receive the walk indication. In situations like this (pedestrian recall, when push-button use is not needed), actuated pedestrian calls (A45B) best predict pedestrian crossing volumes. But when push-button use is needed (not on pedestrian recall), pedestrian detections (A90C) best predict crossing volumes. It makes sense that crossings (on pedestrian recall) at signals with higher pedestrian activity have larger coefficients but poorer model fits, since they likely see larger but more variable group

sizes, which may make it harder to use pedestrian signal data to predict crossing volumes. It is also sensible for crossings (not on pedestrian recall) with longer cycle lengths to have larger multipliers, since the increased wait time can allow larger pedestrian platoons to form.

Overall, the findings from our large-scale validation effort show that traffic-signal big data can be successfully used to estimate pedestrian crossing volumes at intersections. This conclusion is an incredible boon for research and engineering/planning tasks that rely on pedestrian monitoring and pedestrian data collection, especially given the ubiquity of traffic signals in the US. If a region has connected traffic signals (with pedestrian push-buttons) and is archiving controller log data—as many dozens of agencies throughout the country are doing through ATSPM systems—this research immediately opens up these active sensors as a source of pedestrian big data that is available continuously (in time) and for as many locations as there are signals.

5.3 Limitations and Challenges

Despite the promise of pedestrian signal data, they are not without limitations. Like other automatic sensors, they are not perfect measurements of pedestrian volumes, and they are subject to outages or contamination due to equipment malfunctions. More fundamentally, they are only a potential data source at signals with pedestrian push-buttons and where pedestrian actuation is sometimes or always required to get the walk indication. Many downtown areas and older cities do not have or require the use of push-buttons at signalized intersections. However, this signal configuration is common in suburban locations where information about walking is often scarce. Although our models appear to work in the variety of locations and conditions we studied, more research could validate their applicability in areas outside of Utah.

Finally, we would like to note that the ability to estimate pedestrian volumes from traffic signal data should not be a justification for making pedestrian crossings actuated (having people press the pedestrian push-button). There are many operational reasons to do this, and many other reasons to put signal phases on pedestrian recall (and rest in walk). Our view is that this ability is a fortunate side -ffect of the standard practice of pedestrian operations at US traffic signals.

6.0 RECOMMENDATIONS AND IMPLEMENTATION

6.1 Recommendations

This research directly addresses one of UDOT's "top 10" goals: for real-time full situational awareness of the performance and operation of Utah's transportation system. UDOT is a national leader in developing and deploying the ATSPM system for real-time management and archived performance assessment of traffic signals throughout the state. Although these systems record pedestrian actuations and include a pedestrian delay performance measure, little else is done with these data. This project collected true pedestrian counts and successfully validated the signal actuations, in the process developing methods to estimate pedestrian volumes and thus glean more information from these data, extending their usefulness into the planning realm. Overall, improved estimates of pedestrian activity can be used for traffic signal operations and timing, pedestrian traffic safety analyses, and health assessments, all of which help to promote UDOT's mission of enhancing quality of life.

Because UDOT is a leader in ATSPM, validating pedestrian actuation data and investigating the applicability of these data for pedestrian planning and analysis will be useful for DOTs in the rest of the country. Agencies in other states may be able to borrow the measures and methods used (or developed) in this research project to utilize pedestrian-signal actuation data for improved pedestrian planning in their own jurisdictions.

There are many potential uses of pedestrian signal data and the models developed in this research project. Investments in pedestrian infrastructure (such as sidewalks and improved crossings) can be prioritized in areas with higher levels of pedestrian activity. Pedestrian volumes can be used as a measure of exposure in safety studies to more accurately identify pedestrian crash risk factors or evaluate the effectiveness of a safety treatment. They could also be associated with attributes of the surrounding area to identify stronger (and temporal variations in) built environment relationships with walking, thus informing the design of walk-friendly communities. Level-/quality-of-service calculations require pedestrian flow rates. Temporal patterns in pedestrian signal detection can offer guidelines for reconfiguring traffic signal operations for pedestrians, such as implementing pedestrian recall, leading pedestrian intervals,

no right turns on red, or even pedestrian scrambles (Barnes dances). This information could be correlated with weather and air-quality data to learn more about how atmospheric and environmental conditions affect pedestrian behaviors, or tracked over time to notice changes due to major events, natural disasters, or outbreaks of infectious diseases. Pedestrian traffic signal data can also be used to develop improved temporal factors and factor groups that allow for the extrapolation of short-duration pedestrian counts into average pedestrian volumes.

6.1.1 Future Research

We envision several potentially fruitful avenues for future research. First, it would be useful to validate the models and methods developed in this project both in other states (cross-sectionally) and ongoing (over time). Although our approach validated the use of traffic signal data as a reasonably accurate data source to predict pedestrian-intersection crossing volumes using orders-of-magnitude more data than had previously been collected, regression methods like those employed in this research can be subject to overfitting, thus yielding slightly less accurate predictions when applied in new situations (other locations not studied) or in the future (due to behavioral changes). Thus, conducting a similar data collection and validation process in other states would determine the generalizability of the project's key findings and products to other locations and contexts outside of Utah. We expect our models to perform fairly well in other contexts, but there could be a few special circumstances where they are not as appropriate. We also recommend that the findings should be validated with periodic additional data collection, even within Utah. Doing some modest data collection (at 5-10 locations, perhaps every 5 years) and comparison of pedestrian signal data against observed pedestrian crossings will ensure that the factors and methods remain valid into the future. This work is especially important during and after major disruptions—like the coronavirus pandemic—that have the potential to change pedestrian behaviors. Similarly, if traffic signal operations change significantly, then similar validation work should be done.

Although we hope our prototype visualizations are useful, more research could be done to determine the desired functionalities and user interactivity needs from a diverse set of pedestrian data users. This research would involve several rounds of user testing with stakeholders, to develop potential ways to quantify and visualize pedestrian data. For example, people working in

traffic signal operations need to be able to drill down to individual crossings and specific hours on particular days, to understand how pedestrians are interacting with the traffic signal. On the other hand, transportation planners may want more averaged and aggregated data and the ability to compare pedestrian volumes at different times of time, days of the week, or seasons. These diverse needs may require different types of user interfaces, customized to each set of potential applications of pedestrian signal data.

Finally, we think that this work will open up new avenues of transportation research, especially in areas that were lacking for pedestrian volume data. Research on pedestrian safety requires measures of pedestrian "exposure" to traffic (exposure is usually measured as volumes), but pedestrian exposure data for safety analysis often must rely on short-duration counts or proxies (e.g., neighborhood socio-demographics). Research on the built environment correlates of walking activity also suffers from a lack of long-term data on pedestrian volumes. Due to the temporal coverage of traffic signal data (collected continuously over long time periods), we expect that pedestrian signal data will help to identify associations between pedestrian activity and shorter-term (e.g., daily or hourly) measures of weather and air quality, to better understand behavioral sensitivities to temperature, precipitation, and air pollution. Finally, the empirical clustering method we applied to develop typologies of pedestrian activity could be adapted to help develop pedestrian factor groups for improved pedestrian travel monitoring. Such work to identify similar patterns and expansion factors would improve the ability to convert short-duration (e.g., peak hour) pedestrian counts into estimates of annual average daily pedestrian volumes at many locations.

6.2 Implementation Plan

Currently, our procedures for processing pedestrian traffic signal data, applying the regression models to estimate pedestrian crossing volumes, and visualizing these data have been implemented on local machines using manually-downloaded ATSPM data and customized scripts coded in the open-source programming language R. (See Appendix B.) In order to fully implement the products of this research (regression models and interactive visualizations), several general steps are needed.

First, the linkages between the ATSPM traffic signal data, the processing scripts, and the visualizations needs to be automated. Currently, our R scripts take manually downloaded high-resolution traffic-signal controller log data, clean and process the data, apply the regression models, and save the output for later visualization. Other visualization scripts then query those saved pedestrian data for displaying in the user interfaces. Ideally, scripts would directly query (at pre-determined intervals, such as every hour, day, or week) the ATSPM data server, process those data in the background, and perhaps save the (processed and somewhat aggregated) results in one or a set of intermediate pedestrian databases. The user interfaces could then query the intermediate pedestrian databases much faster on demand. All of this work would involve the deployment of servers to host the intermediate pedestrian databases, websites for the interactive user interfaces, and the processing scripts that can communicate with the ATSPM data server. Although this could be integrated into the existing ATSPM system, we envision a broad set of potential users and uses, so a separate (linked) system could be useful.

Second, the processing scripts could be improved in several ways. For one, R is likely not the most efficient programming language for rapidly processing large quantities of data. Converting our existing R scripts and processes into other programming languages (such as Python) could make the procedures more efficient. Also, we need to develop better algorithms for measuring pedestrian traffic-signal data quality and identifying missing or erroneous data. Data could be missing (for an entire signal, or for a particular crossing) in various time periods due to gaps in connectivity, malfunctioning equipment, or maintenance and construction work. Alternatively, malfunctioning push-buttons could send false signals of pedestrian activity. Distinguishing an hour of missing data from an overnight hour with no traffic can be challenging, as can be distinguishing a stuck push-button from the true activity of a large pedestrian crowd. If data are not carefully cleaned, estimates of pedestrian volumes could be biased (too high or too low). We developed some ad-hoc methods to detect missing data, but these rely on subjective time-difference calculations and may be too sensitive when actual volumes are low. At this time, we have not developed automated methods to detect erroneous (stuck push-button) data. These tasks are necessary to fully and accurately implement the methods and tools developed through this research project.

REFERENCES

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering – A decade review. *Information Systems, 53,* 16–38. https://doi.org/10.1016/j.is.2015.04.007

Aköz, Ö., & Karsligil, M. E. (2014). Traffic event classification at intersections based on the severity of abnormality. *Machine Vision and Applications, 25*(3), 613–632. https://doi.org/10.1007/s00138-011-0390-4

ATKINS. (2016). *Automated Traffic Signal Performance Measures Reporting Details*. Atlanta, GA: Georgia Department of Transportation. https://udottraffic.utah.gov/ATSPM/Images/ATSPM_Reporting_Details.pdf

Blanc, B., Johnson, P., Figliozzi, M., Monsere, C., & Nordback, K. (2015). Leveraging signal infrastructure for nonmotorized counts in a statewide program: Pilot study. *Transportation Research Record: Journal of the Transportation Research Board, 2527,* 69–79. https://doi.org/10.3141/2527-08

Chouakria, A. D., & Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification, 1*(1), 5–21. https://doi.org/10.1007/s11634-006-0004-6

Côme, E. & Oukhellou, L. (2014). Model-based count series clustering for bike sharing system usage mining: A case study with the Velib' system of Paris. *ACM Transactions on Intelligent Systems and Technology, 5*(3), 39–39. https://doi.org/10.1145/2560188

Datesh, J., Scherer, W. T., & Smith, B. L. (2011). Using k-means clustering to improve traffic signal efficacy in an IntelliDrive SM environment. Presented at the 2011 IEEE Forum on Integrated and Sustainable Transportation Systems, Vienna, AT. https://doi.org/10.1109/FISTS.2011.5973659

Day, C. M., Premachandra, H., & Bullock, D. M. (2011). Rate of pedestrian signal phase actuation as a proxy measurement of pedestrian demand. Transportation Research

Record. Presented at the 90th Annual Meeting of the Transportation Research Board, Washington, DC. https://docs.lib.purdue.edu/civeng/24/

Day, C. M., Bullock, D. M., Li, H., Remias, S. M., Hainen, A. M., Freije, R. S., ... & Brennan, T. M. (2014). *Performance measures for traffic signal systems: An outcome-oriented approach*. West Lafayette, IN: Purdue University. https://doi.org/10.5703/1288284315333

Day, C. M., Taylor, M., Mackey, J., Clayton, R., Patel, S. K., Xie, G., ... & Bullock, D. (2016). Implementation of Automated Traffic Signal Performance Measures. *ITE Journal, 86*(8), 26–34.

Federal Highway Administration (FHWA). (2016). *Traffic monitoring guide*. Washington, DC: U.S. Department of Transportation. https://www.fhwa.dot.gov/policyinformation/tmguide/

Hankey, S., Lindsey, G., Wang, X., Borah, J., Hoff, K., Utecht, B., & Xu, Z. (2012). Estimating use of non-motorized infrastructure: Models of bicycle and pedestrian traffic in Minneapolis, MN. *Landscape and Urban Planning, 107*(3), 307–316. https://doi.org/10.1016/j.landurbplan.2012.06.005

Kassambara, A. (2017). P*ractical guide to cluster analysis in R: Unsupervised machine learning. SDHTA*. https://www.datanovia.com/en/wp-content/uploads/dn-tutorials/book-preview/clustering_en_preview.pdf

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies, 1*(6), 90–95.

Kothuri, S., Nordback, K., Schrope, A., Phillips, T., & Figliozzi, M. (2017). Bicycle and pedestrian counts at signalized intersections using existing infrastructure: Opportunities and challenges. *Transportation Research Record: Journal of the Transportation Research Board, 2644*, 11–18. https://doi.org/10.3141/2644-02

Liao, T. W. (2005). Clustering of time series data—A survey. *Pattern recognition, 38*(11), 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025

Montero, P., & Vilar, J. A. (2014). TSclust: An R package for time series clustering. *Journal of Statistical Software, 62*(1), 1–43. https://doi.org/10.18637/jss.v062.i01

Ramsey, K., & Bell, A. (2014). *Smart Location Database: Version 2.0 User Guide*. Washington, DC: U.S. Environmental Protection Agency. https://edg.epa.gov/data/PUBLIC/OP/SLD

Ryus, P., Ferguson, E., Laustsen, K. M., Proulx, F. R., Schneider, R. J., Hull, T., & Miranda-Moreno, L. (2014). *Methods and technologies for pedestrian and bicycle volume data collection* (NCHRP Web-Only Document 205). Washington, DC: Transportation Research Board. https://doi.org/10.17226/23429

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence, 2*(2), 34–38.

Smaglik, E. J., Sharma, A., Bullock, D. M., Sturdevant, J. R., & Duncan, G. (2007). Event-based data collection for generating actuated controller performance measures. *Transportation Research Record: Journal of the Transportation Research Board, 2035*, 97–106. https://doi.org/10.3141/2035-11

Smith, B. L., & Demetsky, M. J. (1994). Short-term traffic flow prediction models-a comparison of neural network and nonparametric regression approaches. Presented at the 1994 IEEE International Conference on Systems, Man and Cybernetics, San Antonio, TX. https://doi.org/10.1109/ICSMC.1994.400094

StreetLight Data. (2019). *StreetLight InSight Active Transportation Metrics: Methodology and Validation*. StreetLight Data. https://learn.streetlightdata.com/proven-metrics-for-active-transportation-plans

Sturdevant, J. R., Overman, T., Raamot, E., Deer, R., Miller, D., Bullock, D. M., ... & Remias, S. M. (2012). *Indiana traffic signal hi resolution data logger enumerations*. West Lafayette, IN: Purdue University. http://dx.doi.org/10.4231/K4RN35SH

Taylor, M., & Mackey, J. (2018). Automated Traffic Signal Performance Measures. Presented at the 2018 UDOT Annual Conference, Sandy, UT. https://udottraffic.utah.gov/ATSPM/Images/Session%2027_ATSPMs_UDOT%20Conference_20181106.pdf

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of Royal Statistical Society, 63*(2), 411-423. https://doi.org/10.1111/1467-9868.00293

Urbanik, T., Tanaka, A., Lozner, B., Lindstrom, E., Lee, K., Quayle, S., ... & Sunkari, S. (2015). *Signal timing manual: Second edition* (NCHRP Report 812). Washington, DC: Transportation Research Board. https://doi.org/10.17226/22097

Wang, X., Cottrell, W., & Mu, S. (2005). Using k-means clustering to identify time-of-day break points for traffic signal timing plans. Presented at the 2005 IEEE Intelligent Transportation Systems, Vienna, AT. https://doi.org/10.1109/ITSC.2005.1520102

APPENDIX A: LIST OF TRAFFIC SIGNALS USED IN VIDEO DATA COLLECTION

Table A.1 List of Signalized Intersections Studied with Video Data Collection

Signal ID	Street E/W	Street N/S	City
1021	1300 S	300 W	Salt Lake City
1045	700 S	Main St	Salt Lake City
1094	South Temple	"I" St / 700 E	Salt Lake City
1225	800 S	1300 E	Salt Lake City
1229	2100 S	1300 E	Salt Lake City
1801	South Temple	50 E (HAWK)	Salt Lake City
1803	800 S	1250 E (HAWK)	Salt Lake City
4011	3900 S	2300 E	Salt Lake County/Holladay
4020	3900 S	210 W	South Salt Lake/Salt Lake County
4024	Fort Union Blvd (6910 S)	1300 E	Cottonwood Heights
4130	6200 S	Jordan Canal Rd / Margray Dr (1950 W)	Taylorsville
4158	Highland Dr (14700 S)	Minuteman Dr (100 W)	Draper
4301	Fort Union Blvd (7000 S)	Union Park Ave (1090 E)	Midvale/Cottonwood Heights
4406	7755 S/Forbush Ln	1300 E (Union Park Ave)	Sandy/Cottonwood Heights
4502	3100 S	Constitution Blvd (2700 W)	West Valley City
4511	4100 S	3200 W	West Valley City
4522	3100 S	Decker Lake Dr (2210 W)	West Valley City
4662	Herriman Pkwy (12600 S)	Herriman Main St (5100 W)	Herriman
4895	New Bingham Hwy (8200 S)	4800 W	West Jordan
5024	24th St	Washington (US-89)	Ogden
5030	12th St (SR-39)	Washington (US-89)	Ogden
5053	36th St	Harrison Blvd (SR-203)	Ogden
5093	4800 S	1900 W (SR-126)	Roy
5108	Antelope Dr (2000 N)	Hill Field Rd (SR-232)	Layton
5170	200 N (SR-273)	Main St (SR-273)	Kaysville
5179	Washington (US-89)	Harrison Blvd (SR-203)	South Ogden
5221	205 S (SR-193)	2000 W (SR-108)	Syracuse
5260	Syracuse JHS HAWK	2000 W (SR-108)	Syracuse
5299	Main St (SR-142)	US-91 (200 W)	Richmond
5305	200 N (SR-30)	Main St (US-89 / US-91)	Logan
5306	400 N (US-89)	Main St (US-89 / US-91)	Logan
5311	1400 N	Main St (US-91)	Logan

Signal ID	Street E/W	Street N/S	City
5315	600 S	1000 W (SR-252)	Logan
5330	1700 S / 800 W	US-89/US-91	Logan
5332	1200 S	Main St (SR-165)	Logan/Providence
5349	2600 S (SR-93)	US-89	Woods Cross/North Salt Lake/Bountiful
5363	400 N (SR-106)	500 W (US-89)	Bountiful/West Bountiful
5393	Antelope Dr (1700 S / SR-127 / SR-108)	2000 W (SR-108)	Syracuse
5618	Gentile St	Flint St	Layton
5624	Wasatch Dr	Fairfield Rd	Layton
5702	500 S	Main St	Bountiful
6038	Pioneer Crossing (SR-145)	2300 W (9550 W)	Lehi
6047	Arrowhead Trail Rd (SR-164)	Main St (SR-198)	Spanish Fork
6078	Pony Express Pkwy	Redwood Rd (SR-68)	Saratoga Springs
6125	Main St (US-40)	Vernal Ave (US-191)	Vernal
6146	Cory Wride Hwy (SR-73)	Ranches Pkwy	Eagle Mountain
6168	1400 S	SR-198	Payson
6303	800 N (SR-52)	State St (US-89)	Orem
6307	800 N (SR-52)	Palisade Dr	Orem
6393	1600 N	State St (US-89)	Orem
6407	Center St	University Ave (US-189)	Provo
6436	550 W/2230 N	University Pkwy (SR-265)	Provo
6446	1230 N (Bulldog) / Columbia Ln	500 W (US-89)	Provo
7041	South Campus Dr (SR-282)	1725 E	Salt Lake City
7060	3500 S (SR-171)	Bangerter Hwy (SR-154)	West Valley City
7086	North Temple	Redwood Rd (SR-68)	Salt Lake City
7099	2320 S	Redwood Rd (SR-68)	West Valley City
7110	5400 S (SR-173) CFI Master	Redwood Rd (SR-68)	Taylorsville
7119	12600 S (SR-71)	Redwood Rd (SR-68)	Riverton
7126	South Temple	300 W (US-89)	Salt Lake City
7160	5300 S (SR-173)	State St (US-89)	Murray
7164	6400 S (Winchester)	State St (US-89)	Murray
7184	900 S	700 E (SR-71)	Salt Lake City
7218	Wakara Wy	Foothill Blvd (SR-186)	Salt Lake City
7285	3500 S (SR-171)	3200 W	West Valley City
7328	5400 S (SR-173)	4015 W	Salt Lake County/Taylorsville

Signal ID	Street E/W	Street N/S	City
7332	5400 S (SR-173)	2200 W TSC - Flex Lanes	Taylorsville
7355	13800 S	Bangerter Hwy (SR-154)	Draper
7381	3500 S (SR-171)	5600 W (SR-172)	West Valley City
7382	4100 S	5600 W (SR-172)	West Valley City
7464	5415 S (SR-173)	4420 W	Salt Lake County
7475	50 S HAWK	300 W (US-89)	Salt Lake City
7511	13400 S	SR-85 NB (Mountain View)	Riverton
7610	5415 S (SR-173)	4800 W	Salt Lake County
7622	11400 S (SR-175)	Redwood Rd (SR-68)	South Jordan
7719	HAWK (Homestake Rd)	Park Ave (SR-224)	Park City
7812	2000 N	SR-36 (Main St)	Tooele
8105	Sunset Blvd (SR-8)	Dixie Dr	St. George
8113	Main St/Hilton Dr	Bluff St (SR-18)	St. George
8117	St. George Blvd (SR-34)	Main St	St. George
8119	St. George Blvd (SR-34)	400 E	St. George
8124	St. George Blvd (SR-34)	Red Cliffs Dr / River Rd	St. George
8208	200 N (SR-56)	Main St (SR-130)	Cedar City
8222	200 N (SR-56)	I-15 NB Ramps/1225 W	Cedar City
8302	Center St	Main St (US-191)	Moab
8601	100 S	Main St	St. George
8627	850 N	3050 E	St. George
8634	Brigham Rd	River Rd	St. George
8725	Pioneer Pkwy	Rachel Dr	Santa Clara
8828	Red Cliffs Dr / Telegraph St	Green Springs Dr	Washington

APPENDIX B: LIST OF ASSOCIATED DATA, SCRIPTS, AND DOCUMENTATION

The following datasets, scripts, and associated documentation are made available to UDOT as part of the deliverables of this project. Included in a Data_Scripts folder are the following (with information files containing more details):

- *Typologies*: Scripts, data, and results of the cluster analysis to identify patterns (typologies) of pedestrian activity at traffic signals.
- *Example Data Collect*: Example scripts, raw data, video, and combined data showing how to collect and process pedestrian event data and combine it with signal data.
- *Data*: Data collected and assembled, including about videos, pedestrian crossing events, and combined with signal data.
- *Models*: Scripts, data, and results of the regression modeling to develop (factoring) methods to estimate pedestrian volumes from signal data.
- *Example Apply Models*: Example scripts, data, and results of applying the regression models to raw signal data and estimating pedestrian volumes.
- *Visualization*: Scripts, data, and interfaces that create a prototype to visualize pedestrian signal activity.

These scripts were written in R. To use, download R (https://cloud.r-project.org/) and then download RStudio (https://rstudio.com/products/rstudio/download/#download). Then, follow additional instructions in each folder.