MOUNTAIN-PLAINS CONSORTIUM

MPC 24-525 | P. A. Singleton, A. Rafe, P. Humagain, F. Runa, A. Islam and M. Mekker

UTILIZING TRAFFIC SIGNAL PEDESTRIAN PUSH-BUTTON DATA FOR PEDESTRIAN PLANNING AND SAFETY ANALYSIS

A University Transportation Center sponsored by the U.S. Department of Transportation serving the Mountain-Plains Region. Consortium members:

Colorado State University
North Dakota State University
South Dakota State University
University of Colorado Denver
University of Denver
University of Utah
Utah State University
University of Wyoming

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
MPC-622		

4. Title and Subtitle	5. Report Date
Utilizing Traffic Signal Pedestrian Push-Button Data for Pedestrian Planning and Safety Analysis	June 2024
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Patrick Singleton Amir Rafe Prasanna Humagain Ferdousy Runa Ahadul Islam Michelle Mekker	MPC 24-525

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Utah State University 4110 Old Main Hill Logan, UT, 84322-4110	11. Contract or Grant No.

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Mountain-Plains Consortium North Dakota State University PO Box 6050, Fargo, ND 58108	Final Report
	14. Sponsoring Agency Code

15. Supplementary Notes

Supported by a grant from the US DOT, University Transportation Centers Program

16. Abstract

Transportation planning, traffic monitoring, and traffic safety analysis require detailed information about pedestrian volumes, but such data are usually lacking. Fortunately, recent research has demonstrated the accuracy of pedestrian volumes estimated from push-button data contained within high-resolution traffic signal controller log data. Such data are available continuously for many locations. This project takes advantage of these novel pedestrian traffic signal data to advance pedestrian traffic monitoring and improve pedestrian traffic safety by applying them as estimates of volume and exposure, often alongside advanced machine learning techniques. Through a series of five studies, we identify temporal patterns in pedestrian activity; study the accuracy of pedestrian volume estimation methods over time; use machine learning methods to improve the quality and completeness of pedestrian time-series data; analyze crashes to identify a “safety in numbers” effect for pedestrians; and apply a new deep learning model to better understand factors affecting pedestrian crash severity. Altogether, this work leverages novel pedestrian traffic signal data to further research and efforts in pedestrian traffic monitoring and safety.

17. Key Word	18. Distribution Statement
built environment, data collection, pedestrian actuated controllers, pedestrian safety, signalized intersections, traffic volume, transportation planning	Public distribution

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	105	n/a

Utilizing Traffic Signal Pedestrian Push-Button Data for Pedestrian Planning and Safety Analysis

Patrick A. Singleton
Associate Professor
Civil and Environmental Engineering
Utah State University

Amir Rafe
Graduate Research Assistant
Civil and Environmental Engineering
Utah State University

Prasanna Humagain
Graduate Research Assistant
Civil and Environmental Engineering
Utah State University

Ferdousy Runa
Graduate Research Assistant
Civil and Environmental Engineering
Utah State University

Ahadul Islam
Graduate Research Assistant
Civil and Environmental Engineering
Utah State University

Michelle Mekker
Assistant Professor
Civil and Environmental Engineering
Utah State University

June 2024

Acknowledgements

The work in Chapters 2 and 3 was supported in part by the Utah Department of Transportation (Research Project UT-18.602), as was the work in Chapter 5 (UT-19.316). Thanks to Doo Hong Lee and Keunhyun Park of Utah State University for calculating the land use, built environment, and sociodemographic characteristics used in multiple chapters. Several anonymous peer reviewers offered constructive comments that improved the work documented in this report.

Disclaimer

ABSTRACT

Transportation planning, traffic monitoring, and traffic safety analysis require detailed information about pedestrian volumes, but such data are usually lacking. Fortunately, recent research has demonstrated the accuracy of pedestrian volumes estimated from push-button data contained within high-resolution traffic signal controller log data. Such data are available continuously for many locations. This project takes advantage of these novel pedestrian traffic signal data to advance pedestrian traffic monitoring and improve pedestrian traffic safety by applying them as estimates of volume and exposure, often alongside advanced machine learning techniques. Through a series of five studies, we identify temporal patterns in pedestrian activity; study the accuracy of pedestrian volume estimation methods over time; use machine learning methods to improve the quality and completeness of pedestrian time-series data; analyze crashes to identify a "safety in numbers" effect for pedestrians; and apply a new deep learning model to better understand factors affecting pedestrian crash severity. Altogether, this work leverages novel pedestrian traffic signal data to further research and efforts in pedestrian traffic monitoring and safety.

TABLE OF CONTENTS

1. INTRODUCTION 1
1.1 Research Objectives 2
1.2 Research Approach and Overview 2
1.3 References 3
2. ADVANCES IN PEDESTRIAN TRAVEL MONITORING: TEMPORAL PATTERNS AND SPATIAL CHARACTERISTICS USING PEDESTRIAN PUSH-BUTTON DATA FROM UTAH TRAFFIC SIGNALS 4
2.1 Abstract 4
2.2 Introduction 4
2.2.1 Pedestrian Push-button Data to Measure Pedestrian Activity 5
2.2.2 Applications of Continuously-measured Pedestrian Data 5
2.2.3 Pedestrian Expansion Factors 6
2.2.4 Research Objectives 6
2.3 Data and Methods 6
2.3.1 Data Sources and Preparation 7
2.3.2 Analysis Methods 9
2.4 Results 10
2.4.1 Hourly/weekday Patterns 10
2.4.2 Monthly Patterns 15
2.4.3 Cross-classification of Hourly/weekday and Monthly Clusters 19
2.5 Discussion and Conclusions 19
2.5.1 Limitations 22
2.5.2 Future Research 23
2.6 References 23
3. IMPACTS OF THE COVID-19 PANDEMIC ON PEDESTRIAN PUSH-BUTTON UTILIZATION AND PEDESTRIAN VOLUME MODEL ACCURACY IN UTAH 26
3.1 Abstract 26
3.2 Introduction 26
3.2.1 Background and Research Questions 27
3.3 Data and Methods 28
3.3.1 Pedestrian Data Collection 28
3.3.2 Pedestrian Push-button Data Assembly 29
3.3.3 Analysis of Changes in Pedestrian Push-button Utilization 30
3.3.4 Analysis of Changes in the Accuracy of Pedestrian Crossing Volume Estimates 30
3.4 Results and Discussion 30
3.4.1 Analysis of Changes in Pedestrian Push-button Utilization 30
3.4.2 Analysis of Changes in the Accuracy of Pedestrian Crossing Volume Estimates 32
3.5 Conclusion 34
3.6 References 35
4. IMPUTING TIME SERIES PEDESTRIAN VOLUME DATA WITH CONSIDERATION OF EPIDEMIOLOGICAL-ENVIRONMENTAL (EPIENV) VARIABLES 37
4.1 Abstract 37
4.2 Introduction 37
4.3 Literature Review 38
4.4 Data 41
4.4.1 Estimated Pedestrian Volumes from Traffic Signal Data 41
4.4.2 Epidemiological-environmental (EpiEnv) Data 42

4.5 Method 45
4.5.1 Anomaly Detection 45
4.5.2 Imputation 48
4.6 Results 49
4.7 Discussion 53
4.8 Conclusion and Future Work 54
4.9 References 56
5. EVALUATING PEDESTRIAN "SAFETY IN NUMBERS" AT SIGNALIZED INTERSECTIONS IN UTAH WITH PEDESTRIAN EXPOSURE DATA FROM TRAFFIC SIGNALS 60
5.1 Abstract 60
5.2 Introduction 60
5.3 Literature Review 61
5.3.1 Factors Affecting Pedestrian Crash Frequency 61
5.3.2 Factors Affecting Pedestrian Crash Severity 62
5.3.3 Safety in Numbers 62
5.3.4 Summary of Literature Review 63
5.4 Data 63
5.4.1 Traffic Signals and Intersection Data 63
5.4.2 Pedestrian Crash Data 63
5.4.3 Pedestrian Exposure Data 65
5.4.4 Descriptive Statistics 65
5.5 Methods 68
5.5.1 Pedestrian Crash Frequency Modeling 68
5.5.2 Pedestrian Crash Severity Modeling 68
5.6 Results 69
5.6.1 Pedestrian Crash Frequency 69
5.6.2 Pedestrian Crash Severity 71
5.7 Discussion 73
5.8 Conclusion 74
5.9 References 74
6. EXPLORING FACTORS AFFECTING PEDESTRIAN CRASH SEVERITY USING TABNET: A DEEP LEARNING APPROACH 77
6.1 Abstract 77
6.2 Introduction 77
6.3 Literature Review 78
6.4 Data and Method 79
6.4.1 Data and Variables 79
6.4.2 Method 83
6.5 Model Results 84
6.6 Model Interpretation and Discussion 86
6.7 Conclusion 95
6.8 References 95

LIST OF TABLES

Table 2.1 Descriptive statistics for land use, built environment, and socio-economic attributes (N = 1,161) 8
Table 2.2 Fit statistics for various numbers of clusters 10
Table 2.3 Summary of hourly/weekly cluster results 11
Table 2.4 Multinomial logit model results of hourly/weekday cluster membership 14
Table 2.5 Multinomial logit model results of monthly cluster membership 18
Table 2.6 Cross-classification of hourly/weekday and monthly clusters (N = 1,446) 19
Table 3.1 Details about data collection in 2019 and 2020 29
Table 3.2 Pedestrians push-button utilization, 2019 vs. 2020, by signal and overall 31
Table 3.3 Pedestrian volume model prediction errors, 2019 vs. 2020, by signal 33
Table 4.1 Descriptive statistics of estimated daily pedestrian volume data, by year 41
Table 4.2 Descriptive statistics for built environment variables 44
Table 4.3 Performance comparison of different anomaly detection methods 50
Table 4.4 Optimum hyperparameters for anomaly detection methods in this study 50
Table 4.5 The performance evaluation of imputation methods for each missing value pattern 51
Table 5.1 Descriptive statistics of variables in the frequency analysis 66
Table 5.2 Descriptive statistics of independent variables in the severity analysis 67
Table 5.3 ZINB model results for pedestrian crash frequency (N = 1,038) 70
Table 5.4 Ordered logit model results for pedestrian crash severity (N = 1,572) 72
Table 6.1 Summary of benefits and limitations of various techniques for pedestrian crash severity analysis 79
Table 6.2 Descriptive statistics of the variables 81
Table 6.3 Optimum hyperparameters of the TabNet models in this study 86
Table 6.4 Performance evaluation metrics 86

LIST OF FIGURES

Figure 2.1 Mean and distribution of pedestrian activity patterns by hourly/weekday cluster 13
Figure 2.2 Expansion accuracy for hourly/weekday clusters 15
Figure 2.3 Means of pedestrian activity patterns by monthly cluster 16
Figure 2.4 Map of signalized intersections and climate divisions 17
Figure 2.5 Expansion accuracy for monthly clusters 18
Figure 2.6 Means of pedestrian activity patterns by hourly/weekday cluster: 1 (N = 871), 2 (N = 278), 3 (N = 302), 4 (N = 188), and 5 (N = 58) 21
Figure 3.1 Map of locations with data collected in 2019 and 2020 28
Figure 3.2 Pedestrian push-button use, 2019 vs. 2020, for signals 1021, 1229, 7184, and 8302 32
Figure 4.1 The dispersion of investigated traffic signals through a tree and point map 41
Figure 4.2 Time series of daily high and low temperature (above) and precipitation (below) around a traffic signal in Cache County 42
Figure 4.3 Heatmap of AQI (yearly-monthly matrix) around a traffic signal in Utah 43
Figure 4.4 The yearly AQI categories in Utah 43
Figure 4.5 The daily COVID-19 case rate per 100,000 in Utah 44
Figure 4.6 Conceptual framework of anomaly detection imputation for pedestrian volume data 46
Figure 4.7 The pedestrian volume data with and without the injected outliers for a sample traffic signal 47
Figure 4.8 The cleaned pedestrian volume data (replace anomalies with missing values) with DBSCAN method in a sample traffic signal in Cache County 50
Figure 4.9 The polar plot of performance evaluation data of imputation methods based on MAE (left) and RMSE (right) 51
Figure 4.10 The results of imputation performed by random forest (a), LSTM (b) and GRU (c) on various sample traffic signal in Utah 52
Figure 5.1 Distributions of dependent variables 64
Figure 5.2 Demonstration of the “safety in numbers” effect for pedestrians at signals 73
Figure 6.1 The spatial configuration of pedestrian crashes 80
Figure 6.2 The structure of the TabNet model for predicting crash severity levels using various EVs 83
Figure 6.3 The importance of each EV for each crash severity class in TabNet model 85
Figure 6.4 The SHAP summary plot for each crash severity class in TabNet model 91
Figure 6.5 The SHAP values, explaining the contribution of EVs to the raw TabNet model output for a specific observation 94

EXECUTIVE SUMMARY

Multimodal transportation planning, traffic monitoring, and traffic safety analyses all require information on how many people walk in various locations throughout the day. Unfortunately, pedestrian volumes are rarely available for these purposes. Luckily, a novel "big data" source of pedestrian information that is relatively ubiquitous in both time and space (24/7 at many locations) is now available: pedestrian push-button actuations recorded in high-definition data logs from traffic signal controllers at signalized intersections. A recent research project studied pedestrian traffic signal data at 90 signalized intersections in Utah, compared push-button presses to video-based ground-truth pedestrian counts, and demonstrated that pedestrian traffic signal data can estimate pedestrian volumes at signalized intersections with reasonable levels of accuracy.

Armed with these novel pedestrian data, this study had three primary objectives: (1) advance pedestrian traffic monitoring by developing methods and models for estimating pedestrian volumes at intersections, based on traffic signal pedestrian push-button data and environmental characteristics; (2) improve pedestrian traffic safety by developing methods and models for analyzing pedestrian crashes at intersections, utilizing traffic signal pedestrian push-button data and environmental characteristics; (3) apply novel machine learning techniques to aid in the advancement of pedestrian traffic monitoring and improvement of pedestrian traffic safety. To achieve these objectives, we conducted a series of five studies, all centered around pedestrian traffic monitoring, pedestrian safety, pedestrian traffic signal data, and/or machine learning methods.

This work advances pedestrian traffic monitoring. The first study takes pedestrian traffic signal data from around 1,500 intersections in Utah, uses time series clustering to identify temporal patterns, and links these patterns with spatial characteristics. The second study investigates how pedestrian push-button press behavior changed at 11 Utah intersections from 2019 to 2020 (during the COVID-19 pandemic), finding no degradation in the accuracy of pedestrian volume estimation models developed before the pandemic when applied to data collected during the pandemic. The third study applies a variety of statistical, machine learning (ML), and deep learning (DL) methods to pedestrian traffic signal data, finding that ML and DL methods—alongside environmental and epidemiological information about temperature, precipitation, air quality, and COVID-19 case rates—can aid in improving the quality and completeness of pedestrian time series data.

This work also improves an understanding of pedestrian traffic safety. The fourth study analyzes pedestrian crash frequency and severity at over 1,000 Utah intersections—utilizing exposure measures obtained from pedestrian traffic signal data—and identifies a "safety in numbers" effect for pedestrians, in which crash rates decrease with increasing pedestrian volumes. The fifth study applies a new deep learning model (TabNet) to the analysis of pedestrian crash severity data, identifying important factors affecting pedestrian injury severity when involved in a traffic collision.

1. INTRODUCTION

Multimodal transportation planning, traffic monitoring, and traffic safety analyses all require information on how many people walk in various locations throughout the day. Pedestrian volumes can be used as outputs when developing pedestrian travel demand forecasting models and as measures of walking exposure when conducting pedestrian safety and health analyses. Unfortunately, traditional data collection methods for levels of pedestrian activity are insufficient for these purposes (FHWA, 2022; Ryus et al., 2014). Manual counts on intersection or street segments are time consuming and often infeasible to conduct over long periods of time. Instruments such as infrared counters can record continuous data on non-motorized trail users, but they are costly to deploy across multiple locations. Video-based pedestrian data collection methods via computer image processing are promising, but video cameras are also costly to install everywhere in a network.

Fortunately, a novel "big data" source of pedestrian information that is relatively ubiquitous in both time and space (available 24/7 at many locations) is now available: pedestrian push-button actuations recorded in high-definition data logs from traffic signal controllers at signalized intersections. Thanks to recent advances (Smaglik et al., 2007), archived and near-real-time pedestrian push-button data can be more easily accessed, such as through the Automated Traffic Signal Performance Measures (ATSPM) system (Day et al., 2016). A recent research project in Utah studied pedestrian traffic signal data at 90 signalized intersections in Utah (Singleton, Runa, & Humagain, 2020). The authors collected ground-truth video-based pedestrian counts, compared them to pedestrian push-button data, and developed factoring and adjustment methods. They demonstrated that pedestrian traffic signal data can estimate pedestrian volumes at signalized intersections with reasonable levels of accuracy (Singleton & Runa, 2021). This advancement opens opportunities to use estimated pedestrian volumes obtained from traffic signal data for a variety of pedestrian planning and safety tasks.

In transportation planning, traffic monitoring is a critical activity that involves collecting data and doing calculations to understand how the use of different transportation modes varies over time and in different places. As described above, pedestrian traffic monitoring lags behind data collection for motor vehicles (Ryus, 2014) because of a lack of permanent stations continuously counting pedestrian volumes. Pedestrian traffic signal data offer an opportunity to improve pedestrian traffic monitoring activities because push-button data are collected 24/7 at many locations. Such data could be used to extract common temporal patterns (by time-of-day, day-of-week, and season) of pedestrian activity, often called "factor groups" in the traffic monitoring literature (FHWA, 2022). These patterns (and factors derived from them) are then used to convert short-term counts into long-term averages, such as annual average daily pedestrian volumes. In this way, pedestrian traffic signal data could be used to advance pedestrian traffic monitoring.

Pedestrian traffic signal data can also overcome a major obstacle to improved pedestrian safety: the lack of pedestrian volume exposure data. Pedestrian safety is a critical current issue given the troubling national trend of increased numbers and shares of pedestrian injuries and fatalities (NHTSA, 2023). Safety predictive methods—safety performance functions, crash modification factors, and crash severity models—usually require the use of exposure data for estimation and application. While motor vehicle volumes are often available, pedestrian volumes rarely are, thus limiting our understanding of and ability to address pedestrian safety issues. Ubiquitous pedestrian signal data can help safety analyses include more robust data on pedestrian exposure, including the existence and magnitude of a potential "safety in numbers" effect (Jacobsen, 2015): as pedestrian volumes increase, walking gets safer and pedestrian crash rates (crashes per exposure) decrease.

At the same time, it is critical to assure the quality of pedestrian volume data being used for such purposes, including for traffic monitoring and safety analysis. First, the estimated pedestrian volumes from traffic signal data rely on empirical relationships established in Utah in one year (2019). Whether these relationships continue to hold over time—especially during and after societal disruptions due to the COVID-19 pandemic—is an important question. Second, all sensor-based longitudinal datasets are susceptible to erroneous or missing data; pedestrian traffic signal data are no exception. It is important to help develop quality control methods to identify potentially erroneous and missing records and, if desired, impute the missing values. With all of these activities (traffic monitoring, traffic safety, data quality), new analytical tools like machine learning methods are starting to be applied that can provide enhanced understanding and predictability across multiple domains, possibly including pedestrian travel.

1.1 Research Objectives

This study has three primary research objectives:

1. Advance pedestrian traffic monitoring by developing methods and models for estimating pedestrian volumes at intersections, based on traffic signal pedestrian push-button data and environmental characteristics.
2. Improve pedestrian traffic safety by developing methods and models for analyzing pedestrian crashes at intersections, utilizing traffic signal pedestrian push-button data and environmental characteristics.
3. Apply novel machine learning techniques to aid in the advancement of pedestrian traffic monitoring and improvement of pedestrian traffic safety.

1.2 Research Approach and Overview

To achieve these objectives, we conducted a series of five studies, all centered around pedestrian traffic monitoring, pedestrian safety, pedestrian traffic signal data, and/or machine learning methods. Taken as a whole, they help achieve the three research objectives. A summary of each study (one per chapter) is contained in the following paragraphs.

Chapter 2, "Advances in pedestrian travel monitoring: Temporal patterns and spatial characteristics using pedestrian push-button data from Utah traffic signals" takes pedestrian traffic signal data from around 1,500 intersections in Utah, uses time series clustering to identify temporal patterns, and links these patterns with spatial characteristics. This work demonstrates the great utility of pedestrian traffic signal data for advancing understanding of pedestrian behaviors and improving the transportation planning practice of pedestrian traffic monitoring.

Chapter 3, "Impacts of the COVID-19 pandemic on pedestrian push-button utilization and pedestrian volume model accuracy in Utah," investigates how pedestrian push-button press behavior changed at 11 Utah intersections from 2019 to 2020 during the COVID-19 pandemic. This work finds no degradation in the accuracy of pedestrian volume estimation models developed before the pandemic when applied to data collected during the pandemic, suggesting that the prior models (Singleton & Runa, 2021) may still work acceptably well for estimating pedestrian volumes from traffic signal data during/after COVID.

Chapter 4, "Imputing time series pedestrian volume data with consideration of epidemiological-environmental (EpiEnv) variables," applies a variety of statistical, machine learning, and deep learning methods to pedestrian traffic signal data, thus informing an important topic in transportation planning and traffic monitoring: detecting anomalies and imputing missing data. Results suggest that ML and DL methods—alongside environmental and epidemiological information about temperature, precipitation, air

quality, and COVID-19 case rates—can aid in improving the quality and completeness of pedestrian time series data.

Chapter 5, "Evaluating pedestrian "safety in numbers" at signalized intersections in Utah with pedestrian exposure data from traffic signals," analyzes pedestrian crash frequency and severity at over 1,000 Utah intersections, utilizing exposure measures obtained from pedestrian traffic signal data. Importantly, this work identifies a "safety in numbers" effect for pedestrians, in which crash rates decrease with increasing pedestrian volumes.

Chapter 6, "Exploring factors affecting pedestrian crash severity using TabNet: A deep learning approach," applies a new deep learning model (TabNet) to the analysis of pedestrian crash severity data. In addition to identifying important factors affecting pedestrian injury severity when involved in a traffic collision, linear and nonlinear results from this "black box "model are interpreted to aid in understandability. This work highlights emerging advanced computational methods that are helping to better understand and address pedestrian safety issues.

Note that Chapters 2, 3, and 4 have been previously published as peer-reviewed manuscripts in academic journals. They are being reprinted here with permission.

1.3 References

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. https://trid.trb.org/View/1418795

Federal Highway Administration (FHWA). (2022). *Traffic Monitoring Guide*. https://www.fhwa.dot.gov/policyinformation/tmguide/

Jacobsen, P. L. (2015). "Safety in numbers: More walkers and bicyclists, safer walking and bicycling." *Injury Prevention, 21*(4), 271-275. https://doi.org/10.1136/ip.9.3.205rep

National Highway Traffic Safety Administration (NHTSA). (2023). *Pedestrian safety: Prevent Pedestrian Crashes* (accessed 22 July 2023). https://www.nhtsa.gov/road-safety/pedestrian-safety

Ryus, P., Ferguson, E., Laustsen, K. M., Schneider, R. J., Proulx, F. R., Hull, T., & Miranda-Moreno, L. (2014). *Guidebook on Pedestrian and Bicycle Volume Data Collection* (NCHRP Report 797). Transportation Research Board. https://doi.org/10.17226/22223

Singleton, P. A., & Runa, F. (2021). "Pedestrian traffic signal data accurately estimates pedestrian crossing volumes." *Transportation Research Record: Journal of the Transportation Research Board, 2675*(6), 429-440. https://doi.org/10.1177/0361198121994126

Singleton, P. A., Runa, F., & Humagain, P. (2020). *Utilizing Archived Traffic Signal Performance Measures for Pedestrian Planning & Analysis*. Utah Department of Transportation. https://rosap.ntl.bts.gov/view/dot/54924

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*(1), 97–106. https://doi.org/10.3141/2035-11

2. ADVANCES IN PEDESTRIAN TRAVEL MONITORING: TEMPORAL PATTERNS AND SPATIAL CHARACTERISTICS USING PEDESTRIAN PUSH-BUTTON DATA FROM UTAH TRAFFIC SIGNALS

- Humagain, P., & Singleton, P. A. (2021). "Advances in pedestrian travel monitoring: Temporal patterns and spatial characteristics using pedestrian push-button data from Utah traffic signals." *Journal of Transport and Land Use, 14*(1), 1341-1360. https://doi.org/10.5198/jtlu.2021.2112

2.1 Abstract

In this study, we advanced pedestrian travel monitoring using a novel data source: pedestrian push-button presses obtained from archived traffic signal controller logs at more than 1,500 signalized intersections in Utah over one year. The purposes of the study were to: (1) quantify pedestrian activity patterns, (2) create factor groups and expansion/adjustment factors from these temporal patterns, and (3) explore relationships between patterns and spatial characteristics. Using empirical clustering, we classified signals into five groups based on normalized hourly/weekly counts (each hour's proportion of weekly totals, or the inverse of the expansion factors), and three clusters with similar monthly adjustment factors. We also used multinomial logit models to identify spatial characteristics (land use, built environment, socio-economic characteristics, and climatic regions) associated with different temporal patterns. For example, we found that signals near schools were much more likely to have bimodal daily peak hours and lower pedestrian activity during out-of-school months. Despite these good results, our hourly/weekday patterns differed less than in past research, highlighting the limits of existing infrastructure for capturing all kinds of activity patterns. Nevertheless, we demonstrate that signals with push-button data are a useful supplement to existing permanent counters within a broader pedestrian traffic monitoring program.

2.2 Introduction

Despite recent advances and interest from researchers and practitioners, pedestrian monitoring and data collection remains incomplete and insufficient, especially compared with motorized data collection. In practice, there are two methods for counting pedestrian activity at intersections: manual and automatic approaches (Greene-Roesel et al., 2008; FHWA, 2016). Manual counts involve collecting pedestrian volumes (by an observer) in real-time in situ or later using video recordings. Although manual counts are accurate and advantageous for making modal distinctions (walking vs. cycling) and determining directional flows (left, straight, or right), the count accuracy depends upon the observer's characteristics (e.g., attentiveness) (Diogenes et al., 2007). More critically, manual counts are infeasible over longer time periods because of the need for direct human supervision. Another method based on automated instruments—such as microwave, ultrasonic, or infrared—are feasible for longer-duration counts, thus are ideal for identifying variations in pedestrian activity over time (Bu et al., 2007; Green-Roesel et al., 2008). However, automated counts can be susceptible to adverse weather or crowding, are expensive to install, and sometimes require periodic validation from manual counts (FHWA, 2016; Ryus et al., 2014).

2.2.1 Pedestrian Push-button Data to Measure Pedestrian Activity

Alternatively, one novel source of pedestrian data is from pedestrian push-buttons at signalized intersections. Many (but not all) traffic signals require people walking who want to cross an approach to press a pedestrian push-button to request (actuate) the walk phase. Given readily available hardware and software, each pedestrian push-button press event can be timestamped, logged (Smaglik et al., 2007; Sturdevant et al., 2012), archived, and made available (for example) through the Automated Traffic Signal Performance Measures (ATSPM) system (ATKINS, 2016; Day et al., 2014, 2016). Such high-resolution traffic signal controller log data are relatively ubiquitous in both time and space (available 24/7 at many intersections), making them a potentially rich source of information about pedestrian activity levels. Some of the limitations of existing methods—such as requirement of manual labor or upfront costs for installation of automated counters—and the lack of pedestrian data could be addressed by the use of this novel pedestrian data source.

Until recently, few studies investigated the use of pedestrian data from traffic signal controller logs to estimate walking activity. Day et al. (2011) analyzed data on pedestrian actuations per hour at one signalized intersection in Indiana over an 18-month period, finding impacts of time-of-day, day-of-week, weather and other seasonal effects, special events, and a change in pedestrian phase configuration on pedestrian actuations. Similarly, Blanc et al. (2015) and Kothuri et al. (2017) conducted studies of pedestrian activity at one intersection in Oregon that had actuated pedestrian crossings (using push-button detection) for all four crosswalks. The two Oregon studies used video data to manually count pedestrians, which they then compared to pedestrian actuations for each crosswalk, usually finding correlations of around 0.80 or greater. Recently, a large-scale validation study of pedestrian push-button use and walking activity at signalized intersections was conducted in Utah by Singleton, Runa, and Humagain (2020). The authors compared hourly pedestrian signal activity metrics derived from push-button presses against observed pedestrian counts—obtained from manual counts of over 20,000 hours of videos recorded in 2019 for 320 crosswalks at 90 signalized intersections—using simple nonlinear regression models. The models' estimated pedestrian volumes were strongly correlated with observed pedestrian crossing volumes (0.84) and had a low mean absolute error (3.0 pedestrians per hour) (Singleton, Runa, & Humagain, 2020; Singleton & Runa, 2021). Overall, these studies demonstrate that traffic signal data can be used to estimate pedestrian crossing volumes and monitor levels of pedestrian activity at intersections. We utilize pedestrian data from traffic signals in this chapter.

2.2.2 Applications of Continuously-measured Pedestrian Data

Temporally rich pedestrian data—measured continuously over time—has many applications. Quantifying and understanding the characteristics of pedestrian activity patterns in different spatial locations over time can assist planners and/or researchers in any (or all) of the following ways:

1. *Planning*: Pedestrian data can help planners prioritize pedestrian infrastructure investments in specific areas and predict the impacts of new transportation or urban development projects on walking.
2. *Safety*: Pedestrian safety analysis could use temporal patterns of pedestrian activity to better quantify risks related to exposure to traffic at crossings.
3. *Traffic operations*: Hourly distributions of pedestrian activity by location can assist with optimizing traffic signal timing for pedestrian delay or safety, as well as scheduling/permitting maintenance or construction work for areas and times with low pedestrian activity.
4. *Traffic monitoring*: Automated pedestrian counters cannot be deployed in all areas, so long-term count data are used to develop expansion factors that translate short-duration (e.g., manual) counts into estimated average annual daily pedestrian volumes, information which is useful for all of the activities listed above.

This chapter contributes to the fourth application, traffic monitoring.

2.2.3 Pedestrian Expansion Factors

To develop expansion factors, locations with similar pedestrian activity patterns (quantified either daily or weekly) are often grouped together into "factor groups" (Medury et al., 2019; FHWA, 2016; Ryus et al., 2017), each with a unique set of expansion factors. Short-duration pedestrian volume measurements (e.g., manual peak-period or daily counts) are then multiplied by the expansion factors—for the specific factor group to which that short-duration count location best belongs—in order to estimate long-term average pedestrian volumes more precisely (FHWA, 2016; Ryus et al., 2017). In current practice, there are two common approaches to constructing factor groups of multiple locations with similar pedestrian activity patterns. The first method is the *land use classification* approach (Medury et al., 2019), which involves classifying locations based upon their surrounding land use characteristics, under the assumption that locations with similar land uses will generate similar pedestrian activity patterns. Studies implementing this approach have identified distinct patterns for commercial areas, employment areas, university areas, trail areas, and others. (Schneider et al., 2009; Medury et al., 2019). The second method is the data-driven *empirical clustering* approach, which essentially groups locations based upon their pedestrian activity patterns, referred to as clusters. In short, the clustering algorithm works by minimizing differences in patterns within each cluster while simultaneously maximizing differences in patterns between clusters. Miranda-Moreno and Lahti (2013) classified bicycle traffic patterns into four distinct groups as utilitarian, mixed-utilitarian, mixed-recreational, and recreational. Griswold et al. (2018) compared land use and empirical clustering approach and concluded that both approaches provided better results than a "single factor" method (where all locations are combined into single factor group). No matter the approach, the process of constructing factor groups is limited by the number and variety of locations with long-duration pedestrian count data.

2.2.4 Research Objectives

In this study, we aim to overcome limitations surrounding the lack of long-term automated pedestrian count data for traffic monitoring through the use of pedestrian push-button information from hundreds of sites in one U.S. state. Specifically, we utilize high-resolution data collected from traffic signal controller logs at more than 1,500 signalized intersections throughout Utah—available from the Utah Department of Transportation (UDOT)'s ATSPM system—to investigate the temporal patterns of pedestrian activity and develop expansion factors and factor groups that relate to spatial characteristics. As such, the objectives of this chapter are threefold, to:

1. Quantify and understand pedestrian activity patterns at signalized intersections, using continuous, archived data from pedestrian push-buttons at more than 1,500 signalized intersections in Utah.
2. Calculate time-of-day/day-of-week expansion factors and create factor groups based on empirical clustering of pedestrian activity patterns at signalized intersections, while accounting for seasonal variation.
3. Explore relationships between pedestrian factor groups and land use, built environment, and socio-economic neighborhood characteristics.

2.3 Data and Methods

In this section, we present the data and an overview of the analysis methods. First, we describe calculating the pedestrian activity metrics for two temporal dimensions—hourly/weekday patterns and monthly (seasonal) patterns—from traffic signal controller log data, as well as assembling data on spatial characteristics from various sources. Second, we explain the analysis methods employed, including

empirical clustering, regression modeling, and expansion/adjustment factor accuracy. The data and scripts used in this chapter are publicly available (Singleton, Runa, & Humagain, 2021).

2.3.1 Data Sources and Preparation

2.3.1.1 Pedestrian traffic signal data

Traffic signal controller log data from most of the over 2,000 signalized intersections in Utah were collected from UDOT's ATSPM system (UDOT, 2020) for one full year (July 2017 through June 2018). In total, data from 1,697 signals with pedestrian push-buttons were usable. The remainder of the signals either did not have pedestrian push-buttons (either in isolated rural/industrial locations or in the heart of downtown Salt Lake City) or were missing data for a significant portion of the year.

In order to prepare time series pedestrian datasets for clustering, a suitable metric that defines intersection-level "pedestrian activity" from traffic signal data was required. For this purpose, we relied upon the research by Singleton, Runa, and Humagain (2020) that validated pedestrian push-button data against observed pedestrian counts using over 20,000 crossing-hours of observations in Utah. That research, using regression models and various fit statistics, determined that a new pedestrian activity metric of imputed pedestrian calls registered, "45B" was the best predictor of actual pedestrian crossing volumes in many cases. More details about this validation and modeling process can be found elsewhere (Singleton, Runa, & Humagain, 2020; Singleton & Runa, 2021). Specifically, the 45B pedestrian activity metric is defined as:

- For each pedestrian phase, in a time-ordered sequence of traffic signal controller events with just events {0, 21, 90}, the number of 90 events immediately preceded by a 0 or 21 event, where:
    - Event 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 21, Pedestrian Begin Walk: This event occurs with the activation of the walk indication for a particular phase.
    - Event 90, Pedestrian Detector On: This event occurs when the signal from the pedestrian push-button is activated, after any delay or extension is processed, for a particular pedestrian detector channel.

In simple terms, the pedestrian activity metric 45B (imputed pedestrian calls registered) counts the number of times the walk signal appeared as a result of a pedestrian push-button press.

We analyzed two types of temporal patterns in pedestrian activity: hourly/weekday patterns and monthly (or seasonal) patterns. For *hourly/weekday patterns*, we did the following for each intersection i: First, we calculated the pedestrian activity metric (45B) for all pedestrian phases over the entire year, removing any hours with missing data (i.e., due to communication outages or maintenance work). Second, we averaged these year-long hourly observations into 168 values $v_{i,t,d}$, one for each of the unique hour-of-day t and day-of-week d combinations (e.g., 4–5 p.m. Mondays). Third, we calculated normalized counts $\bar{\upsilon}_{i,t,d}$ according to the following equation (Ryus et al., 2014; Griswold et al., 2018):

$$\bar{\upsilon}_{i,t,d} = \frac{v_{i,t,d}}{\sum_{t=1}^{24}\sum_{d=1}^{7} v_{i,t,d}} \quad (1)$$

These normalized counts $\bar{\upsilon}_{i,t,d}$ are really the average hourly counts as a proportion of total average weekly counts of pedestrian activity at each signal, or essentially the (inverse of) hour-to-week expansion factors specific to each intersection. By averaging across the entire year, this process mitigates some of the effects of temporal variation caused by special events, abnormal weather, or other unusual occurrences. These intersection-specific normalized counts (inverse expansion factors) of pedestrian

activity (45B) were used as the data input into the empirical clustering analysis for hourly/weekday patterns.

For *monthly (seasonal) patterns*, we did the following for each intersection *i*: First, we took the whole-year hourly pedestrian activity (45B) dataset from the first step of the previous paragraph and summed the hourly values to generate 365 daily totals. Second, for each month *m*, we calculated the monthly average daily volume ($d_{i,m}$); we also calculated the overall annual average daily volume ($y_{i,y}$). Third, we calculated the 12 monthly adjustment factors ($m_{i,m}$) according to the following equation:

$$m_{i,m} = \frac{d_{i,m}}{y_{i,y}} \tag{2}$$

These intersection-specific monthly adjustment factors were used as the data input into the empirical clustering analysis for monthly (seasonal) patterns.

2.3.1.2 Spatial data

To explore relationships between temporal patterns in pedestrian activity and spatial characteristics, we assembled land use, built environment, and socio-economic attributes for the area surrounding each signalized intersection. Specifically, measures were calculated using quarter-mile network buffers. Data came from various sources, including population and employment data from the 2013-2017 American Community Survey and the 2017 Longitudinal Employer-Household Dynamics dataset for Census block groups, as well as 2019 land use and transportation data from the Utah Automated Geographic Reference Center. (See Singleton, Park, and Lee [2021] for details on these data.) Due to a lack of data, this information was available for only 1,161 signals. Descriptive statistics for these attributes are presented in Table 2.1.

Table 2.1 Descriptive statistics for land use, built environment, and socio-economic attributes (N = 1,161)

Attribute	*Mean*	*SD*
Land use attributes		
Residential land use (%)	31.264	22.006
Commercial land use (%)	30.756	19.360
Industrial land use (%)	2.007	8.190
Schools (#)	0.344	0.665
Places of worship (#)	0.593	0.850
Parks (acres)	1.537	3.622
Built environment attributes		
Population density (1,000/mi^2)	5.263	2.923
Employment density (1,000/mi^2)	8.365	12.888
Intersection density (#/mi^2)	105.228	46.549
4-way intersections (%)	30.988	20.036
Transit stops (#)	43.065	23.491
Socio-economic attributes		
Vehicle ownership (#, mean)	1.646	0.418
Household size (#, mean)	2.972	0.850
Household income ($1,000)	57.655	20.152

2.3.2 Analysis Methods

For each type of temporal pattern (hourly/weekday, monthly), we conducted a series of analyses: (1) empirical cluster analysis to identify clusters of signals with similar temporal patterns; (2) multinomial logit regression modeling to understand spatial factors associated with temporal clusters; and (3) calculating the accuracy of applying the expansion/adjustment factors.

2.3.2.1 Identifying temporal patterns using empirical cluster analysis

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 in this section.

Because the objective of this study focuses on grouping intersections based on similar hourly/weekly patterns, we used a structural-based (dis)similarity measure—temporal correlation (CORT)—since it allows us to compare the relative trajectories of normalized counts across intersections. Basically, CORT measures the proximity of temporal variation between two time series, which aligns better with our objective than other conventional distance measures such as Euclidean distance, which works with the difference in magnitude between data points (Montero & Vilar, 2014). The equation for CORT is:

$$d_{CORT}(F_i, F_j) = \frac{\sum_{t=1}^{T-1}(F_{i(t+1)} - F_{it})(F_{j(t+1)} - F_{jt})}{\sqrt{\sum_{t=1}^{T-1}(F_{i(t+1)} - F_{it})^2}\sqrt{\sum_{t=1}^{T-1}(F_{j(t+1)} - F_{jt})^2}} \quad (3)$$

where F_i and F_j represent two time series *i* and *j*, measured over T time points *t*.

In terms of clustering algorithms, we applied a k-means algorithm, which basically has the objective of minimizing differences within each cluster and maximizing differences between other clusters. The k-means algorithm is found to be computationally efficient and can group intersections based on subtle nuances in temporal patterns of pedestrian activity.

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. Common tools to assist in the selection of the number of clusters include the following:

- Calinski-Harabasz (CH) criterion: This is the ratio of between-cluster variation to within-cluster variation, so a higher value reflects distinct clusters.
- Sum of squared differences (SSD): This calculates the sum of squared differences between each observation's values and the mean values of the cluster to which the observation belongs. Smaller differences indicate more homogenous clusters.
- Average silhouette width (ASW): This measures the similarity (ranging from −1 to +1) of patterns of observations within each cluster (cohesion) compared with observations of other clusters (separation). A more positive value implies that observations are well matched within clusters and poorly matched to neighboring clusters.

By performing k-means clustering for various numbers of clusters (2 to 7 for hourly/weekday, 2 to 5 for monthly), we computed the various fit statics and also visualized the patterns of resulting clusters. Generally, Table 2.2 shows that CH, SSD, and ASW values decreased with increased numbers of clusters, although not exclusively so. More clusters generally means an increase in both the distinctiveness and compactness of clusters; hence, the optimum number of clusters can be the number at which more/fewer clusters provides neither a significant improvement nor degradation in the fit statistics. Hence, the optimum number of hourly/weekday clusters was determined to be five, as it provides reasonable fit statistics and a relatively lower decrease in SSD (compared with six clusters). Similarly, the optimum

number of monthly clusters was determined to be three, as it provides a reasonable tradeoff between low SSD and higher ASW.

Table 2.2 Fit statistics for various numbers of clusters

	Hourly/weekday clusters			*Monthly (seasonal) clusters*		
# clusters	*CH*	*SSD*	*ASW*	*CH*	*SSD*	*ASW*
2	1,424.71	4,715.61	0.72	700.38	668.42	0.54
3	1,082.93	3,226.85	0.54	**585.93**	**539.92**	**0.40**
4	1,052.84	3,612.17	0.54	539.65	475.59	0.30
5	**985.61**	**2,285.66**	**0.57**	502.74	307.44	0.27
6	973.47	1,970.04	0.31			
7	995.75	1,520.18	0.28			

2.3.2.2 Predicting clusters with spatial factors using multinomial logit regression models

In order to explain which signals belonged in each cluster on temporal patterns of pedestrian activity, we performed multinomial logit regression using the spatial factors shown in Table 2.1 (land use, built environment, and socio-economic attributes) as explanatory variables. Variables with statistically significant coefficients indicate spatial characteristics associated with signals having different temporal patterns in pedestrian activity.

2.3.2.3 Assessing the accuracy of expansion/adjustment factors

When applying the cluster results and expansion/adjustment factors to convert short-duration counts to longer-term average volumes, there will be some discrepancy even with the same data due to using expansion/adjustment factors based on the cluster mean values. For hourly/weekday factors, this expansion accuracy is expressed as the absolute percentage error of the expanded weekly counts at a location relative to the average expanded weekly counts of the cluster to which that location belongs. As presented in Griswold et al. (2018) and Medury et al. (2019), the expansion accuracy $\in_{i,t,d}^{c}$ for hourly/weekday expansion factors is given by the following equation:

$$\in_{i,t,d}^{c} = \left| \frac{\bar{\upsilon}_{i,t,d}}{\gamma_{i,t,d}^{c}} \right| \times 100 \qquad (4)$$

where: $\bar{\upsilon}_{i,t,d}$ is the normalized count and $\gamma_{i,t,d}^{c}$ is the applicable expansion factor for cluster c, location i, and time period t. A similar equation applies to the accuracy of monthly adjustment factors.

2.4 Results

2.4.1 Hourly/weekday Patterns

2.4.1.1 Results of hourly/weekly clusters

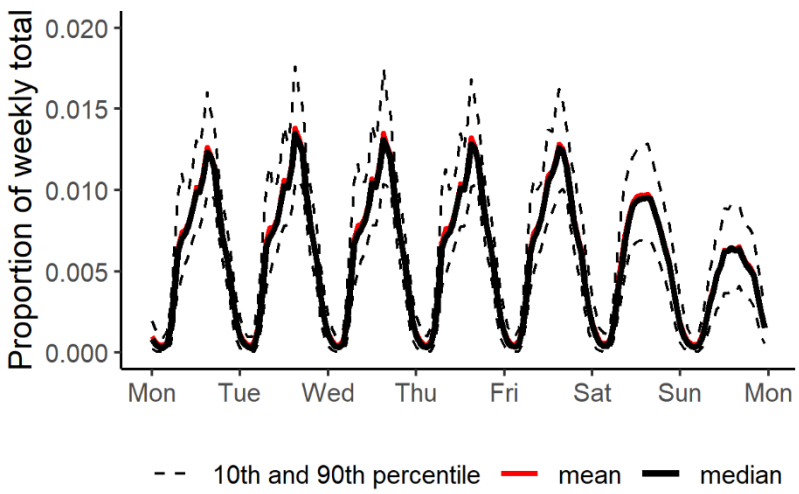
To recap, we used the CORT (dis)similarity measure and the k-means algorithm to classify the normalized counts of pedestrian activity at 1,697 signalized intersections into five clusters. The cluster analysis results are summarized in Table 2.3 and the text below, and the mean and distributions of the hourly/weekly patterns are depicted in Figure 2.1.

Table 2.3 Summary of hourly/weekly cluster results

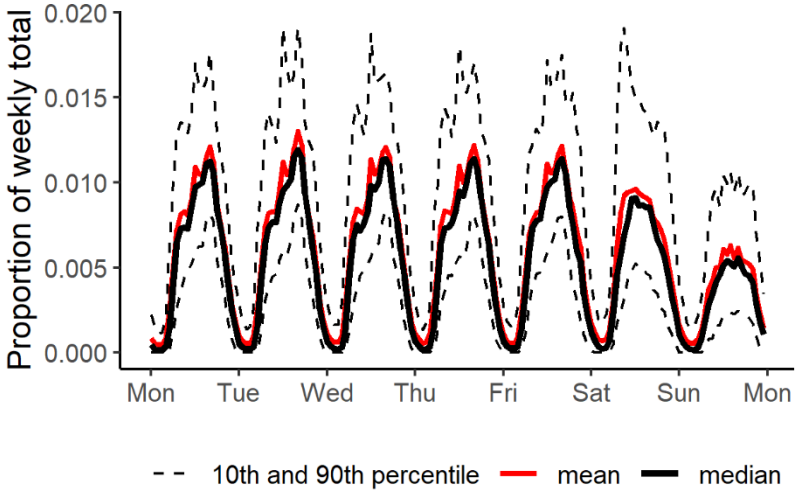
Pattern	*Cluster (#, %)*	*Visual characteristics*
Uniform	1 (871, 51.3)	Evening peak, increase from morning to evening, weekdays > weekends, peak hour volume ~1-1.5% of weekly volume
	2 (278, 16.4)	Evening peak, increase from morning to evening, weekdays > weekends, peak hour volume ~1-1.5% of weekly volume
Bimodal	3 (302, 17.8)	Morning and evening peaks, evening > morning, weekdays > weekends, peak hourly volume ~1.5-2% of weekly volume
	4 (188, 11.1)	Morning and evening peaks, evening > morning, weekdays > weekends, peak hourly volume ~1.5-2% of weekly volume
	5 (58, 3.4)	Morning and evening peaks, evening > morning, weekdays > weekends, peak hourly volume ~2-2.5 of weekly volume

Overall, the hourly/weekday clusters can be classified into two general patterns of pedestrian activity: (a) unimodal, with one (usually evening) peak hour that is approximately 1% to 1.5% of the weekly total, and (b) bimodal, with two distinct peak hours (usually evening is greater than morning) and where the (usually evening) peak hour is approximately 1.5% to 2% of the weekly total. Besides these general observations, the clusters themselves show some (albeit more minor) differences. Unimodal clusters 1 and 2 are slightly differentiated in their daytime vs. evening patterns: for cluster 1, the pattern is somewhat more uniform (or smooth) than for cluster 2, and the mean is slightly more peaked. The bimodal clusters 3, 4, and 5 are distinguished by the magnitude of their peaks—cluster 5's peaks are more than 2% of the weekly total, whereas peaks for clusters 3 and 4 are 1.5% to 2%—and somewhat by the difference between the morning and evening peaks (difference: cluster 5 > cluster 4 > cluster 3).

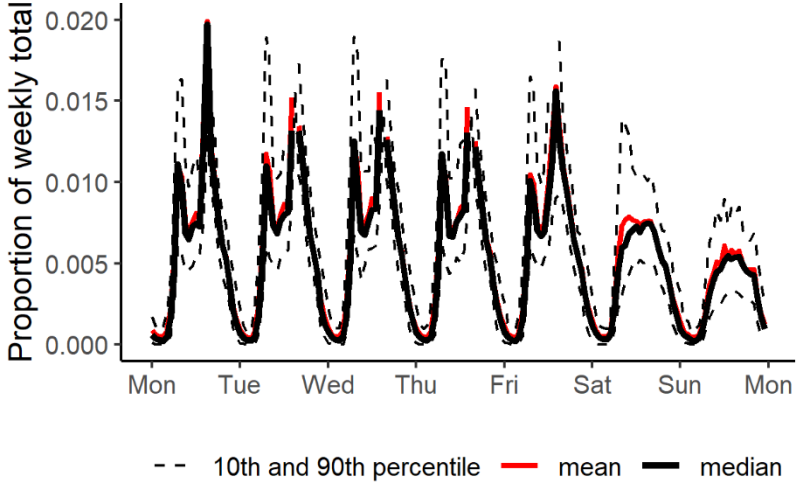
There were also some similarities between all the hourly/weekday clusters. Unsurprisingly, pedestrian activity was highest during daytime and evening hours, with most intersections recording little to no activity overnight. Peak pedestrian hours were more common in the afternoon and early evening than in the morning. Weekend pedestrian activity (especially on Sundays) was lower than on weekdays, but often without a clear single peak hour (usually midday). Tuesdays often had the largest peak hour of pedestrian activity, while Mondays and Fridays tended to have slightly lower peaks than other weekdays (although Mondays had the highest peaks for clusters 3 and 4).



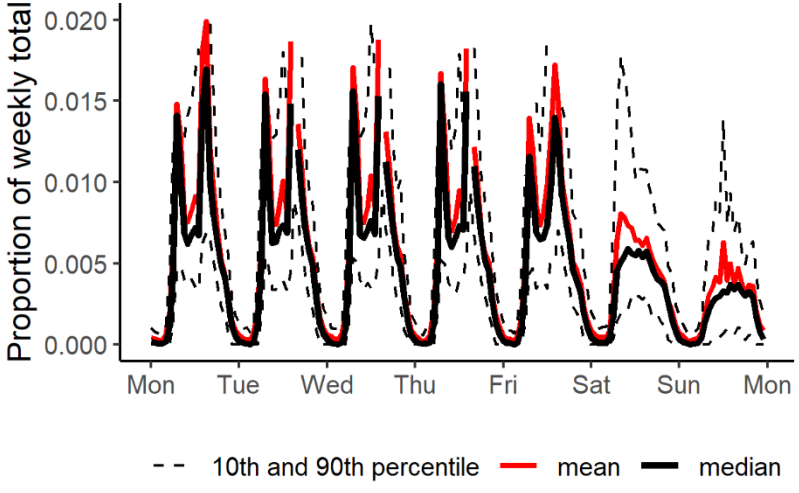
i. Unimodal – Cluster 1



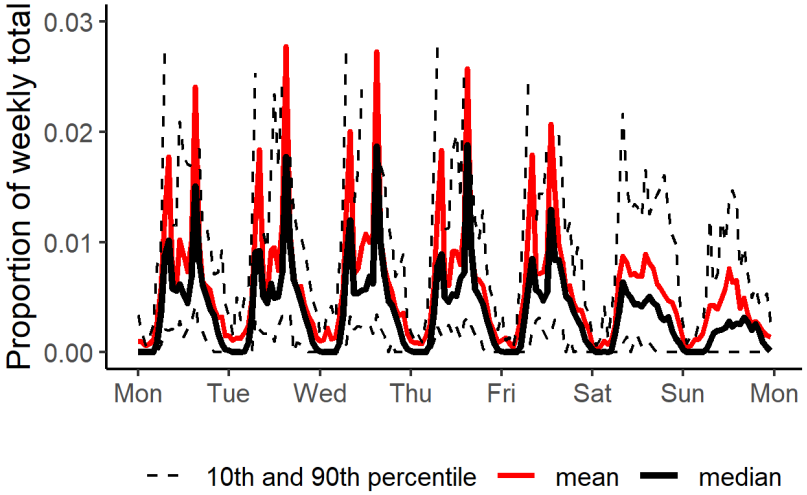
ii. Unimodal – Cluster 2



iii. Bimodal – Cluster 3



iv. Bimodal – Cluster 4



v. Bimodal – Cluster 5

Figure 2.1 Mean and distribution of pedestrian activity patterns by hourly/weekday cluster

2.4.1.2 Spatial Factors Affecting Hourly/weekly Patterns

Some past studies have investigated the influential role of land use, the built environment, and socio-economic characteristics in shaping temporal patterns of pedestrian activity across locations (e.g., Hankey et al., 2012; Medury et al., 2019; Schneider et al., 2009). The presence of offices, schools/colleges, and different land use characteristics surrounding count locations are found to influence the hourly/weekday temporal patterns. For instance, locations near schools displayed multiple peaks on weekdays and relatively lower pedestrian activity during weekends, whereas recreational trails had higher activity during weekday evenings and relatively higher pedestrian activity during weekends (Medury et al., 2019). We add to this literature using our larger dataset of over 1,000 signalized intersections in Utah.

To understand the relationships between pedestrian activity patterns and spatial characteristics, we estimated a multinomial logit model on 1,161 signalized intersections with such data, where membership in an hourly/weekly cluster (1 to 5) was the dependent variable and spatial characteristics were the independent variables. Results are presented in Table 2.4 and described below.

Table 2.4 Multinomial logit model results of hourly/weekday cluster membership

	Cluster-specific coefficients (ref. = 1)			
Variable	*2*	*3*	*4*	*5*
Intercept	0.751	1.409	0.006	-0.576
Population density (1,000/mi^2)	**-0.234**	**-0.240**	**-0.373**	
Residential land use (%)	**-0.035**			
Commercial land use (%)	**-0.046**	**-0.053**	**-0.055**	
Industrial land use (%)		**-0.034**		**0.058**
Intersection density (#/mi^2)		**-0.005**	**-0.010**	**-0.019**
4-way intersections (%)		**-0.017**	**-0.018**	**-0.039**
Schools (#)		**0.515**	**1.039**	**1.166**
Vehicle ownership (#, mean)	**-0.555**			
Household size (#, mean)	**0.275**		*0.325*	
Household income ($1,000)	**0.021**	**0.012**	**0.016**	
McFadden pseudo-R^2	0.170			
Sample size (N)	1,161			

Statistical significance: **bold** for $p < 0.05$, *italics* for $p < 0.10$, not shown for $p > 0.10$.

The model results help to explain why we see some of the differences in the pedestrian activity patterns across clusters. Notably, the bimodal patterns (multiple peaks) of clusters 3, 4 and 5 is partially explained by the result that these locations were much more likely to located within a quarter-mile walking distance of one or more schools, indicated by the significant positive coefficients for number of schools. Signals were also more likely to have a bimodal pattern in areas with less street network connectivity, as shown by the significant negative coefficients for intersection density and percentage of four-way intersections. Based on the results for population density, percentage of commercial land use, and household income, signals were more likely to have the smooth unimodal pattern of cluster 1 when they were in areas with greater population density, more commercial land uses, and lower household incomes. Looking at differences between unimodal signals, belonging to cluster 1 was more likely in neighborhoods with more residential land uses, greater vehicle ownership, and smaller household sizes.

2.4.1.3 Expansion factor accuracy

Figure 2.2 displays the average expansion accuracy for each hour in the week by hourly/weekly cluster. Overall, expansion accuracy is greater (lower error) for clusters and during times with higher pedestrian activity levels. Average error is less than 75% to 100% for all clusters when counts are taken during

daytime hours, but greater than 75% (and as much as 150%) when overnight counts are expanded. Expansion errors for daytime counts at signals in clusters 1 and 3 (the largest clusters) are around 25% or less, suggesting that only a few hours of counts at these locations may be enough to accurately estimate longer-term pedestrian volumes. Conversely, counts taken at signals in the smallest cluster (5) may need to be of a longer duration in order to produce similarly accurate estimates of pedestrian activity.

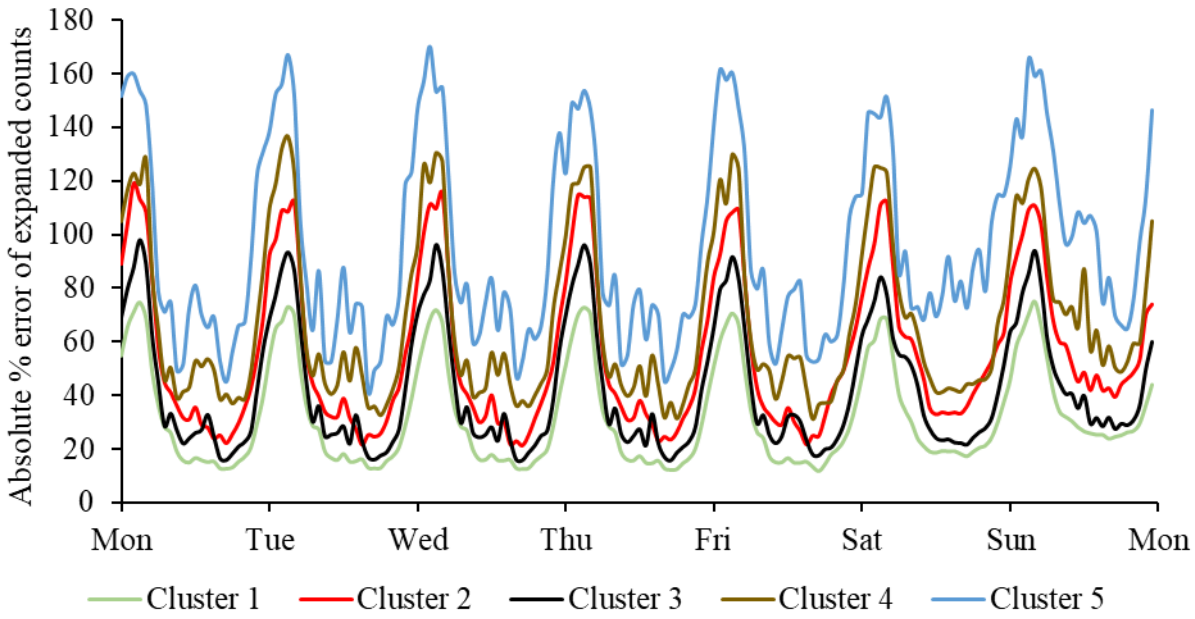


Figure 2.2 Expansion accuracy for hourly/weekday clusters

2.4.2 Monthly Patterns

2.4.2.1 Seasonal (monthly) Clusters

The normalized counts (inverse expansion factors) for the hourly/weekday clusters shown in Figure 2.1 depict average hourly and weekday pedestrian activity patterns expressed as a proportion of weekly totals. Homogeneity within those clusters may obscure other sources of temporal variations in pedestrian activity patterns between locations, such as those differences due to seasonal variation. In fact, factoring processes in traffic monitoring to convert short-duration counts to annual average daily volumes require seasonal adjustment factors as well.

Therefore, we performed a similar empirical clustering process to generate monthly clusters of similar seasonal pedestrian activity patterns. In traffic monitoring, such seasonal variation in activity at intersections is addressed during the calculation of annual volume by using monthly adjustment factors. After calculating the monthly adjustment factors as described in section 2.3.1.1, some of the intersections were removed due to unusually high factors above 3.0, which could have resulted from missing data or technical errors, resulting in total of 1,446 intersections. To recap, the optimum number of monthly (seasonal) clusters was determined to be three (based on fit statistics and visualization).

The mean values of the adjustment factors for the three monthly clusters are shown in Figure 2.3. On average, the largest cluster A (1,076, 74.4%) has the least variation in pedestrian activity patterns from month to month, peaking in the summer months (especially June, but with a slight decrease in July) and bottoming out in the winter months (especially December). Higher pedestrian activity during the summer

is warranted because of pleasant weather (warm and dry) throughout most of Utah. Similarly, cold temperature and snow conditions result in lower levels of pedestrian activity during Utah winter months. The other two clusters, B and C, show greater (and different) variations on this trend. Signals in cluster B (111, 7.7%) have something of a more extreme version of cluster A's pattern, with even higher relative volumes in the summer (from June through September) and lower volumes in winter (especially February). The pattern suggests that pedestrian activity for signals in cluster B is more sensitive to weather. Conversely, cluster C (259, 17.9%) shows a distinctly different pattern, peaking in September and having less than average pedestrian activity during the summer months (June through August). Signals in cluster C could be near schools or universities, which are usually not in session during these three summer months. These hunches about the reasons motivating seasonal variations in pedestrian activity could be confirmed through comparisons with spatial characteristics, as presented in the following section.

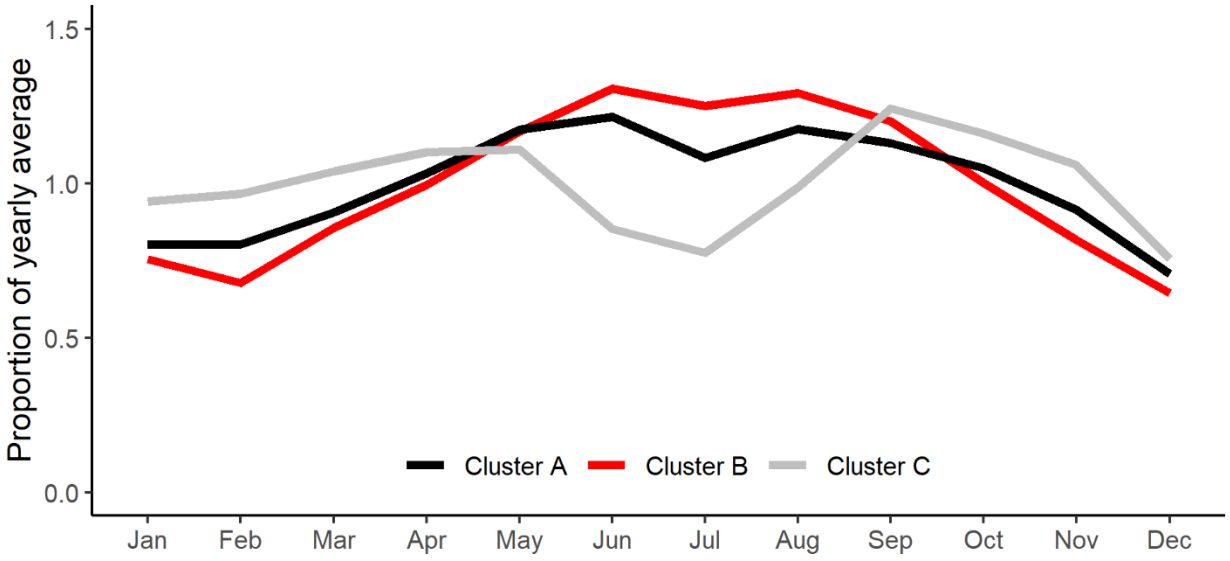


Figure 2.3 Means of pedestrian activity patterns by monthly cluster

2.4.2.2 Spatial factors affecting seasonal (monthly) patterns

As described earlier, seasonal variation in pedestrian activity patterns is mostly influenced by climatic conditions (i.e., snow, temperature, rainfall) (Runa, 2020). To help explain these seasonal variations in Utah, we estimated another multinomial logit model on 1,161 signalized intersections, this time where the dependent variable was membership in a monthly cluster (A, B, or C). In addition to the same land use, built environment, and socio-economic characteristics as used previously, we added climatic division classifications from the National Climate Data Center. The assumption is that signals in the same climatic division experience similar weather patterns throughout the year. As shown in Figure 2.4, Utah contains seven climatic divisions, although most signals in our dataset are located in the North Central region (also known as the Wasatch Front), with some in the Northern Mountains and Dixie regions. (We did not have spatial characteristics data for signals in the other climatic divisions.) Note that although many of the signals appear to lie along the border of the North Central and Northern Mountains divisions, almost all are truly located in the urbanized valleys of Utah's Wasatch Front (such as in Salt Lake City). These signals experience weather patterns that are much more similar to each other than they are with higher-elevation locations in the Northern Mountains (such as in Park City).

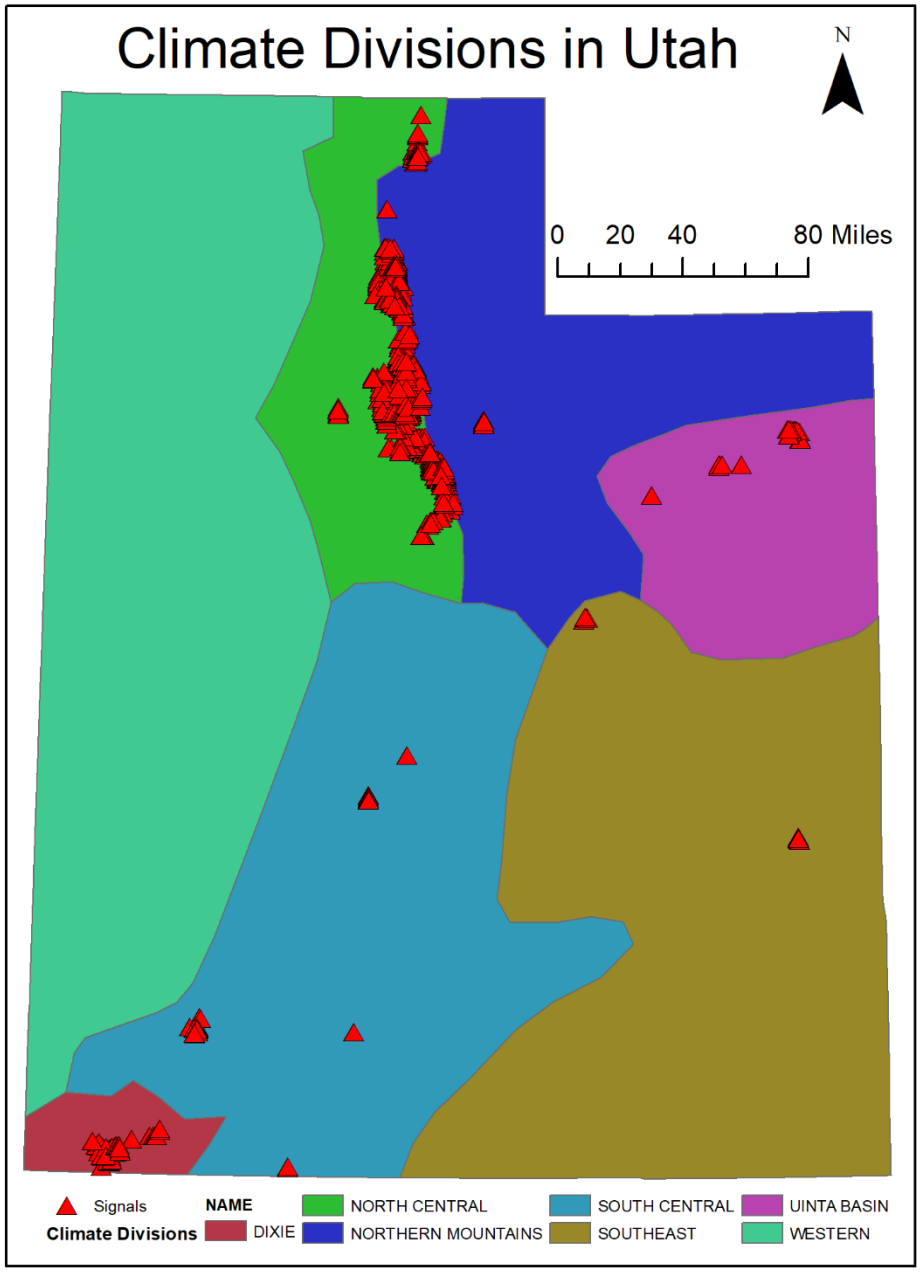


Figure 2.4 Map of signalized intersections and climate divisions

The results of the multinomial logit model for monthly clusters are shown in Table 2.5; the largest cluster A is the reference alternative. Signals were less likely to be in cluster A and more likely to be in clusters B or C if they were in areas with greater employment density, lower traffic volumes, and larger household sizes. Cluster B was associated with less commercial land use, greater intersection density, and fewer places of worship. Signals in the Dixie climatic region were much more likely to belong to cluster B or C than to A. Typically, this region experiences hotter summers and milder winters than other areas of Utah, which could explain the greater monthly variation for cluster B and the summer trough of cluster C. Also, signals near schools were much more likely to be in cluster C, which supports our hypothesis about a lack of school attendance being an explanation for lower pedestrian activity levels in summer months.

Table 2.5 Multinomial logit model results of monthly cluster membership

	Cluster-specific coefficients (ref. = A)	
Variables	*B*	*C*
Intercept	*-2.233*	**-3.521**
Climate division: Dixie (ref. = North Central)	**2.852**	**4.666**
Employment density (1,000/mi^2)	**0.029**	**0.046**
Commercial land uses (%)	**-0.031**	
4-way intersections (%)	**0.019**	
Schools (#)		**0.734**
Places of worship (#)	*-0.399*	
Vehicle ownership (#, mean)	**-1.001**	**-0.880**
Household size (#, mean)	**0.722**	**1.053**
McFadden pseudo-R^2	0.186	
Sample Size (N)	1,161	

Statistical significance: **bold** for $p < 0.05$, *italics* for $p < 0.10$, not shown for $p > 0.10$.

2.4.2.3 Adjustment factor accuracy

The average error (expansion/adjustment accuracy) for the three monthly clusters is shown in Figure 2.5. Accuracy is remarkably good (around 10% error) for the largest cluster A all year round, but even cluster C has reasonably good expansion/adjustment accuracy (10% to 20% error) throughout most of the year. The smallest cluster B shows the potential for more error (30% to 40%). As with the hourly/weekday clusters, accuracy for the monthly clusters tends to be worse during months with lower levels of pedestrian activity (December for all clusters, July for cluster C).

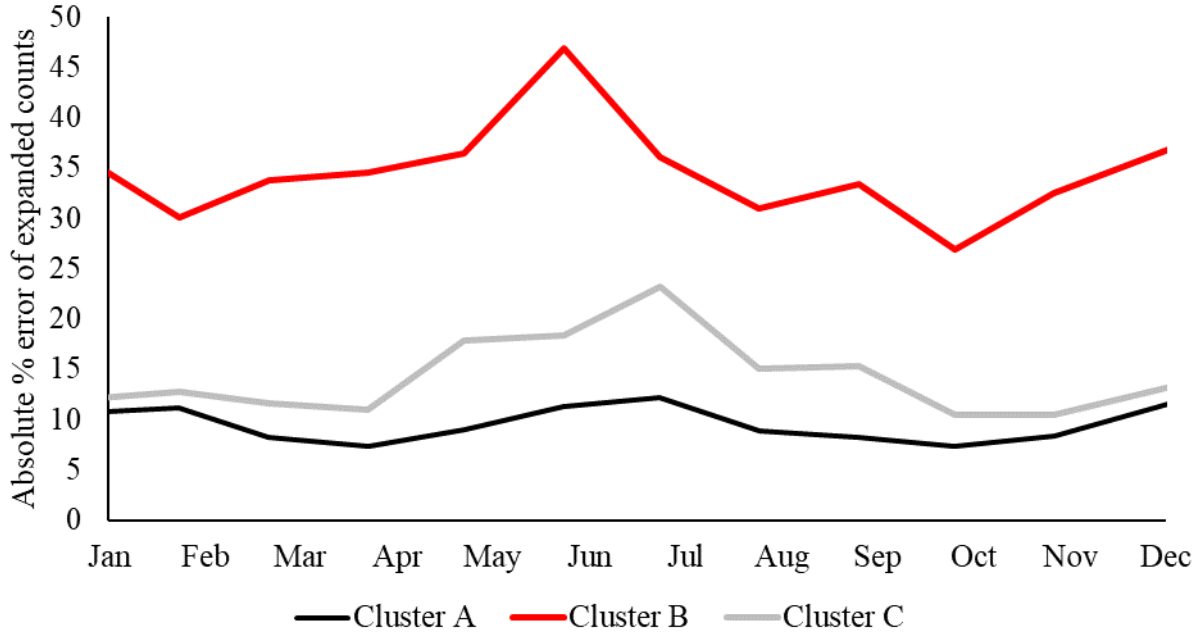


Figure 2.5 Expansion accuracy for monthly clusters

2.4.3 Cross-classification of Hourly/weekday and Monthly Clusters

As hourly/weekday and monthly pedestrian activity pattern trends differ in nature (due to variations in climatic conditions and some spatial characteristics), it is useful to examine the frequency of signals in each of the hourly/weekday clusters that belonged to a particular seasonal trend produced by the monthly clusters. Therefore, we cross-tabulated the number of intersections that belonged to each combination of hourly/weekday and monthly clusters. Table 2.6 indicates that most unimodal hourly/weekday clusters (1 and 2) pertain to the more uniform seasonal variations of cluster A. However, most of the intersections that belonged to the bimodal hourly/weekday clusters 4 and 5 (those with high relative peak pedestrian activity levels) were grouped under monthly cluster C, where pedestrian activity dips during summer months. This makes sense, given that a bimodal daily pattern and a summer lull in pedestrian activity are both indicative of school-driven pedestrian activity patterns.

Table 2.6 Cross-classification of hourly/weekday and monthly clusters (N = 1,446)

Hourly/weekday clusters	*Monthly Clusters*		
	A	*B*	*C*
1	661	38	85
2	152	36	39
3	199	11	50
4	53	19	70
5	11	7	15

2.5 Discussion and Conclusions

In this study, we provided an empirical clustering approach to grouping locations with similar long-term pedestrian activity patterns using pedestrian push-button data from over 1,500 signalized intersections in Utah. After calculating the proxy measure of pedestrian activity (imputed pedestrian calls registered), we performed cluster analysis to classify signals based on the normalized hourly/weekly counts (each hour's proportion of weekly totals, or the inverse of the expansion factors) and account for seasonal variation using monthly adjustment factors. We also used multinomial logit models to identify spatial and climatic characteristics (land use, built environment, and socio-economic characteristics, as well as climatic regions) that help predict and explain which locations see different temporal patterns in pedestrian activity levels. Finally, we assessed the accuracy of applying the expansion/adjustment factors.

The ultimate objective of this work was to investigate the temporal patterns of pedestrian activity and develop expansion/adjustment factors and factor groups that relate to spatial characteristics. Utilizing the hourly/weekday and monthly clusters that we developed, each with an average temporal pattern, one can expand a short-duration count (for a specific hour, weekday, and month) through multiplication and/or division to get an estimate of the long-term average pedestrian volume at a particular location. In the process of achieving this aim, our chapter made several contributions to travel monitoring for pedestrian travel.

- Most notably, we utilized a much greater quantity of pedestrian data than has been possible to examine before; specifically, one year of data from 1,697 signalized intersections throughout Utah. This larger sample size allowed us to examine more nuanced differences in pedestrian activity patterns and have the power to identify significant associations with spatial characteristics.
- Our use of separate hourly/weekday and monthly clusters allowed us to distinguish the influences of time-of-day and day-of-week from seasonal variations.
- We also considered the expansion/adjustment accuracy for our factor groups (clusters), which has implications for the selection of times-of-day and durations for short-term pedestrian counts.

In the remaining section of this chapter, we discuss key findings, their applications for understanding pedestrian behavior and monitoring pedestrian traffic, and study limitations.

Overall, the clustering of average hourly proportions of weekday counts revealed common hourly/weekday pedestrian activity patterns at most signalized intersections, as shown for the five clusters altogether in Figure 2.6. All clusters saw their highest pedestrian volumes during weekday daytime hours, with peak pedestrian volumes in the afternoon or early evening, lower volumes on weekends, and slightly lower volumes on Mondays and Fridays. In fact, the differences between clusters were more nuanced. Three clusters (3, 4, and 5) showed bimodal morning and evening peaks, with the other two clusters (1 and 2) having just a single evening peak. The weekday peak hours varied from 1% to 2.5% of weekly totals, depending on the cluster. The fact that the clustering algorithm picked up even these small differences in temporal patterns highlights the utility of an empirical data-driven approach to constructing pedestrian factor groups.

On the other hand, the small differences between locations shown here suggests that a coarser factor grouping might not result in significantly inferior count expansion results. Compared with past research defining pedestrian or non-motorized count factor groups (Schneider et al., 2009; Miranda-Moreno & Lahti, 2013; Medury et al., 2019), the hourly/weekday factor groups (clusters) we identified are not as distinct. This is likely due to several factors, most notably the limitation of our source data. Research relying on permanent non-motorized counters can cover a wider range of temporal use pattern types—commuting vs. recreational vs. mixed—precisely because the locations where counters are deployed were selected to capture a variety of behaviors and uses. In this study, we relied upon existing infrastructure (traffic signals) in locations that were not chosen with pedestrian count programs in mind. In other words, location selection was exogenous to our study purpose. Since most traffic signals are located in areas with higher traffic volumes (such as along arterials) or where walking for utilitarian/transportation purposes is expected, they may not be able to detect the full variety of pedestrian temporal patterns that exist, such as along trails or in recreational areas.

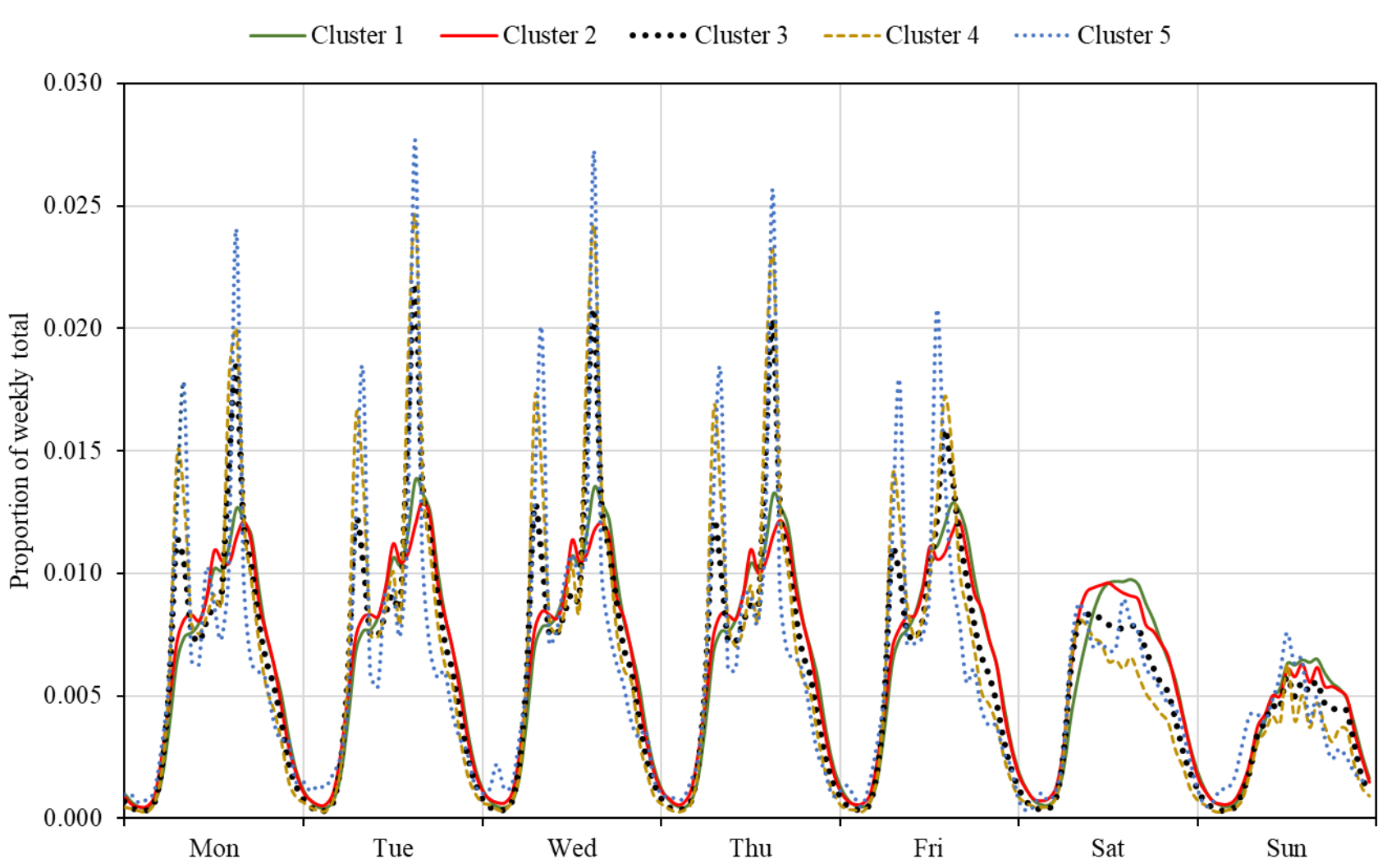


Figure 2.6 Means of pedestrian activity patterns by hourly/weekday cluster: 1 ($N = 871$), 2 ($N = 278$), 3 ($N = 302$), 4 ($N = 188$), and 5 ($N = 58$)

Despite the less noticeable differences between hourly/weekday pedestrian activity patterns in different clusters, the model illuminated land use, built environment, and socio-economic differences that help to explain these pattern variations. Notably, having schools nearby significantly increased the chances of a signal having a bimodal pedestrian activity pattern, coincident with the morning and afternoon/early evening time periods bracketing the school day. Several other built environment attributes—population density, residential and commercial land uses, connected street networks—were linked to signals belonging to cluster 1, with unimodal, smoother, and less varied weekday pedestrian activity patterns. This information about spatial characteristics—along with insight into expansion accuracy—aids planners and pedestrian traffic monitoring program managers by suggesting the types of locations where different expansion/adjustment factors are needed or even where longer or shorter short-duration counts are needed to provide accurate pedestrian volume estimates.

Our clustering of monthly adjustment factors showed more significant differences in the seasonal patterns of pedestrian activity across intersections (see Figure 2.3) than were found for hourly/weekday patterns. Specifically, we captured both the general seasonal trend (higher pedestrian activity in summer, lower activity during winter months) and also trends specific to certain locations, such as the drop in pedestrian activity during out-of-school months from June through August. By linking these monthly groupings to spatial characteristics, we confirmed that this latter specific pattern occurred more often near schools and universities. We also demonstrated that pedestrian activity was more sensitive to weather in certain regions (southwestern Utah) with higher summer temperatures and mild winters.

Another finding of this study is about expansion factor accuracy. Confirming past research, our results suggest that the expansion accuracy is cyclical in nature, with higher errors during low-volume overnight hours and greater accuracy during daytime. This implies that manual counts should be conducted at intersections during the daytime and longer counts may be beneficial, especially during off-peak hours and at locations with more variability (i.e., cluster 5). The expansion accuracy from empirical clustering should be more accurate than the "single factor" approach of having just one factor group, as shown by previous studies (Griswold et al., 2018; Medury et al., 2019).

2.5.1 Limitations

The study is not without additional limitations. First, the pedestrian activity metrics derived from pedestrian push-button data do not provide the actual pedestrian volume, and may contain errors, because of imperfect correlation and nonlinearities. However, correlation between push-button data and volumes is high (Singleton, Runa, & Humagain, 2020; Singleton & Runa, 2021), and we suspect the benefits of being able to analyze data from hundreds if not thousands of locations outweigh the inaccuracies of the source data. Second, the clustering approach produced five clusters, but in reality there were only two distinct groupings (unimodal and bimodal). However, the tradeoff between cluster fit and the number of clusters is difficult to control for, and empirical clustering does offer the benefit of identifying smaller differences that may be obscured using a different approach. Third, the sample size decreased somewhat when incorporating spatial characteristics, which might cause bias or lack of generalizability of the spatial analysis results. Fourth, the relatively coarse nature of the climatic divisions used in the study could not exactly pinpoint what causes the variation in seasonal patterns (i.e., snow, rainfall, wind), and there may be additional influential weather variations within each climatic division. Fifth, the use of empirical clustering and Utah-specific climate zones may limit the generalizability of these temporal patterns to areas outside of Utah. Nevertheless, a similar process may be useful for developing pedestrian expansion and adjustment factors in other states. In summary, despite these limitations, we have demonstrated that traffic signals with pedestrian push-button data are a very useful supplement to—but not a complete replacement for—existing permanent counters within a broader pedestrian traffic monitoring program.

2.5.2 Future Research

There are many opportunities to refine this research or extend it in new directions. Future research could look at using alternative data sources (such as the Google Places API) to calculate the attributes near intersections, which could provide more detailed insights into specific land uses or place types than some of the aggregate metrics used in our study. Correlating traffic signal-based pedestrian activity levels with weather is a potential fruitful area of inquiry. In this regard, more fine-grained data—temperature, humidity, snow, air quality—collected from nearby weather stations could be assembled and correlated with pedestrian activity patterns at intersections (Runa & Singleton, 2021). If there are common patterns in how pedestrian activity changes when it, for example, snows or rains, it may be possible to develop expansion/adjustment factors that work for short-duration counts conducted during mildly inclement weather. Additionally, another promising area for future research could be the investigation of the effects of major events such as concerts or sporting events on changes in pedestrian activity compared with normal days. More research could also be done using these pedestrian traffic signal data to inform the duration and timing of short-term pedestrian counts. For example, one could extend this study to determine the average expansion/adjustment accuracy of different count durations (anywhere from one hour to one week) in an attempt to find the optimum tradeoff between cost and accuracy. Finally, research also could look at using pedestrian push-button data to calculate and compare other types of traffic monitoring count data expansion factors, such as hour-to-year or day-to-year, as these have been suggested as potentially more accurate alternatives for estimating annual average volumes (Medury et al., 2019).

2.6 References

ATKINS. (2016). *Automated Traffic Signal Performance Measures Reporting Details*. 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

Bu, F., Greene-Roesel, R., Diogenes, M. C., & Ragland, D. R. (2007). "Estimating Pedestrian Accident Exposure: Automated Pedestrian Counting Devices Report." UC Berkeley Traffic Safety Center. https://escholarship.org/uc/item/0p27154n

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." 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. https://trid.trb.org/view/1418795

Diogenes, M. C., Greene-Roesel, R., Arnold, L. S., & Ragland, D. R. (2007). "Pedestrian counting methods at intersections: A comparative study." *Transportation Research Record: Journal of the Transportation Research Board, 2002*(1), 26–30. https://doi.org/10.3141/2002-04

Federal Highway Administration (FHWA). (2016). *Traffic Monitoring Guide*. U.S. Department of Transportation. https://www.fhwa.dot.gov/policyinformation/tmguide/

Greene-Roesel, R., Diogenes, M. C., Ragland, D. R., & Lindau, L. A. (2008). "Effectiveness of a commercially available automated pedestrian counting device in urban environments: Comparison with manual counts." Presented at the 87th Annual Meeting of the Transportation Research Board, Washington, DC. https://escholarship.org/uc/item/2n83w1q8

Griswold, J.B., Medury, A., Schneider, R.J. and Grembek, O., 2018. "Comparison of pedestrian count expansion methods: Land use groups versus empirical clusters." *Transportation Research Record: Journal of the Transportation Research Board, 2672*(43), 87–97. https://doi.org/10.1177/0361198118793006

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

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

Medury, A., Griswold, J. B., Huang, L. & Grembek, O. (2019). "Pedestrian count expansion methods: bridging the gap between land use groups and empirical clusters." *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(5), 720-730. https://doi.org/10.1177/0361198119838266

Miranda-Moreno, L. F., & Lahti, A. C. (2013). "Temporal trends and the effect of weather on pedestrian volumes: A case study of Montreal, Canada." *Transportation Research Part D: Transport and Environment, 22*, 54–59. https://doi.org/10.1016/j.trd.2013.02.008

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

Runa, F. (2020). "The effect of weather on pedestrian activity at signalized intersections in Utah" (master's thesis). Utah State University. https://doi.org/10.26076/cdbb-1171

Runa, F., & Singleton, P. A. (2021). "Assessing the impacts of weather on pedestrian signal activity at 49 signalized intersections in Northern Utah." *Transportation Research Record: Journal of the Transportation Research Board, 2675*(6), 406-419. https://doi.org/10.1177/0361198121994111

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). Transportation Research Board. https://doi.org/10.17226/23429

Ryus, P., Butsick, A., Proulx, F. R., Schneider, R. J., & Hull, T. (2017). "Methods and technologies for pedestrian and bicycle volume data collection: Phase 2" (NCHRP Web-Only Document 205). Transportation Research Board. https://doi.org/10.17226/24732

Schneider, R. J., Arnold, L. S., & Ragland, D. R. (2009). "Methodology for counting pedestrians at intersections: Use of automated counters to extrapolate weekly volumes from short manual counts." *Transportation Research Record: Journal of the Transportation Research Board, 2140*(1), 1–12. https://doi.org/10.3141/2140-01

Singleton, P. A., Park, K., & Lee, D. H. (2021). "Varying influences of the built environment on daily and hourly pedestrian crossing volumes at signalized intersections estimated from traffic signal controller event data." *Journal of Transport Geography, 93*, 103067. https://doi.org/10.1016/j.jtrangeo.2021.103067

Singleton, P. A., & Runa, F. (2021). "Pedestrian traffic signal data accurately estimates pedestrian crossing volumes." *Transportation Research Record: Journal of the Transportation Research Board, 2675*(6), 429-440. https://doi.org/10.1177/0361198121994126

Singleton, P. A., Runa, F., & Humagain, P. (2020). "Utilizing archived traffic signal performance measures for pedestrian planning & analysis." Utah Department of Transportation. https://rosap.ntl.bts.gov/view/dot/54924

Singleton, P., Runa, F., & Humagain, P. (2021). "singletonpa/ped-signal-data: Update with temporal patterns" (Dataset). Zenodo. https://doi.org/10.5281/zenodo.4759088

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*(1), 97–106. https://doi.org/10.3141/2035-11

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.” Purdue University. http://doi.org/10.4231/K4RN35SH

Utah Department of Transportation (UDOT). 2020. “Automated Traffic Signal Performance Measures.” https://udottraffic.utah.gov/ATSPM/

3. IMPACTS OF THE COVID-19 PANDEMIC ON PEDESTRIAN PUSH-BUTTON UTILIZATION AND PEDESTRIAN VOLUME MODEL ACCURACY IN UTAH

This chapter is the accepted manuscript of an article published by the Transportation Research Board in the journal *Transportation Research Record: Journal of the Transportation Research Board*. It is reprinted here in accordance with Sage's Author Archiving and Re-Use Guidelines. To cite, please use this reference:

- Runa, F., & Singleton, P. A. (2023). "Impacts of the COVID-19 pandemic on pedestrian push-button utilization and pedestrian volume model accuracy in Utah." *Transportation Research Record: Journal of the Transportation Research Board, 2677*(4), 494-502. https://doi.org/10.1177/03611981221089935

3.1 Abstract

This work investigated the impacts of COVID-19 on pedestrian behavior, answering two research questions using pedestrian push-button data from Utah traffic signals: How did push-button utilization change during the early pandemic due to concerns over disease spread through high-touch surfaces? How did the accuracy of pedestrian volume estimation models (developed pre-COVID based on push-button traffic signal data) change during the early pandemic? To answer these questions, we first recorded videos, counted pedestrians, and collected push-button data from traffic signal controllers at 11 intersections in Utah in 2019 and 2020. We then compared changes in push-button presses per pedestrian (to measure utilization), as well as model prediction errors (to measure accuracy), between the two years. Our first hypothesis of decreased push-button utilization was partially supported. The changes in utilization at most (seven) signals were not statistically significant; yet the aggregate results (using 10 of 11 signals) saw a decrease from 2.1 to 1.5 presses per person. Our second hypothesis of no degradation of model accuracy was supported. There was not statistically significant change in accuracy when aggregating across nine signals, and the models were actually more accurate in 2020 for the other two signals. Overall, we conclude that COVID-19 did not significantly deter people from using push-buttons at most signals in Utah, and that the pedestrian volume estimation methods developed in 2019 likely do not need to be re-calibrated to work for COVID conditions. This information may be useful for public health actions, signal operations, and pedestrian planning.

3.2 Introduction

The outbreak of the coronavirus disease COVID-19 first started in Wuhan, China, in December 2019. In March 2020, the World Health Organization (WHO) announced COVID-19 as a global pandemic after it spread rapidly all over the world, including in the United States. To slow the spread of the virus, different countries took various public health actions. Many U.S. states and communities implemented stay-at-home orders or recommendations, schools and restaurants were closed or limited, working from home became the norm in some fields, and many public events and large gatherings were canceled. Mandates or recommendations also often included social distancing (6 ft or 2 m), face coverings (masks), and frequent hand washing and surface cleaning.

The COVID-19 pandemic led to major changes in travel patterns around the world and across multiple modes (De Vos, 2020; Beck & Hensher, 2020; Jenelius & Cebecauer, 2020; Shamshiripour et al., 2020, Shakibaei et al., 2021). Travel restrictions also appear to have resulted in significant changes in walking activity in Utah (Singleton Transportation Lab, 2020). There is a need and desire to accurately monitor traffic patterns, including pedestrian activity, in order to inform agencies in their traffic management and

other operational and planning decisions. It is also of scientific interest to know how pedestrian behavior changed in response to COVID-19 concerns. In this chapter, we focus on one specific area of COVID-19 influences on pedestrian behavior: pedestrians' utilization of push-buttons at traffic signals (button-press behavior), and corresponding impacts on the accuracy of pedestrian volume estimation from push-button traffic signal data.

3.2.1 Background and Research Questions

Early in the pandemic, fears over the virus spread; the fear of contracting COVID-19 by interacting with high-touch public surfaces—including pedestrian push-buttons at traffic signals—led some transportation agencies to eliminate the need to press the push-button in order to get a walk indication (Combs, 2020), switching to a signal timing technique called pedestrian recall. For example, Salt Lake City and the Utah Department of Transportation (UDOT) placed several dozen signals in downtown Salt Lake City on pedestrian recall from April through June 2020 (Singleton et al., 2023). These actions were in response to fears that could have manifested in different pedestrian behaviors when interacting with traffic signals that had not been switched to pedestrian recall. Specifically, people may have been less willing to press the pedestrian push-button in times and locations during community spread of COVID-19. However, these actions were taken without knowing whether pedestrian push-button utilization or button-press behavior actually changed. In fact, later in the pandemic, studies showed that infected surfaces (especially those exposed to sunlight) were not a leading cause of COVID-19 spread.

- **Research Question 1**: How did the utilization of pedestrian push-buttons at traffic signals change during the early months of the COVID-19 pandemic?
- *Hypothesis 1*: Pedestrians were slightly less likely to press pedestrian push-buttons due to concerns about COVID-19.

Recent research in Utah has investigated the use of pedestrian push-button data from traffic signals for pedestrian traffic monitoring and pedestrian volume estimation (Singleton et al., 2020). UDOT has been a leader in developing the Automated Traffic Signal Performance Measures (ATSPM) system (UDOT, 2021; ATKINS, 2016), which allows access to high-resolution traffic signal controller event logs (Smaglik et al., 2007), including information about pedestrian push-button presses (Sturdevant et al., 2012). Work by Singleton and Runa (2021) in 2019 recorded more than 22,000 crossing-hours of video and collected observed counts of over 170,000 pedestrians at 90 signals throughout Utah. Comparisons of pedestrian counts to pedestrian signal data (including pedestrian actuations and time-filtered pedestrian push-button presses) used simple quadratic or piecewise linear regression models applied to different situations (e.g., pedestrian recall or not, short vs. long cycle lengths). Overall, the model-estimated pedestrian crossing volumes had a low error (±3.0 pedestrians per hour) and were strongly correlated (0.84) with observed volumes (Singleton & Runa, 2021).

The application of these models allows for the estimation of pedestrian volumes (directly from traffic signal data) at around 1,500 signals throughout Utah, providing information that is useful for pedestrian planning and safety analysis efforts (Singleton, Park, & Lee, 2021; Singleton, Mekker, & Islam, 2021). However, these models rely on empirically derived relationships from 2019 about pedestrian behavior at signals, specifically, the utilization of pedestrian push-buttons. Any change in pedestrian push-button press behavior due to COVID-19 might yield less accurate volume estimates and require a recalibration of these pedestrian volume estimation methods.

- **Research Question 2**: How did the accuracy of pedestrian volume estimation models based on traffic signal data (developed pre-COVID) change during the early months of the COVID-19 pandemic?

- *Hypothesis 2*: Pedestrian push-button utilization (button-press behavior) did not change enough to degrade the accuracy of the pedestrian volume estimation models.

3.3 Data and Methods

In order to answer our research questions, we had to first collect pedestrian data from recorded videos, then assemble pedestrian push-button data from traffic signals, and finally analyze push-button utilization and the accuracy of the pedestrian volume estimation models.

3.3.1 Pedestrian Data Collection

Observed pedestrian data were obtained from recorded videos at different signals in Utah. In 2019, we collected pedestrian volume data at 90 signals (Singleton et al., 2020). For 2020, we collected data at 11 signals (see Figure 3.1 and Table 3.1), where we had the most 2019 data, in order to increase the likelihood that any differences in pedestrian behavior were not due to random chance. These locations also captured a range of estimated traffic volumes—annual average daily pedestrian (AADP) crossing volumes and entering motor vehicle volumes as measured by annual average daily traffic (AADT)—as well as different regions and urban contexts. (None of these locations were placed under continuous pedestrian recall by Salt Lake City or UDOT.) For each location in each year, we used UDOT traffic cameras to record more than 200 crossing-hours of video. We then watched the videos and tabulated the number of pedestrians (walking, running, on a skateboard, or in a wheelchair) using each crossing in each hour.

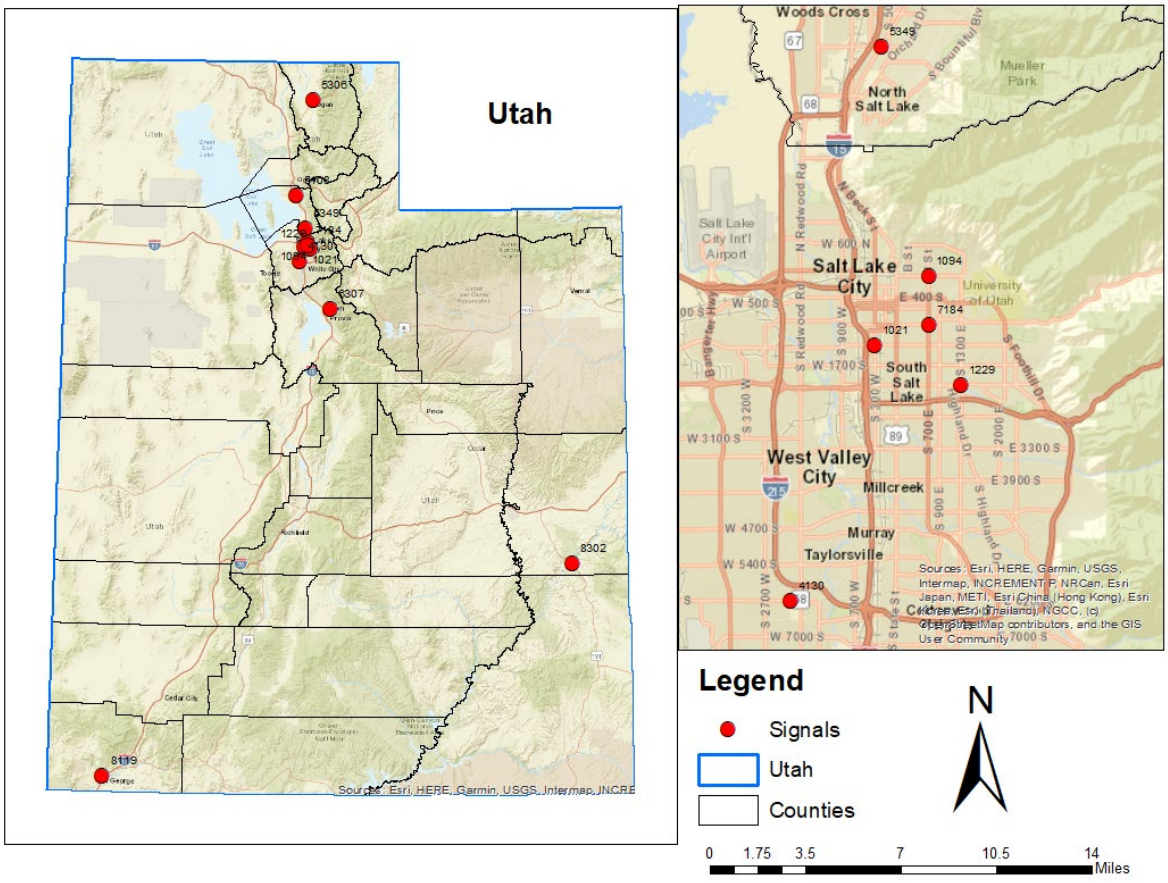


Figure 3.1 Map of locations with data collected in 2019 and 2020

Table 3.1 Details about data collection in 2019 and 2020

		Crossing-hours		*Months*		*Crossing AADP* [a]		*Entering AADT* [b]	
Signal	*Location*	*2019*	*2020*	*2019*	*2020*	*2019*	*2020*	*2019*	*2020*
1021	1300 S & 300 W, Salt Lake City	489	310	03	06, 07	1,871	1,578	37,100	33,200
1094	S Temple & 700 E, Salt Lake City	294	212	04	06, 07	305	196	41,800	37,300
1229	2100 S & 1300 E, Salt Lake City	274	253	03	06, 07	1,489	1,274	62,800	56,100
4130	6200 S & Margray Dr, Taylorsville	471	215	02, 07, 08, 11	05	57	86	24,800	22,200
5108	Antelope Dr & Hill Field Rd, Layton	416	326	03, 10	06	470	318	44,500	39,800
5306	400 N & Main St, Logan	498	251	01, 02, 09, 11	04, 05	291	225	50,900	44,300
5349	2600 S & US-89, Bountiful	508	288	07, 11	05	380	277	47,300	42,300
6307	800 N & Palisades Dr, Orem	281	353	07	06, 07	83	137	37,200	32,400
7184	900 S & 700 E, Salt Lake City	558	694	01, 02, 08, 11	04, 05, 06	822	790	54,700	44,200
8119	St. George Blvd & 400 E, St. George	730	368	01, 02, 08, 11	04, 05	102	84	27,100	26,900
8302	Center St & Main St, Moab	789	358	02, 03, 06, 07, 10	05	6,146	5,163	18,500	17,600

[a] Estimated AADP values were calculated by applying the modeling methods developed by Singleton and Runa (2021) to a full year of pedestrian push-button data. Pedestrian crossing volumes across the 11 signals decreased an average 16% from 2019 to 2020.

[b] Estimated AADT values were obtained from products of UDOT's traffic monitoring program. Motor vehicle traffic volumes across the 11 signals decreased an average 11% from 2019 to 2020.

3.3.2 Pedestrian Push-button Data Assembly

Time-stamped pedestrian push-button presses are recorded in high-resolution traffic signal controller log data (Smaglik et al., 2007). We used UDOT's ATSPM system (UDOT, 2021) to obtain push-button data for the time periods corresponding to the videos at each signal. Based on an earlier work by Singleton et al. (2020), we then calculated—for each hour and pedestrian phase (crossing)—several different measures of pedestrian traffic signal activity.

- **Pedestrian push-button presses**: The most direct measure of pedestrian push-button utilization or button-press behavior is event code 90 ("pedestrian detector on"). This occurs whenever a pedestrian push-button is activated (pressed), which could happen multiple times per cycle.
- **"Unique" pedestrian push-button presses**: Because one person may press the push-button multiple times in quick succession, we used time filters to remove successive push-button presses within a certain time interval. Testing indicated that a 15-second filter was the best fit to the observed volume data (Singleton & Runa, 2021).
- **Pedestrian actuations**: Other research (Kothuri et al., 2017) has used actuations rather than push-button presses to correlate pedestrian volumes. An actuation occurs upon the first time a push-button is pressed before being served, so usually just once per cycle. This measure was the best predictor of pedestrian volumes for crossings when on pedestrian recall (Singleton & Runa, 2021).

3.3.3 Analysis of Changes in Pedestrian Push-button Utilization

To determine if pedestrian push-button utilization changed during the COVID-19 pandemic, we compared the ratio of pedestrian push-button presses to pedestrian crossing volumes, which we define as the push-button use rate or utilization (presses per person). Our use of rates to measure pedestrian push-button utilization behavior is admittedly simplistic, but it was appropriate for our hourly data collection method, and it provides a first-stage look at COVID-related changes. Also, our second analysis (model prediction accuracy) better addresses the (nonlinear) relationship between push-button presses and pedestrian volumes.

To statistically analyze changes in utilization between the two years, we estimated a fixed-effects multilevel linear regression model (hours i at level one, signals j at level two) with no intercept ($Y_{ij} = \beta_j \times X_{ij}$), predicting hourly pedestrian volumes (Y_{ij}) as a function of pedestrian push-button presses (X_{ij}) across all crossings/phases, where the slopes (β_j) are fixed parameters that vary across signals j. Note that the slope (β) is the average number of pedestrians per push-button press, while the inverse slope ($1/\beta$) is the average number of push-button presses per pedestrian, our utilization rate. We allowed β_j to be different for each signal in each year ($Y_{ij} = \beta_{j,2019} \times X_{ij,2019} + \beta_{j,2020} \times X_{ij,2020}$); also, by dummy coding for 2020 ($Y_{ij} = \beta_j \times X_{ij} + \beta_{j,\Delta 2020} \times X_{ij,2020}$), null hypothesis tests of $\beta_{j,\Delta 2020}$ provided statistical significance of the change in slope at each signal from 2019 to 2020. Specifically, a decrease in the utilization rate (an increase in the slope) would suggest that people may have been avoiding push-buttons out of fears of contracting COVID-19.

3.3.4 Analysis of Changes in the Accuracy of Pedestrian Crossing Volume Estimates

To assess any changes in the accuracy of the pedestrian volume estimation models, we compared the model prediction errors between the two years. First, we applied the models developed by Singleton and Runa (2021) to estimate hourly pedestrian crossing volumes from traffic signal and pedestrian push-button data for both 2019 and 2020 and calculated the prediction errors (observed minus estimated). As previously mentioned, Singleton and Runa (2021) developed five piecewise linear or quadratic linear regression models for different situations: pedestrian hybrid beacon signals, crossings with pedestrian recall at high or low volume signals, and crossings with pedestrian recall at signals with short or long cycle lengths. To aid with application, the models used just one independent variable: whichever pedestrian signal activity measure best fit the data (unique push-button presses or pedestrian actuations).

Then, for each signal, we performed a Welch's (unequal variances independent samples) t-test on the model prediction errors for 2019 vs. 2020. Specifically, a significant difference (especially an increase) in prediction error would suggest that the pedestrian volume estimation models may need to be adjusted to remain accurate during the COVID-19 pandemic.

3.4 Results and Discussion

3.4.1 Analysis of Changes in Pedestrian Push-button Utilization

Table 2 summarizes the findings of the first analysis of changes in pedestrian push-button utilization. For most signals (seven out of 11 signals), the change in the utilization rate (push-button presses per person) from 2019 to 2020 was not statistically significant (change in slope: $p > 0.10$). Two other signals (1094 and 7184) had significant decreases in push-button utilization, while the utilization rate increased significantly at the final two signals (5108 and 8302). Aggregating across all 11 signals, the utilization rate increased from 1.08 in 2019 to 1.40 in 2020, which would suggest that people were pressing push-

buttons more often at signals during the early months of the COVID-19 pandemic. However, aggregate results appeared to have been greatly influenced by the noticeably different results at signal 8302; when removing that signal, the new aggregate results (for 10 signals) indicated a statistically significant reduction in push-button utilization from 2.11 to 1.47 presses per pedestrian.

Table 3.2 Pedestrians push-button utilization, 2019 vs. 2020, by signal and overall

	Push-button presses per pedestrian (utilization rate)			*Pedestrians per push-button press (β from model)*				
Signal	*2019*	*2020*	Δ	*2019*	*2020*	Δ	Δ *SE.*	*p*
1021	1.18	1.10	n.s.	0.85	0.91	0.06	0.04	0.104
1094	3.62	2.08	−	0.28	0.48	0.21	0.10	0.043
1229	1.26	1.35	n.s.	0.80	0.74	-0.06	0.08	0.511
4130	3.13	2.83	n.s.	0.32	0.35	0.03	0.31	0.912
5108	2.44	3.65	+	0.41	0.27	-0.14	0.06	0.032
5306	3.34	2.50	n.s.	0.30	0.40	0.10	0.09	0.259
5349	3.19	2.56	n.s.	0.31	0.39	0.08	0.07	0.286
6307	5.84	4.10	n.s.	0.17	0.24	0.07	0.10	0.479
7184	2.16	1.38	−	0.46	0.72	0.26	0.03	<0.001
8119	6.05	3.24	n.s.	0.17	0.31	0.14	0.16	0.381
8302	0.50	0.63	+	2.01	1.58	-0.43	0.07	<0.001
All 11 signals	1.08	1.40	+	0.93	0.71	-0.21	0.02	<0.001
10 signals (not 8302)	2.11	1.47	−	0.47	0.68	0.20	0.01	<0.001
9 signals (not 7184, 8302)	2.10	1.64	−	0.48	0.61	0.13	0.01	<0.001

Notes: n.s. = not significant

Figure 3.2, which shows plots of the relationships between pedestrians and push-button utilization at the four signals with the highest pedestrian activity in our study, illustrates these varied findings. Signal 1021, located in a transit-accessible area of Salt Lake City with numerous big-box stores, experienced a small (but not statistically significant) decrease in push-button utilization (from 1.18 to 1.10 push-button presses per pedestrian). Signal 1229, located in a neighborhood commercial district in Salt Lake City, saw a small (but not statistically significant) increase in the utilization rate (from 1.26 to 1.35). In both cases, 2020 observations generally fell in the same range as 2019 observations.

Signal 7184, located in a residential neighborhood of Salt Lake City near a large park, saw a significant decrease in pedestrian push-button utilization that is also apparent from the increased slope in the figure. Most pedestrians crossing at this intersection were observed going to/from the park, so it may be that people who were walking for recreation (rather than for transportation purposes) during the early pandemic were more cautious and concerned about COVID-19 spread from touching pedestrian push-buttons.

Signal 8302, located in downtown Moab in eastern Utah, was one of two signals to see a significant increase in utilization rate (push-button presses per person) from 2019 to 2020. Due to its proximity to popular Arches and Canyonlands National Parks, Moab is a tourist-oriented small city that attracts many visitors annually, making signal 8302 one of the highest pedestrian volume intersections in Utah (Singleton & Runa, 2021). The COVID-19 pandemic hit Moab hard after the National Park Service closed the parks to all visitors on March 28, 2020. Thus, the most noticeable difference for this signal in Figure 3.2 is that 2020 saw dramatically fewer pedestrians (during the months studied) than in 2019. We suspect that the "increase" in pedestrian push-button utilization found in our analysis (for this signal and overall) is more the result of lighter crowds and smaller pedestrian group sizes (perhaps due to social distancing) than any major change in pedestrians' willingness to press push-buttons due to COVID-19 concerns. In fact, results by Singleton et al. (2020) suggest that the relationship between pedestrians and

push-button presses is nonlinear: the slope increases (more pedestrians per push-button press) as push-button activity (per hour) increases. This also highlights a limitation of our linear analysis method: if overall activity decreased (as it did at signal 8302), then the overall slope would also decrease as well. Thus, we also performed a second analysis (described in the following subsection) that accounted for nonlinear relationships between pedestrian volumes and push-button utilization.

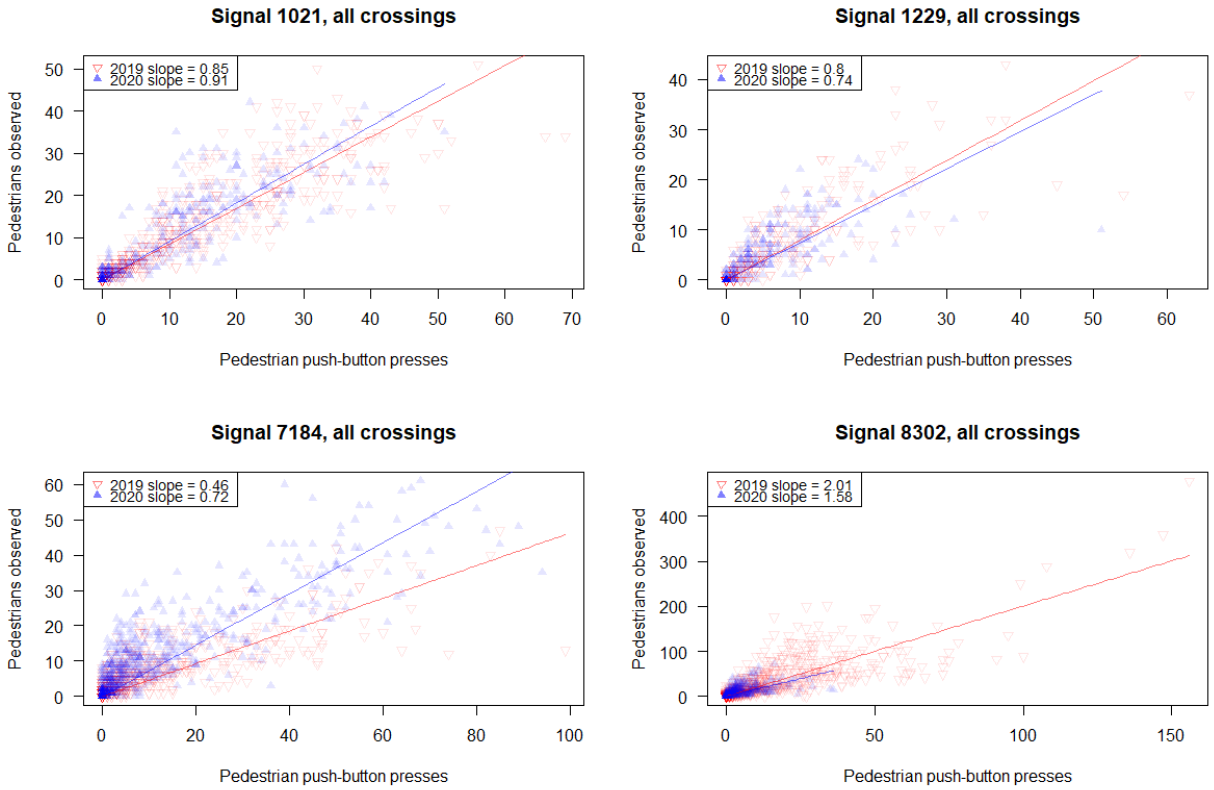


Figure 3.2 Pedestrian push-button use, 2019 vs. 2020, for signals 1021, 1229, 7184, and 8302. Each data point is one crossing observed for one hour on a given day, either in 2019 (empty red downward-pointing triangles) or in 2020 (filled blue upward-pointing triangles).

3.4.2 Analysis of Changes in the Accuracy of Pedestrian Crossing Volume Estimates

Table 3.3 shows the results of the analysis of changes in accuracy of the pedestrian volume estimation models, including the mean and standard deviation of the model prediction errors (observed minus estimated) in 2019 and 2020, and the results from the Welch's t-tests on those errors. Most signals (nine of 11) showed no statistically significant difference ($p > 0.05$) in the average error (pedestrians per hour) between the two years. Furthermore, the small changes in the mean errors did not show a consistent trend for all signals: some moved closer to zero (four) or farther from zero (seven), and some became more negative (five) or more positive (six). Aggregating across all 11 signals, the average error actually became less negative from 2019 to 2020 (−1.38 to −0.68); however, this improvement in accuracy was completely driven by significant differences at two signals (7184, 8302), as discussed later. The change (or lack thereof) in error does not seem to have been caused by more extreme but counteracting (i.e., larger positive and negative) errors, because the overall standard deviation of the prediction errors was smaller in 2020 than in 2019.

Results suggest that, in general, the models were still producing similarly accurate (if not more accurate) estimates of pedestrian crossing volumes during the COVID-19 pandemic. We suspect this may be due to the modest changes in pedestrian push-button utilization behavior identified in the first analysis: no statistically significant change for seven of 11 signals. Also, it could be that the models' methods of time-filtering the push-button data (the 15-second filter for "unique" presses and the use of actuations in some situations) are robust to COVID-induced changes in push-button utilization.

Table 3.3 Pedestrian volume model prediction errors, 2019 vs. 2020, by signal

	2019		*2020*		*Welch's t-test*		
Signal	*Mean*	*SD*	*Mean*	*SD*	*t*	*df*	*p*
1021	-5.02	11.58	-5.23	12.12	0.24	635	0.809
1094	-0.08	5.59	0.36	2.60	-1.17	440	0.241
1229	-4.00	10.00	-3.04	5.33	-1.38	423	0.168
4130	-0.19	1.26	-0.40	1.51	1.81	356	0.071
5108	-0.54	5.61	-0.86	2.06	1.10	549	0.272
5306	-0.16	2.61	-0.32	2.46	0.79	529	0.430
5349	-1.05	2.85	-0.80	1.66	-1.55	793	0.122
6307	-0.50	2.47	-0.85	3.51	1.49	623	0.137
7184	-1.46	5.78	0.20	7.10	-4.55	1,250	<0.001
8119	0.06	1.71	0.25	1.60	-1.78	781	0.076
8302	-2.42	26.51	1.69	9.89	-3.81	1,114	<0.001
All 11 signals	-1.38	11.59	-0.68	6.31	-3.68	8,553	<0.001
10 signals (not 8302)	-1.20	5.91	-0.94	5.72	-1.94	7,170	0.052
9 signals (not 7184, 8302)	-1.16	5.93	-1.24	5.25	0.60	5,951	0.550

As noted, we did find statistically significant differences in the prediction errors for two signals: 7184 and 8302. Nevertheless, we should also note that the absolute value of the mean errors for these two signals in 2020 was smaller than the absolute value of the mean errors in 2019, indicating that the model was actually more accurate (on average) during the COVID-19 pandemic. We have a few potential explanations for why these signals in particular saw changes and why the accuracy of the models might have increased.

As previously mentioned, signal 8302 in Moab saw greatly reduced pedestrian activity during the early months of the COVID-19 pandemic. Smaller crowds and pedestrian group sizes (and lower activity overall) provide fewer opportunities for large prediction errors, and the models tend to be more accurate (smaller magnitude errors) for lower-activity signals (Singleton & Runa, 2021).

In contrast, this explanation cannot account for the improved accuracy at signal 7184, since this location—near a popular large park in Salt Lake City—saw increased pedestrian activity early in the pandemic, especially on days with pleasant weather. One potential explanation unique to this location is that there was an open-streets installation on 900 S (Salt Lake City, 2020) during the study period that converted the outer travel lanes to space for active transportation, including a pop-up bike lane in the WB direction (an EB bike lane already existed). In 2019, we noticed that many people bicycling through this intersection used the crosswalks and push-buttons; thus, these cyclists added push-button presses but were not counted as pedestrians. Therefore, in 2020, perhaps there were fewer people bicycling on the sidewalk and "contaminating" the push-button counts, yielding more accurate model estimates of pedestrian volumes.

3.5 Conclusion

The first objective (Research Question 1) of this research was to examine how the utilization of pedestrian push-buttons at traffic signals changed during the early months of the COVID-19 pandemic. We expected push-button utilization to have decreased slightly due to concerns about using high-touch surfaces. At seven of the 11 signals, the change in utilization rate (push-button presses per pedestrian) was not statistically significant. Push-button utilization decreased at two signals (1094 and 7184) but increased at two others (5108 and 8302). Aggregated across all 11 signals, push-button utilization (presses per person) actually increased slightly in 2020, yet this was mostly driven by unique changes at one signal: 8302. Given this signal's location in a tourist town with high pedestrian volumes, we suspect this result is due to reductions in pedestrian group sizes. If there were fewer crowds or people traveled in smaller groups (due to social distancing), then we expected the observed increase in push-button presses per person (decrease in pedestrians per push-button press). Excluding signal 8302, the new aggregate results (for 10 signals) showed a statistically significant decrease in push-button utilization, from 2.1 to 1.5 presses per person. Thus, Hypothesis 1 was partially supported.

Our second objective (Research Question 2) focused on the accuracy of pedestrian volume estimation models based on traffic signal (push-button) data and developed in 2019, during the early months of the COVID-19 pandemic. We expected that button-press behavior had not changed enough to degrade the accuracy of the models (especially considering the 15-second filtering of extraneous push-button presses). Indeed, nine of the 11 signals saw no statistically significant change in accuracy between 2019 and 2020, while two signals (7184 and 8302) actually had more accurate pedestrian volume estimates in 2020 than in 2019. Aggregated across all 11 signals, the average model prediction error decreased from −1.4 to −0.7; this may have been the result of smaller crowds and pedestrian group sizes at signal 8302. Excluding signals 7184 and 8302, the new aggregate results showed effectively no change in average error (−1.16 to −1.24). Thus, Hypothesis 2 was supported. This is not surprising given the results of the first analysis, but it is still "good news" that the pedestrian volume estimation models seem to be similarly (if not more) accurate and do not need to be recalibrated to work during COVID conditions.

Overall, this research provides insights into the impacts of the COVID-19 pandemic on walking and pedestrian behavior, specifically regarding push-button utilization at traffic signals. Despite this narrow focus, the research addressed an important public health and signal operations question, indicating that, overall, people in Utah were not significantly deterred from using pedestrian push-buttons due to fears of contracting/spreading COVID-19. More recent understanding of COVID transmission sources (more from air than from surfaces) suggests that even modest changes in button-press behavior may not persist post-pandemic. Also, by investigating the accuracy of pedestrian volume estimation models based on traffic signal data, this research also addressed an important question for planning. Our results suggest that pedestrian volume estimates obtained during the COVID-19 pandemic (using models calibrated on pre-pandemic data) are no less accurate and may even be more accurate. Greater model accuracy could be the result of reduced pedestrian activity overall (by an estimated 16%; see Table 3.1) as well as smaller pedestrian group sizes (due to social distancing), both of which reduce large prediction errors. This indicates that the models can continue to be used. Pedestrian volume estimates from traffic signal data have been used for various planning (Singleton, Park, & Lee, 2021) and safety analysis (Singleton, Mekker, & Islam, 2021) purposes.

Still, this study was not without limitations that could be addressed through future work. There may be other factors for which we are not controlling that might explain differences in push-button utilization. Reduced motor vehicle traffic volumes (by an estimated 11%; see Table 1) may have encouraged/allowed some pedestrians to cross against traffic or in mid-block locations without pressing the push-button. Our data collection covered different months in different years, and button-press behavior may vary over the year; however, studying a seasonal effect is challenging because utilization or accuracy differences may

be related to volume differences, and also since we know that pedestrian volumes are affected by weather (Runa & Singleton, 2021) and the models are more accurate for lower-volume locations (Singleton & Runa, 2021). Although we studied 11 signals of different types and in different locations, certain locations could have seen larger or different changes in pedestrian push-button use behavior. Locations with different compositions of users or travel purposes (e.g., walking for recreation vs. transportation) might have seen different results. Even though the models remained similarly accurate during the pandemic, relationships may change or not be applicable outside of Utah. Thus, there is a continued need to validate the models with new data on a periodic basis (every couple of years). Also, our method of data collection and analysis limited us to hourly and more aggregate observations. Studying other pedestrian behaviors at traffic signals that may be of interest during the COVID-19 pandemic—group sizes, social distancing, walking speed, signal violations—would require more fine-grained data collection from videos. We encourage such research to continue to advance our limited understanding of how the COVID-19 pandemic affected pedestrian travel.

3.6 References

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

Beck, M. J., & Hensher, D. A. (2020). "Insights into the impact of COVID-19 on household travel and activities in Australia–The early days under restrictions." *Transport Policy, 96*, 76-93. https://doi.org/10.1016/j.tranpol.2020.07.001

Combs, T. (2020) "Local Actions to Support Walking and Cycling During Social Distancing Dataset." Pedestrian and Bicycle Information Center. http://pedbikeinfo.org/resources/resources_details.cfm?id=5209

De Vos, J. (2020). "The effect of COVID-19 and subsequent social distancing on travel behavior." *Transportation Research Interdisciplinary Perspectives, 5*, 100121. https://doi.org/10.1016/j.trip.2020.100121

Jenelius, E., & Cebecauer, M. (2020). "Impacts of COVID-19 on public transport ridership in Sweden: Analysis of ticket validations, sales and passenger counts." *Transportation Research Interdisciplinary Perspectives, 8*, 100242. https://doi.org/10.1016/j.trip.2020.100242

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*(1), 11-18. https://doi.org/10.3141/2644-02

Runa, F., & Singleton, P. A. (2021). "Assessing the impacts of weather on pedestrian signal activity at 49 signalized intersections in Northern Utah." *Transportation Research Record: Journal of the Transportation Research Board, 2675*(6), 406-419. https://doi.org/10.1177/0361198121994111

Salt Lake City. (2020). *Stay Safe, Stay Active* (accessed Nov 2020). https://www.slc.gov/transportation/2020/04/13/stay-safe-stay-active-streets-response-to-covid-19

Shakibaei, S., De Jong, G. C., Alpkökin, P., & Rashidi, T. H. (2021). "Impact of the COVID-19 pandemic on travel behavior in Istanbul: A panel data analysis." *Sustainable Cities and Society, 65*, 102619. https://doi.org/10.1016/j.scs.2020.102619

Shamshiripour, A., Rahimi, E., Shabanpour, R., & Mohammadian, A. K. (2020). "How is COVID-19 reshaping activity-travel behavior? Evidence from a comprehensive survey in Chicago." *Transportation Research Interdisciplinary Perspectives, 7*, 100216. https://doi.org/10.1016/j.trip.2020.100216

Singleton, P. A., Mekker, M., & Islam, A. (2021). "Safety in numbers? Developing improved safety predictive methods for pedestrian crashes at signalized intersections in Utah using push button-based measures of exposure" (Report No. UT-21.08). Utah Department of Transportation. https://rosap.ntl.bts.gov/view/dot/56362

Singleton, P. A., Park, K., & Lee, D. H. (2021). "Varying influences of the built environment on daily and hourly pedestrian crossing volumes at signalized intersections estimated from traffic signal controller event data." *Journal of Transport Geography, 93*, 103067. https://doi.org/10.1016/j.jtrangeo.2021.103067

Singleton, P. A., & Runa, F. (2021). "Pedestrian traffic signal data accurately estimates pedestrian crossing volumes." *Transportation Research Record: Journal of the Transportation Research Board, 2675*(6), 429-440. https://doi.org/10.1177/0361198121994126

Singleton, P. A., Runa, F., & Humagain, P. (2020). "Utilizing archived traffic signal performance measures for pedestrian planning and analysis" (Report No. UT-20.17). Utah Department of Transportation. https://rosap.ntl.bts.gov/view/dot/54924

Singleton, P. A., Taylor, M., Day, C., Poddar, S., Kothuri, S., & Sharma, A. (2023). "Impact of covid-19 on traffic signal systems: Survey of agency interventions and observed changes in pedestrian activity." *Transportation Research Record: Journal of the Transportation Research Board, 2677*(4), 192-203. https://doi.org/10.1177/03611981211026303

Singleton Transportation Lab. (2020). "Monitoring pedestrian activity in Utah in the time of COVID-19." Utah State University. https://singletonpa.shinyapps.io/ped-covid19/

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*(1), 97-106. https://doi.org/10.3141/2035-11

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." Purdue University. http://doi.org/10.4231/K4RN35SH

Utah Department of Transportation (UDOT). (2021). "Automated Traffic Signal Performance Measures." https://udottraffic.utah.gov/ATSPM

4. IMPUTING TIME SERIES PEDESTRIAN VOLUME DATA WITH CONSIDERATION OF EPIDEMIOLOGICAL-ENVIRONMENTAL (EPIENV) VARIABLES

This chapter is the accepted manuscript of an article published by the Transportation Research Board in the journal *Transportation Research Record: Journal of the Transportation Research Board*. It is reprinted here in accordance with Sage's Author Archiving and Re-Use Guidelines. To cite, please use this reference:

- Rafe, A., & Singleton, P. A. (2024). "Imputing time series pedestrian volume data with consideration of epidemiological-environmental variables." *Transportation Research Record: Journal of the Transportation Research Board*, 03611981241240758. https://doi.org/10.1177/03611981241240758

4.1 Abstract

In this study, we investigate the quality of pedestrian volume data, which holds significance for safety and urban planning purposes. We employ statistical methods, machine learning (ML) methods, and deep learning (DL) methods to first detect anomalies in pedestrian activity data, and then impute missing values. We accomplish this by analyzing daily time series data of pedestrian activity at traffic signals in the state of Utah from 2018 to 2022. Our approach utilizes vector autoregression (VAR) analysis—a multivariate time series analysis—by incorporating epidemiological-environmental (EpiEnv) variables, which consist of average temperature, precipitation, air quality index, and COVID-19 pandemic data. Additionally, we scrutinize the influence of built environment variables when mixed with EpiEnv variables on fluctuations in pedestrian volume data. Our findings suggest that the density-based spatial clustering of applications with noise (DBSCAN) method provides superior performance in anomaly detection, and that the random forest, long short-term memory (LSTM), and gated recurrent unit (GRU) techniques excel at imputing various categories of missing value patterns within temporal-based pedestrian volumes. The VAR analysis results also indicate that EpiEnv variables significantly affect the process of anomaly detection and imputation across all traffic signals. Our findings can assist urban and transportation planners in identifying the most impactful EpiEnv variables on pedestrian activity, which in turn can aid in the development of suitable strategies to promote walking as a mode of transportation.

4.2 Introduction

Walking is a vital part of active transportation, and it is important to understand how pedestrians move around the city. The way intersections are designed and operated can make a big difference for pedestrian safety, accessibility, and comfort. By collecting data on pedestrian volumes, urban planners and transportation engineers can identify where people walk the most and prioritize improvements such as crosswalks, traffic signals, and sidewalk infrastructure. However, collecting reliable and consistent data on pedestrian volume over time and years is challenging and costly (Ryus et al., 2022). Pedestrian volume can vary significantly depending on the location, season, weather, time of day, and other factors (Schneider et al., 2009). Additionally, the methods employed for pedestrian counting, such as manual and automated techniques utilizing video cameras or infrared sensors, pose challenges and can be costly to implement and maintain (Schneider et al., 2009). Consequently, there is a pressing need for innovative and cost-effective approaches to measuring pedestrian volume and utilizing this information for planning purposes.

To address the limitations of collecting real-time and updated pedestrian data for transportation studies, Schneider et al. (2012) and other researchers have developed direct-demand models. These models enable the estimation of pedestrian volumes at intersections by incorporating easily accessible variables, such as land use and socioeconomic features (Sobreira & Hellinga, 2023). However, it is important to note that most of these direct-demand models of pedestrian volume were derived from short-duration manual counts conducted at a limited number of locations, which in turn introduced limitations in the accuracy, generalizability, and sensitivity of the model results (Singleton, Park, & Lee, 2021). To overcome these restrictions, some researchers (Day et al., 2011; Blanc et al., 2015; Kothuri et al., 2017; Singleton & Runa, 2021) have utilized pedestrian data extracted from traffic signal controller logs to estimate walking activity at signalized intersections. Based on push-button presses and validated against observed pedestrian counts (Singleton & Runa, 2021), these estimated pedestrian volumes can be presented as time series data (hourly, daily, weekly, etc.) and are highly valuable for understanding pedestrian behavior in the context of planning (Singleton, Park, & Lee, 2021; Park et al., 2023) and for purposes of safety analysis (Singleton, Mekker, & Islam, 2021).

No matter the source, pedestrian volume data may exhibit anomalies due to a variety of factors, such as measurement inaccuracies, data entry errors, or disruptive events like inclement weather, special events, or construction activities (Ryus et al., 2014; Nordback et al., 2016). These anomalies are observations that significantly deviate from the rest of the dataset. Importantly, such deviations may either indicate true but atypical values—large crowds attending a public event; few pedestrians during a blizzard—or false values due to error sources such as data entry mistakes or faulty measurement tools. Therefore, the identification and handling of these anomalies are crucial to derive accurate conclusions and make appropriate decisions. To this end, we plan to implement rigorous anomaly detection techniques to distinguish between true atypical values and false values. In addition to anomalies, missing data is another challenge in pedestrian volume studies. The absence of data can distort the analysis, leading to misleading conclusions. To address this issue, we plan to employ appropriate imputation techniques to fill in the gaps where data are missing.

The objective of this study is to examine the performance of various methods for detecting and imputing anomalies in pedestrian volume data, taking into consideration the influence of environmental variables and disease epidemics, which we refer to as epidemiological-environmental (EpiEnv) factors. Specifically, we analyzed time series data of pedestrian volume at traffic signals in Utah, along with weather, air quality, and COVID-19 pandemic data from 2018 to 2022. We implemented statistical, ML, and DL methods to detect anomalies and impute missing values for each traffic signal. Through this process, we aim to establish a comprehensive framework for improving pedestrian data collection or estimation and identify the most effective method for detecting true anomalies and impute them using spatiotemporal variables (EpiEnv factors and land uses). In the following sections, we will present a literature review that explores the existing methods for detecting and imputing anomalies in pedestrian activity data. We will then describe our data and methods, present our results, and finally discuss the key findings.

4.3 Literature Review

As previously mentioned, pedestrian volume data, collected through various counting and estimation methods, may contain anomalies and missing data due to factors such as inaccurate or inconsistent data collection methods, environmental influences, alterations in infrastructure or traffic patterns, malfunctioning equipment (like infrared sensors or push-buttons), and other elements that impact pedestrian behavior (Ryus et al., 2022; Ryus et al., 2014). Our research has uncovered a scarcity of literature specifically dedicated to addressing the detection and handling of anomalies in datasets related to pedestrian activity.

Turner and Lasley (2013) proposed an automated method for detecting inaccurate counts of pedestrians and bicyclists. Their approach involved identifying outliers in the count data for each type by utilizing the interquartile range. The authors recommended performing a targeted manual review of select portions of the data by an experienced specialist to identify any anomalies that the automated method might have overlooked. In a separate study, Wang et al. (2013) employed a set of 10 variables, encompassing sociodemographic, built environment, weather, and temporal characteristics, to forecast pedestrian volume on urban multiuse trails. They utilized negative binomial regression models for this purpose, while also suggesting the application of this method to address missing pedestrian count data. Moreover, Ryus et al. (2014) proposed various strategies for detecting and managing anomalies in pedestrian volume data. One approach involves employing quality control procedures to identify and remove outliers or inaccurate data from both manual and automated counts. Another method is to compare count data with historical trends, seasonal patterns, or other sources of information to identify deviations from the expected patterns. Additionally, interpolation or imputation techniques can be utilized to fill in missing or anomalous data points. Finally, temporal and land use adjustment factors or models can be applied to account for variations in weather, day of the week, season, or other factors that may influence pedestrian volume. These approaches provide valuable strategies to enhance the accuracy and reliability of pedestrian volume data.

Recent advancements in quality assurance and quality control (QA/QC) methods for nonmotorized traffic data underscore the need for robust methodologies capable of adapting to diverse data sources and urban contexts. Kothuri et al. (2022) emphasize the integration of emerging data sources like Strava and StreetLight with traditional methods of traffic data collection, such as permanent and short-duration counts, using ML techniques to enhance bicycle volume estimates in various cities. This blending of newer and conventional data sources not only augments but also relies on traditional traffic count methods to achieve a more precise analysis. This approach underscores the importance of traditional data collection in complementing big data, thereby challenging the idea that big data alone is sufficient for comprehensive traffic analysis. Additionally, Lindsey et al. (2024) detail QA methods for hourly nonmotorized traffic counts, addressing the gap in standard procedures for validating such data. Their work with the Minnesota Department of Natural Resources provides a template for employing statistical tests to identify outliers and impute missing counts, thereby enhancing the validity of traffic flow estimates. This empirical approach aligns with the broader push in nonmotorized traffic monitoring toward incorporating QA principles outlined by foundational studies, such as those by Turner and Lasley (2013), and adapting to the variability inherent in nonmotorized data. Furthermore, Jackson et al. (2017) illustrated the practical application of QA/QC principles within North Carolina's Nonmotorized Volume Data Program, outlining a comprehensive approach to data management, including the incorporation of hourly checks to detect outliers. Similarly, Nordback et al. (2015) contributed to the standardization of QA checks, developing measures to address data gaps and anomalies that have been adopted by state departments of transportation. The evolution of QA procedures over the last decade reflects a shift toward more nuanced and comprehensive methodologies that recognize the complex nature of nonmotorized traffic data and the importance of data quality tailored to specific applications.

In broader contexts, a wide range of techniques are employed for detecting anomalies and performing imputation in spatiotemporal traffic and transportation data. These techniques encompass statistical models (Lam et al., 2017), distance measures, pattern analysis, as well as ML (Wang et al., 2015) and DL (Banifakhr & Sadeghi, 2021) learning methods.

In addition to finding suitable methods for detecting anomalies and imputing missing values in pedestrian data, it is crucial to explore the factors that can cause significant shocks, leading to sharp decreases or increases in pedestrian volume within specific time periods. Environmental factors, such as weather, play a substantial role in shaping changes in pedestrian volume and walking behavior. Temperature, precipitation, air quality, and season are among the various factors that can influence the number of

pedestrians and their walking patterns. According to a study conducted by de Montigny et al. (2012), a 5°C increase in temperature was associated with a 14% increase in pedestrian volume. Additionally, research by Runa and Singleton (2021) indicated that very hot maximum temperatures (≥ 90°F) were found to be linked to lower pedestrian activity at approximately one-third of the observed locations, while very low minimum temperatures (< 20°F) also resulted in decreased pedestrian activity. Precipitation is another significant factor affecting pedestrian volume. Rain has been identified as having the most substantial impact on pedestrian volumes at a given location, with cloud cover, wind, and extreme temperatures (both hot and cold) also contributing to decreased pedestrian volumes (Attaset et al., 2010). Moreover, Shaaban and Muley (2016) found that, on average, precipitation reduced the hourly volume of pedestrians by approximately 13%. Additionally, the study by Montigny et al. (2012) demonstrated that a transition from snowy conditions to dry conditions was associated with a notable increase of 23% in pedestrian volume. In addition, Holmes et al. (2009) highlighted the significant impact of weather conditions on walking behavior in their study conducted in Indianapolis, IN. They emphasized that increased hours of sunshine are positively associated with higher urban trail traffic, while rainfall, particularly heavy rainfall, discourages trail use. Interestingly, the research also noted a unique increase in trail traffic during snowfall, suggesting a regional or cultural enthusiasm for snowy conditions.

Regarding the impact of air quality on walking activity, recent studies have underscored the influence of air pollution and air quality warnings on pedestrian travel behavior. Yu and Zhang (2023) revealed that a 10-unit increase in the daily air quality index (AQI) was associated with a reduction in daily physical activity by six minutes of moderate-to-vigorous physical activity and 230 walking steps. Tribby et al. (2013) highlighted that such alerts significantly affect behavior, especially during high-exposure activities like outdoor exercise, though the impact is not uniform and varies by context and activity. A study by Xu et al. (2022) found no direct correlation between air quality alerts and micromobility usage in Austin, TX. However, they observed a decrease in usage on days with high pollution levels, suggesting a nuanced impact of air quality on travel decisions, particularly for short-distance trips. Additionally, Chung et al. (2019) discovered that the concentration of PM10 (particulate matter with a diameter of 10 micrometers or less) influenced individuals' intention to walk and had an impact on the volume of pedestrians at street level. These insights collectively highlight the importance of considering moderate air pollution when studying levels of pedestrian activity. Moreover, Holmes et al. (2009) underscored the influence of air quality on walking, with high levels of ozone and fine particulate matter correlating with reduced trail traffic, indicating that air pollution concerns may deter individuals from engaging in outdoor walking activities.

In addition to the environmental factors, epidemiological issues, such as the COVID-19 pandemic, have a notable impact on pedestrian volume and walking behavior. A study conducted in Utah found that pedestrian crossing volumes decreased by an average of 16% from 2019 to 2020, while motor vehicle traffic volumes decreased by an average of 11% during the same period (Runa & Singleton, 2023). Additionally, Hunter et al. (2021) discovered that COVID-19 response measures resulted in significant declines in walking, particularly utilitarian walking. However, recreational walking has shown a recovery and even surpassed pre-pandemic levels in various U.S. cities (Hunter et al., 2021). On the other hand, Beck and Hensher (2020) conducted a study examining the effects of COVID-19 on travel behavior in Australia. They found that lockdown measures led to a significant decrease in the use of private and public transportation, while active transportation modes like walking and cycling saw an increase from 14% to 20%. As restrictions began to ease, travel activity experienced a 50% increase, reaching 66% of pre-COVID-19 levels; however, the use of active transportation methods continued to trend upward.

Overall, the review of literature highlights a gap in the investigation of time-series pedestrian data and the manipulation of such data in the context of detecting anomalies and imputing missing values, particularly in relation to EpiEnv factors.

4.4 Data

4.4.1 Estimated Pedestrian Volumes from Traffic Signal Data

For this study, we utilized daily estimated pedestrian volume data from January 2018 to December 2022 in Utah. The pedestrian volumes were estimated based on research conducted by Singleton and Runa (2021). In their study, they utilized high-resolution data on pedestrian push-button activations obtained from traffic signal controller event logs in Utah. After collecting over 10,000 hours of observed pedestrian counts at 90 locations, they developed piecewise linear and quadratic regression models to estimate pedestrian volume from pedestrian signal data. The authors found that hourly pedestrian volumes predicted by the model from push-button data were highly correlated ($R^2 > 0.80$) with observed crossing volumes and had low mean absolute error (±3.0 pedestrians per hour) (Singleton & Runa, 2021). These models can estimate the annual average daily pedestrian crossing volumes at signalized intersections and identify locations with high pedestrian volume. Therefore, we employed their model to estimate daily pedestrian volume for 2,113 traffic signals across Utah (see Figure 4.1) using data from the Automated Traffic Signal Performance Measures (ATSPM) system (UDOT, 2023). Most of these signals are situated in the northern region of Utah; 1,031 (49%) are located in Salt Lake County. From 2018 to 2022, across all signals, we estimated approximately 807 million pedestrian crossings.

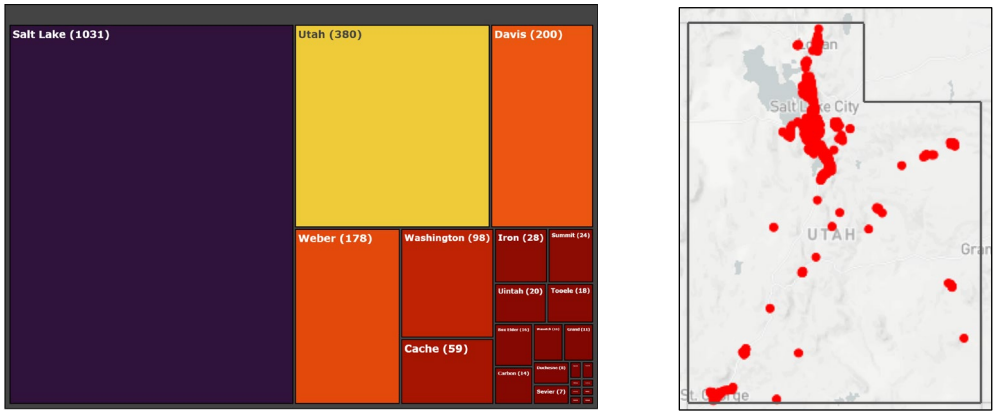


Figure 4.1 The dispersion of investigated traffic signals through a tree and point map

Table 4.1 presents the descriptive statistics of estimated daily pedestrian volume data by year. In 2019, Utah traffic signals exhibited the highest mean daily pedestrian volume: approximately 251 pedestrians per signal per day. However, 2020 witnessed a significant decrease in pedestrian activity, likely attributed to the outbreak of COVID-19. All maximum values were observed in traffic signals located near educational land uses, such as the University of Utah, Utah State University, and certain high schools. Notably, these peaks were observed during January, February, August, and September, which coincide with the beginning of the spring and fall school semesters.

Table 4.1 Descriptive statistics of estimated daily pedestrian volume data, by year

Year	*Min*	*Median*	*Max*	*Mean*	*SD*	*25th*	*75th*
2018	1.09	73.37	53,973.43	230.73	651.01	18.23	192.92
2019	1.11	77.95	41,431.38	251.63	693.08	21.37	200.48
2020	1.06	66.92	53,330.00	187.90	635.12	20.66	155.52
2021	1.05	70.47	57,340.10	205.04	666.71	22.23	168.89
2022	1.08	78.14	50,428.73	235.90	711.80	24.30	191.08

4.4.2 Epidemiological-environmental (EpiEnv) Data

As mentioned earlier, one of our objectives is to investigate the influence of EpiEnv variables on the temporal variation in pedestrian volume. The first category of environmental variables we focused on is weather. We collected information on precipitation, high temperature, and low temperature (ISU, 2023) for each traffic signal from 2018 to 2022. Figure 4.2 displays a sample time series of these data specifically around a traffic signal located in Cache County. During the study period, temperatures ranged from a high of 109°F (in 2021) to a low of -11°F (in 2022).

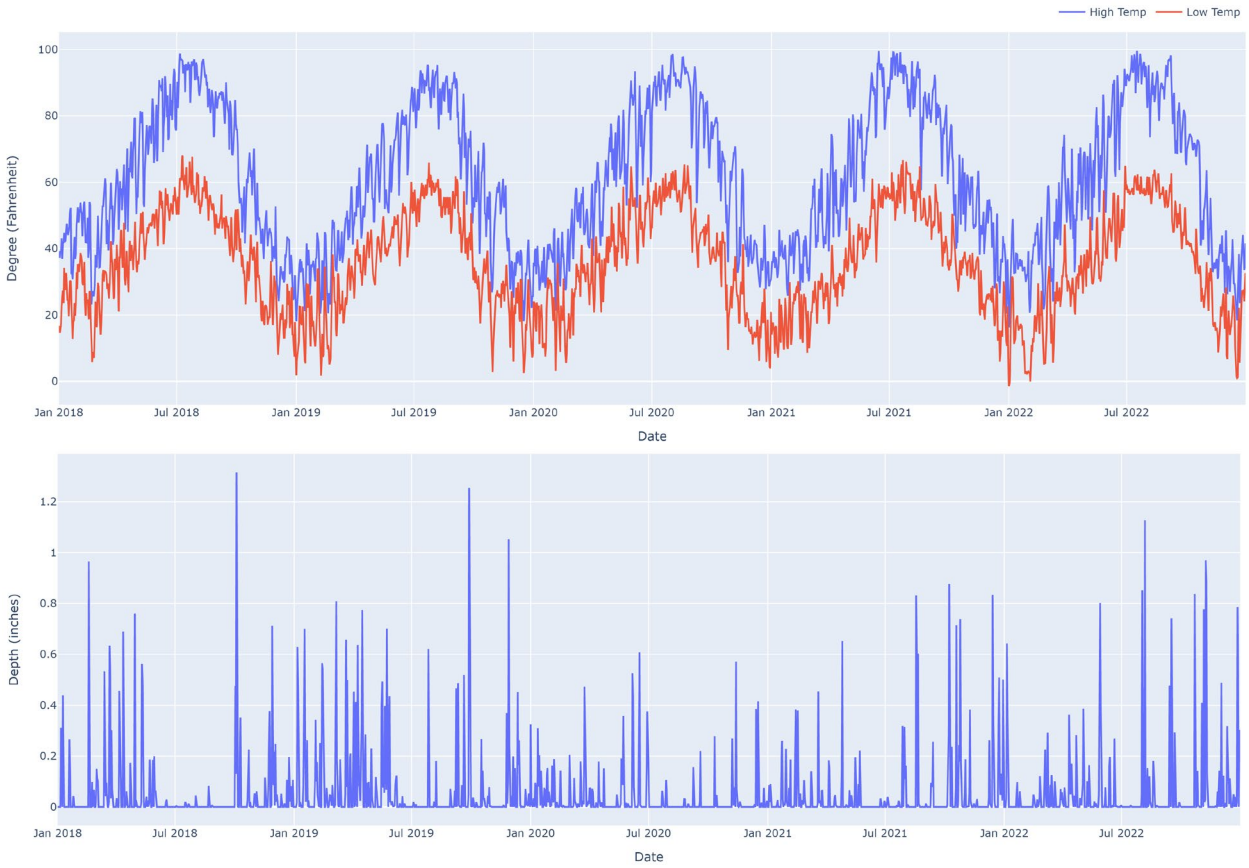


Figure 4.2 Time series of daily high and low temperature (above) and precipitation (below) around a traffic signal in Cache County

The second type of environmental variable we examined is air pollution, measured using the air quality index (AQI). These data were collected using AirNow API (US EPA, 2023) for each signal from 2018 to 2022 and were based on measurements of PM2.5, ozone, NO2, PM10, and CO. Figure 4.3 illustrates a heatmap of the yearly-monthly matrix of average AQI values around traffic signals in Utah, highlighting the variations in AQI levels throughout the year. The heatmap reveals higher AQI values occurring during the months of June, July, and August consistently each year. Additionally, Figure 4.4 presents the AQI categories for each year in Utah, providing an overview of the distribution of air quality conditions. The figure indicates that Utah experienced a higher number of days with good air quality in 2020, while in 2021, there was an increase in the occurrence of unhealthy air quality days for both sensitive and general populations.

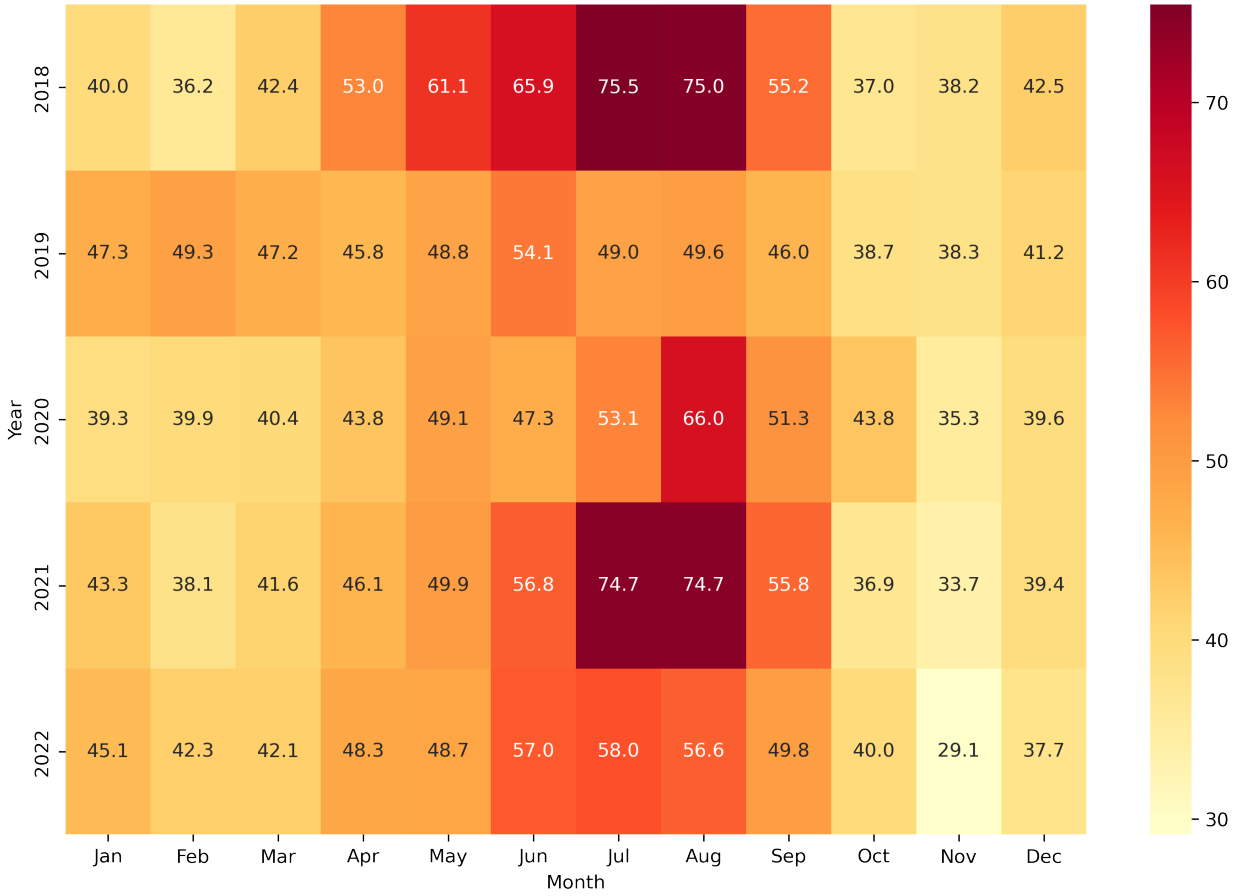


Figure 4.3 Heatmap of AQI (yearly-monthly matrix) around a traffic signal in Utah

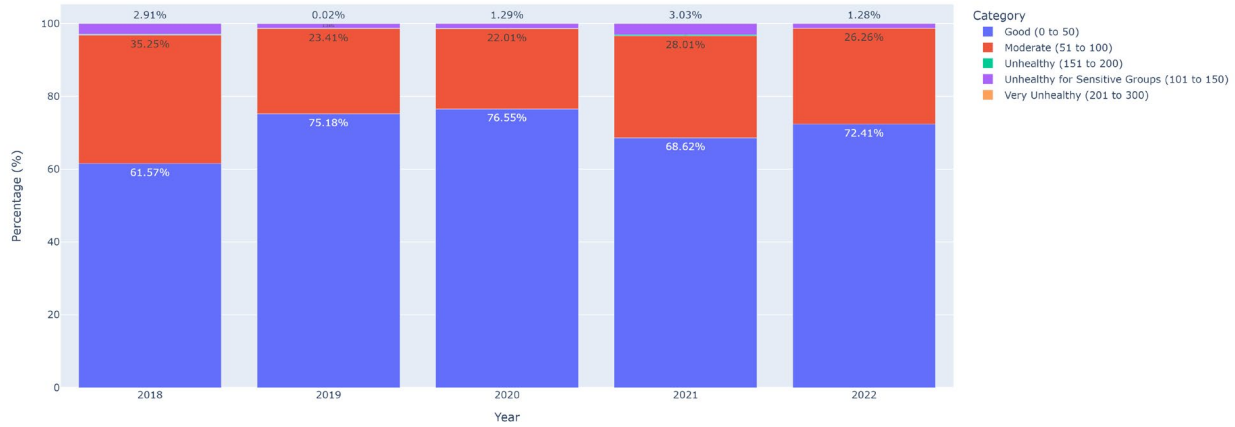


Figure 4.4 The yearly AQI categories in Utah

In addition to these environmental variables, we also extracted information about the built environment surrounding each traffic signal within a quarter-mile buffer, which is considered a typical distance by which the built environment likely affects walking behaviors. This allowed us to analyze the trends in pedestrian volume and identify patterns of anomalies and missing values in relation to built environment variables. To conduct this analysis, we followed the methodology outlined in the Singleton, Park, and Lee (2021) study, which involved extracting several variables related to the built environment, including density, diversity, design, destination accessibility, and distance to transit. Table 4.2 provides descriptive

statistics of the built environment variables that were extracted for each signal. It is important to note that the built environment variables were not utilized in the anomaly detection and imputation steps, but were instead applied in the analysis phase to provide deeper context and enhance our understanding of the results derived from these processes.

Table 4.2 Descriptive statistics for built environment variables

Variable	*Mean*	*SD*
Population density (1000 per sq. mi.)	4.49	3.01
Vehicle ownership (average)	1.73	0.45
% residential land use	31.00	23.72
% commercial land use	28.03	20.92
# schools	0.28	0.59
# places of worship	0.48	0.78
# transit stops	5.94	3.54

In the second part of EpiEnv variables, we obtained COVID-19 data from the Utah coronavirus dashboard (Utah DHHS, 2023) as epidemiological information. We collected these data for 13 local health districts and assigned the data to the respective traffic signal locations within each district. Utah has gone through different phases of COVID-19 response since March 2020. The first phase was the urgent phase, which lasted from mid-March to late April 2020, when the state implemented strict public health measures to slow down the spread of the virus. The second phase was the stabilization phase, which lasted from late April 2020 to early June 2021, when the state gradually eased some restrictions and reopened some sectors of the economy while maintaining social distancing and mask wearing. The third phase was the recovery phase, which lasted from early June 2021 to late February 2022, when the state lifted most of the remaining restrictions and focused on increasing vaccination rates and testing capacity. The fourth and current phase is the endemic phase, which started on March 1, 2022, when the state declared that COVID-19 is no longer a public health emergency but a seasonal respiratory disease that can be managed with routine measures. Figure 4.5 displays the daily COVID-19 case rate per 100,000 in Utah, providing an overview of the trends in COVID-19 cases. Based on the information presented in the figure, the highest number of COVID-19 cases in Utah occurred in January 2022 during Phase 3.

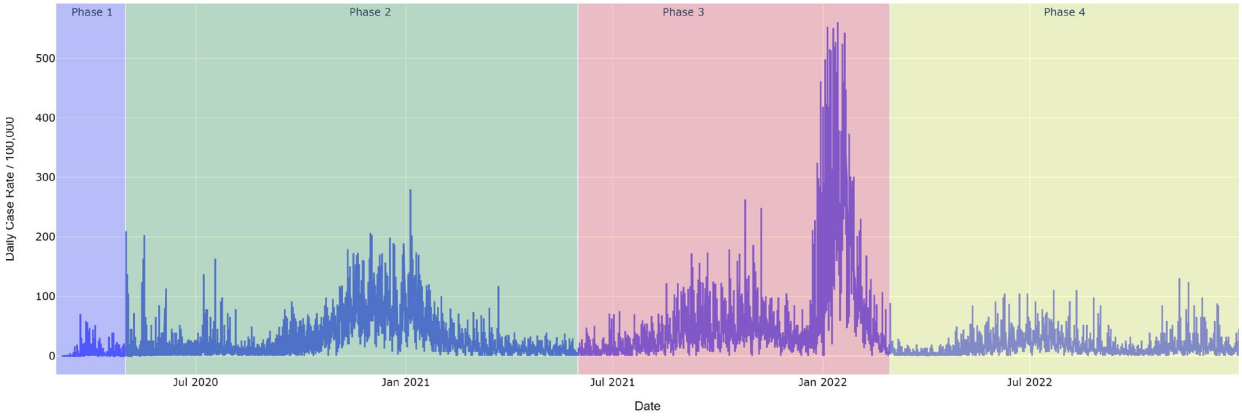


Figure 4.5 The daily COVID-19 case rate per 100,000 in Utah

In interpreting the COVID-19 case rates presented in Figure 4.5, particularly during Phases 3 and 4, it is crucial to consider the effects of the widespread adoption of at-home testing. This trend likely contributed to an underestimation of reported cases as not all at-home test results would be officially recorded (Rader et al., 2022). While this underreporting could influence the absolute case numbers, it does not impact the efficacy of the methodologies employed in our study.

4.5 Method

To develop a comprehensive framework for analyzing pedestrian volume data in this study, we applied and examined the performance of a wide range of anomaly detection and imputation methods. After the data preparation process, as depicted in Figure 4.6, this study consists of two main parts: anomaly detection and imputation.

4.5.1 Anomaly Detection

In this section, we begin by exploring the concept of data and identifying periods during which pedestrian volumes experience significant changes. Since our pedestrian volume data exhibit non-stationarity, we employ the least absolute deviation (LAD) method (Bai, 1995) to detect change points in the time series data. The LAD cost function, $C(s)$, defined in Equation 1, is instrumental in detecting where substantial changes in the median level of pedestrian volume occur within the time series. Here, "substantial" refers to deviations that exceed the typical variability observed in our data, which is informed by historical median values and the interquartile range—providing a robust and tailored threshold for our specific dataset. The segmentation s, the pedestrian volume $P(t)$ at time t and the median value $m(s)$ within segment s are the core components of this function. We assign penalties across segmentations of the time series data to ultimately identify a segmentation pattern that minimizes this cost. A key aspect of our approach is the incorporation of a dynamic programming method to determine an optimal number of change points, which serves as a constraint. The chosen number of change points is crucial as it balances the sensitivity of the method to detect true changes against the risk of overfitting to random fluctuations. This methodology allows for an efficient computation that avoids exhaustive comparisons across all possible segmentations. The dynamic programming method, therefore, offers an optimization process that is computationally manageable and methodologically sound, ensuring the integrity of our anomaly detection process.

$$C(s) = \Sigma\, |P(t) - m(s)| \qquad (1)$$

After identifying the change points and segment periods for each signal, we applied the vector autoregression (VAR) model on pedestrian data within each time segment. This model identifies and extracts time periods during which pedestrian volume is influenced by exogenous EpiEnv variables. For a $(n \times 1)$ vector of $P(t) = [P_1(t), \dots, P_n(t)]$ representing the pedestrian volume time series, the VAR model is formulated as follows:

$$P(t) = c + \sum_{i=1}^{p} a_i P(t-i) + \sum_{i=1}^{p} b_{1i} TE(t-i) + \sum_{i=1}^{p} b_{2i} PE(t-i) + \sum_{i=1}^{p} b_{3i} AQI(t-i) + \sum_{j=1}^{q} b_{4j} COV(t-j) + e_t \qquad (2)$$

where, c is the constant term, t is a $D \times 1$ vector of time series, a_i is the coefficient for the ith lag of the pedestrian volume at time $(t-i)$, b_{1i} , b_{2i} , b_{3i} are the coefficients for the ith lags of the exogenous variables TE (average temperature), PE (precipitation), and AQI at time $(t-i)$ in the lag period p, and b_{4j} for COV (COVID-19) at time $(t-j)$ in the lag period q, and e_t is the error term at time t.

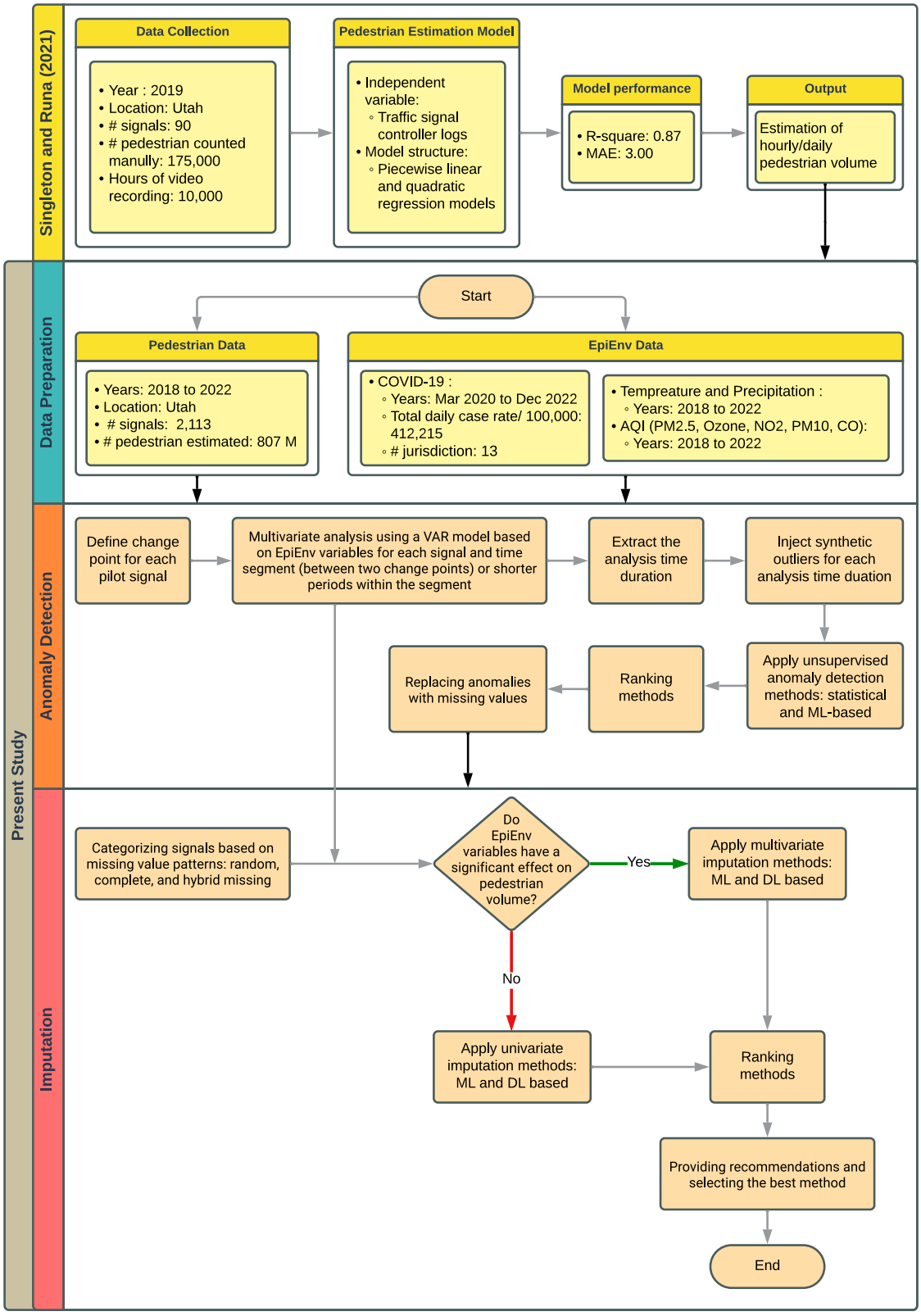


Figure 4.6 Conceptual framework of anomaly detection imputation for pedestrian volume data

Given that the COVID-19 data begin from 2020, we accounted for the different lag periods by determining the appropriate values of p and q based on the smallest Akaike Information Criterion (AIC) (Akaiki, 1974). We conducted the VAR analysis for each segment (time duration between two change points) or shorter period, in some cases, to identify the duration of time where the pedestrian volume is influenced by at least one of the exogenous variables or none of them. These durations, which we call analysis durations (AD), serve as the foundation for subsequent steps.

In the next step, to apply unsupervised anomaly detection methods and assess their performance, we selected a sample dataset and injected synthetic outliers into the data. This allowed us to evaluate which methods are more effective in detecting these outliers. We chose all 59 traffic signal data from Cache County as our sample dataset. For injecting synthetic outliers, we randomly selected values within each AD and increased or changed them to alter the trend in neighboring data points by a random percentage ranging from 50% to 85%. This approach allowed us to account for shocks from exogenous variables that are extracted from change point detection and VAR analysis steps, as well as atypical values. Figure 4.7 presents the original data along with the data containing the injected outliers, while the gray vertical lines represent the boundaries of the ADs.

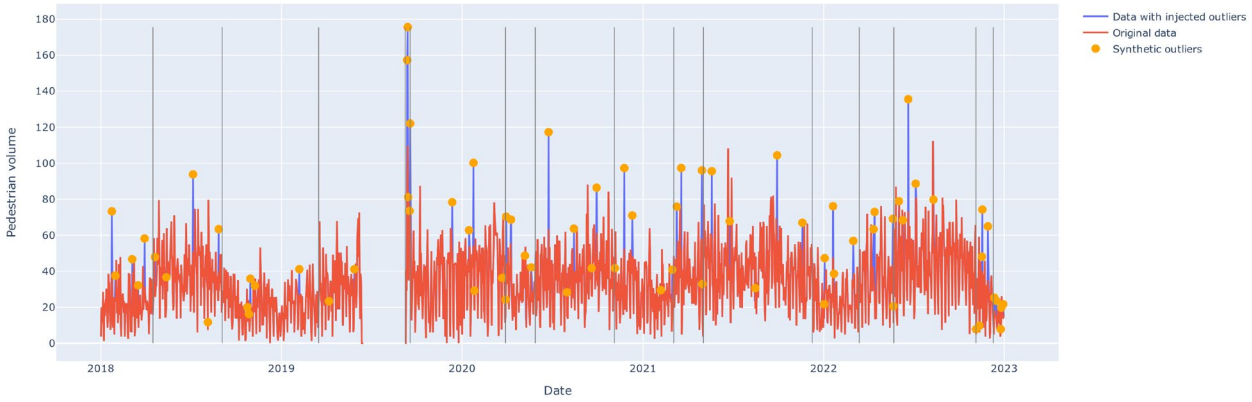


Figure 4.7 The pedestrian volume data with and without the injected outliers for a sample traffic signal

We then applied various statistical, ML, and DL unsupervised anomaly detection methods to the sample data.

- Statistical method
  - *Z-score*: Identifies anomalies based on how many standard deviations away from the mean a data point lies, providing a measure of its abnormality.
- ML algorithms
  - *K-nearest neighbor (KNN)* (Hautamaki et al., 2004): Flags data points as anomalies if they differ significantly from their nearest neighbors in the dataset.
  - *One-Class SVM (OCSVM)* (Ma & Perkins, 2003): Differentiates between "normal" and "anomaly" data points by learning a boundary around most data points.
  - *Isolation Forest (iForest)* (Liu et al., 2008): Detects anomalies by isolating points; the fewer splits required, the more likely a point is an anomaly.
  - *Density-based spatial clustering of applications with noise (DBSCAN)* (Ester et al., 1996): Groups closely packed points and mark low-density areas as anomalies.
  - *Seasonal Autoregressive Integrated Moving Average (SARIMA)* (Hanbanchong & Piromsopa, 2012): A time series modeling technique that can highlight anomalous points by contrasting them with seasonal trends and cycles.

- DL models
  - *Long Short-Term Memory (LSTM)* (Malhotra et al., 2016): An advanced RNN variant capable of learning order dependence in sequence prediction problems, effective in pinpointing anomalies over time.
  - *Generative Adversarial Networks (GANs)* (Li et al., 2019): A pair of neural networks contest with each other to respectively generate potential anomalies and evaluate their authenticity.
  - *Gated Recurrent Units (GRU)* (Li et al., 2019): A streamlined alternative to LSTMs, which also excels at modeling temporal sequences for anomaly detection.

To assess the effectiveness of each method, we utilized performance metrics such as accuracy, precision, recall, and F1 score. The accuracy of a model is determined by the percentage of observations that are correctly classified; precision measures the proportion of true positives among positive predictions; recall measures the proportion of true positives among actual positives; and F1 score is the harmonic mean of precision and recall. These metrics are calculated in Equations 3 to 6:

$$accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive\ +\ True\ Negative\ +\ False\ Positive\ +\ False\ Negative} \quad (3)$$

$$precision\ = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \quad (4)$$

$$recall\ = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \quad (5)$$

$$F1\ score = \frac{2 \times (precision \times recall\ )}{precision + recall} \quad (6)$$

After calculating these metrics, we ranked the methods based on their performance on the sample data (59 signals). The highest-ranked anomaly detection method was then applied to the pedestrian volume data for all signals in Utah. Finally, detected anomalies were replaced with missing values.

During the anomaly detection process on the sample data, in addition to the injected outliers, the methods detected some other outliers in certain signals. To investigate these outliers, we relied on domain knowledge and examined whether there were any special events on the dates identified as anomalies. We utilized web scraping methods and Utah events calendar websites for this purpose. Therefore, in cases where we identified anomalies that were confirmed to be genuine anomalies based on domain knowledge, we classified them as true positives.

4.5.2 Imputation

In this part, the imputation of missing values was carried out by first categorizing the traffic signals based on their missing value patterns: signals with randomly scattered missing values, signals with missing values covering complete periods of time, and signals with a hybrid pattern of missing values (a combination of complete and random missing values). Subsequently, for each AD, we examined whether the EpiEnv variables had a significant effect on pedestrian volume using the VAR model output. If the pedestrian data in an AD with missing values were found to be influenced by the EpiEnv variables, we applied multivariate imputation methods. Conversely, if there was no significant effect, we utilized univariate imputation methods. To impute the missing values, we employed ML-based and DL-based methods.

- ML-based method
  - *Random Forest*: An ensemble learning method that uses a multitude of decision trees to predict missing values based on the patterns found in the data, making no assumptions about data linearity.
- DL-based methods
  - *Recurrent Neural Network (RNN)*: Designed to recognize patterns in sequences of data; RNNs use their internal state (memory) to process variable length sequences; ideal for imputing missing values in time-series data.
  - *Long Short-Term Memory (LSTM)*: A type of RNN that is capable of learning long-term dependencies, making it highly effective for sequential prediction tasks, such as time-series imputation where past information is crucial.
  - *Gated Recurrent Units (GRU)*: Like LSTMs, GRUs streamline the model architecture to improve efficiency while still capturing temporal dependencies necessary for accurately predicting missing values.
  - *Temporal Convolutional Networks (TCN)* (Bai et al., 2018): Utilize a series of dilated convolutions to capture temporal correlations, which makes them well-suited for imputation where the missing data points are in sequences.

To assess the performance of these methods, we employed the mean absolute error (MAE) and root mean square error (RMSE) as evaluation metrics.

$$MAE = \frac{\sum_{i=1}^{n} | P_i - \hat{P}_i |}{n} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - \hat{P}_i)^2}{n}} \quad (8)$$

where, n is the total number of missing items, P_i is the real value of the ith missing item, and $\hat{P}_i$ is the imputed value of the ith missing item. By calculating the MAE and RMSE, we can assess the accuracy and quality of the imputation methods in capturing the differences between the actual and imputed values of the missing items for each missing pattern. To access additional details about our study's methodology, sample data, and analysis scripts, visit our study's GitHub repository (Rafe & Singleton, 2023).

4.6 Results

In the first phase of this study, we performed VAR analysis and identified ADs based on the lag values derived from the smallest AIC. Subsequently, we applied various unsupervised anomaly detection methods to our sample data, the 59 traffic signals in Cache County. The performance results of these methods are presented in Table 4.3. One important metric for evaluating anomaly detection methods is accuracy, which indicates the extent to which the methods are successful in classifying anomalies. The results indicate that DBSCAN achieved the highest accuracy (86.7%), followed by GRU, iForest, and KNN, which had similar accuracy scores. In terms of F1 score, DBSCAN also achieved the highest score (0.836) compared with the other methods, followed by KNN, iForest, and LSTM.

Table 4.3 Performance comparison of different anomaly detection methods on Cache County traffic signals with injected synthetic outliers

Anomaly detection method	*Accuracy*	*Precision*	*Recall*	*F1 score*
Z-score	0.730	0.724	0.681	0.702
KNN	0.753	0.796	0.796	0.796
OCSVM	0.593	0.686	0.661	0.673
iForest	0.760	0.798	0.763	0.780
DBSCAN	**0.867**	**0.836**	**0.836**	**0.836**
SARIMA	0.667	0.754	0.688	0.719
LSTM	0.699	0.891	0.683	0.773
GRU	0.767	0.855	0.611	0.713
GANs	0.645	0.636	0.523	0.574

Another important aspect in the anomaly detection process is the tuning of hyperparameters and the computation time or speed of the methods. To evaluate these parameters, we applied the top three ranked methods (DBSCAN, iForest, and KNN) to the full Utah traffic signals dataset. Our findings indicate that the computation times for DBSCAN and KNN are comparable, but tuning the hyperparameters of DBSCAN can be more challenging compared with KNN. On the other hand, the iForest method exhibited good accuracy in detecting anomalies and had a lower computation time than both DBSCAN and KNN. As a result, there is a tradeoff to consider when selecting the best method, weighing accuracy against the simplicity of hyperparameter tuning and computational speed, with DBSCAN, KNN, and iForest offering different advantages. Table 4.4 displays the optimal hyperparameters for DBSCAN, KNN, and iForest in this study. Additionally, Figure 4.8 illustrates the cleaned pedestrian volume data for a traffic signal in Cache County, where anomalies detected by the DBSCAN method have been replaced with missing values.

Table 4.4 Optimum hyperparameters for anomaly detection methods in this study

Anomaly detection method	*# hyper-parameters*	*Hyperparameters*	*Optimum hyperparameter for the database*
DBSCAN	4	Metric to calculating distance between instances in a feature array, Algorithm compute pointwise distances and find nearest neighbors, Maximum distance between two samples, Number of samples in a neighborhood of a core point	Euclidean, Ball Tree (Omohundro, 1989), 0.5, 11
iForest	2	Number of base estimators, outlier fraction	150, 0.25
KNN	2	Outlier fraction, number of neighbors	0.15, 8

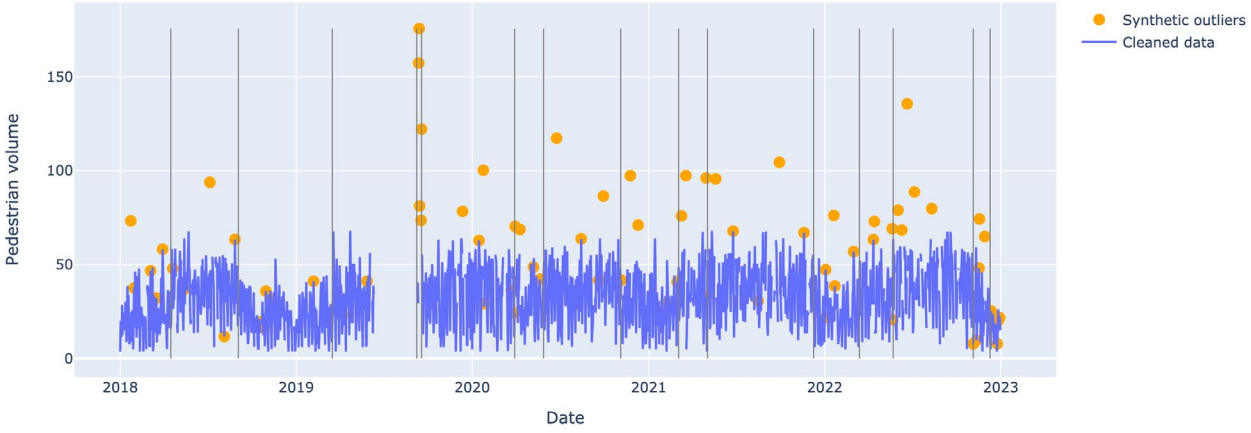


Figure 4.8 The cleaned pedestrian volume data (replace anomalies with missing values) with DBSCAN method in a sample traffic signal in Cache County

In this study's second phase, we applied imputation methods to handle the dataset's missing (and anomalous) values. To begin with, we categorized the traffic signals based on their missing value patterns. Among the 2,113 traffic signals in Utah, 67% exhibited a hybrid missing value pattern, 21% had a complete missing values pattern, 9% had random point missing values, and 3% did not have any missing values. This categorization allowed us to tailor the imputation methods to the specific missingness patterns. Table 4.5 presents the evaluation performance of each imputation method for the different missing value patterns observed. Figure 4.9 illustrates the polar plot for each imputation method, showcasing their performance based on MAE and RMSE.

Table 4.5 The performance evaluation of imputation methods for each missing value pattern

Random point missing value pattern		
Imputation method	*MAE*	*RMSE*
Random Forest	**5.359**	**8.928**
LSTM	15.921	18.717
GRU	10.974	14.530
TCN	29.234	31.921
RNN	13.323	17.179
Complete missing value pattern		
Imputation method	*MAE*	*RMSE*
Random Forest	19.469	23.703
LSTM	**11.686**	**12.366**
GRU	15.902	17.708
TCN	35.628	39.802
RNN	13.543	17.361
Hybrid missing value pattern		
Imputation method	*MAE*	*RMSE*
Random Forest	24.515	27.053
LSTM	9.434	10.785
GRU	**5.358**	**6.619**
TCN	21.047	24.167
RNN	7.734	9.264

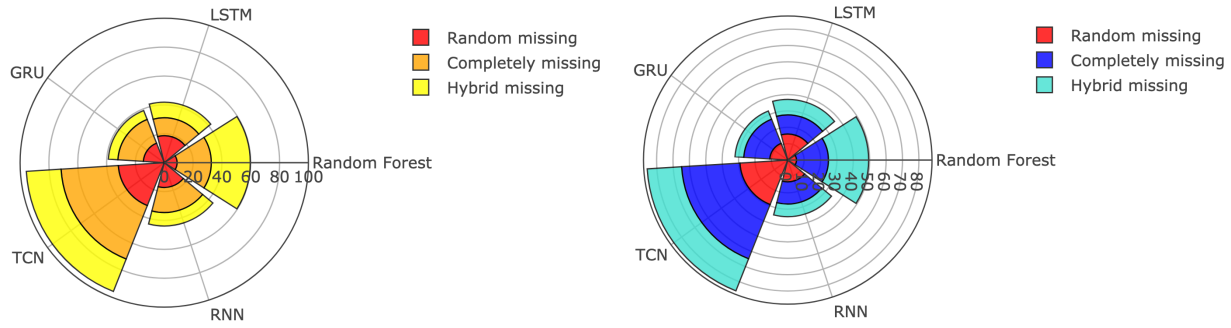
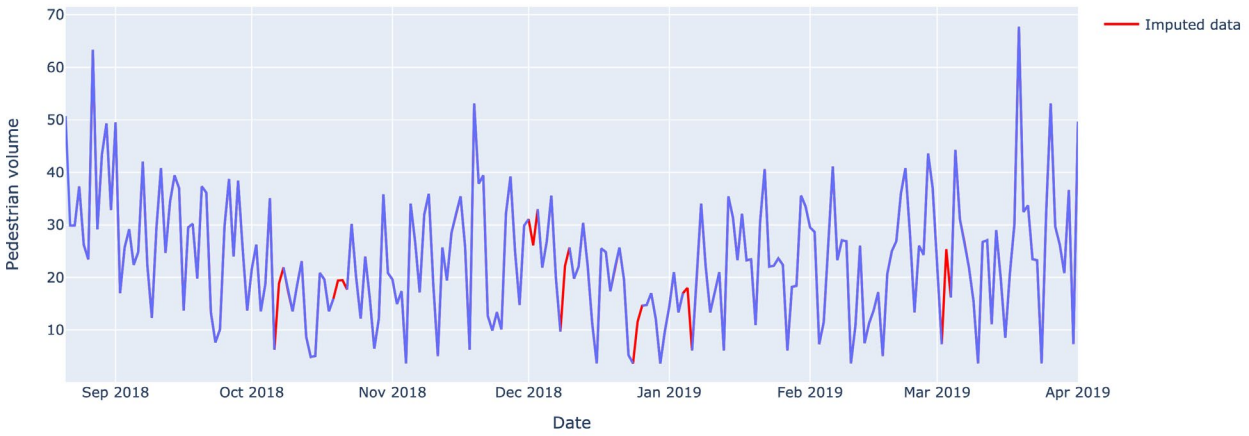


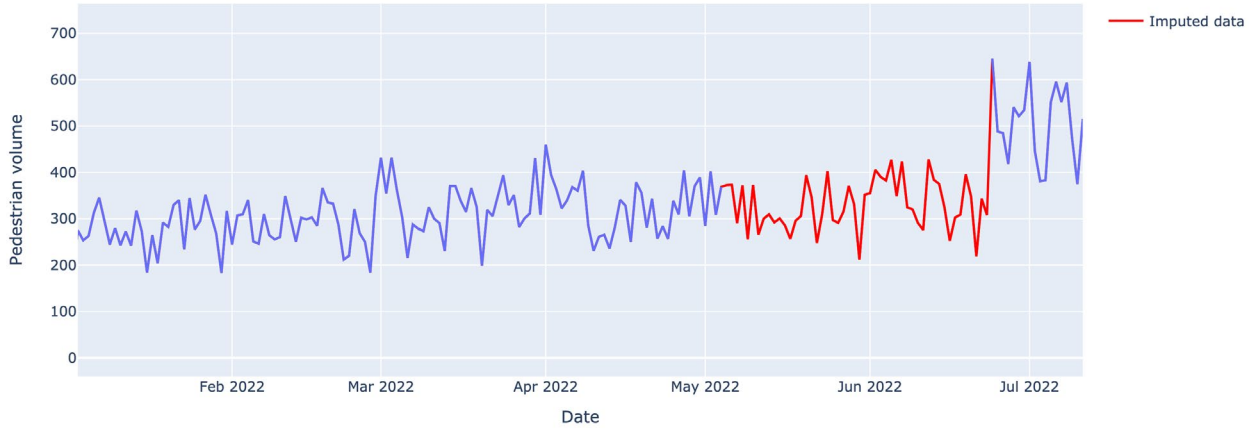
Figure 4.9 The polar plot of performance evaluation data of imputation methods based on MAE (left) and RMSE (right)

As seen in the table and figure, the random forest method performed better in the random point missing value pattern category, the LSTM method demonstrated better performance for complete missing value patterns, and the GRU method showed better performance among hybrid missing value patterns. These results indicate that each method excelled in imputing missing values in specific patterns, highlighting the importance of considering the nature of the missing data when selecting an appropriate imputation

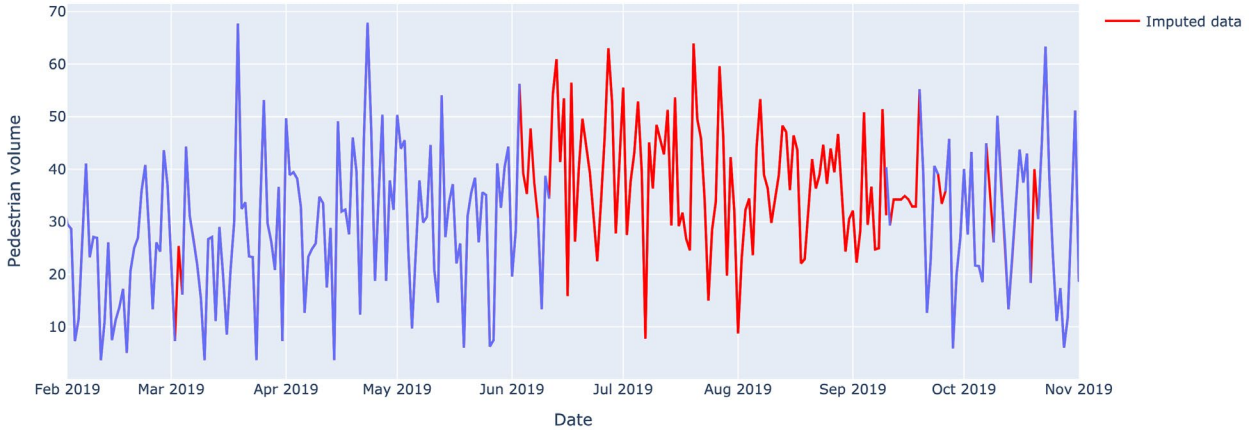
method. Figure 4.10 displays the results of imputation performed by each of the mentioned methods on various sample traffic signals in Utah.



(a) Imputation by Random Forest method on random point missing value pattern



(b) Imputation by LSTM method on complete missing value pattern



(c) Imputation by GRU method on hybrid missing value pattern

Figure 4.10 The results of imputation performed by random forest (a), LSTM (b) and GRU (c) on various sample traffic signal in Utah

4.7 Discussion

Given the unique data imputation methods used in our study, including random forest, LSTM, GRU, TCN, and RNN, the pattern of missing values takes on increased significance. The models employed, such as the LSTM, GRU, and RNN, are sequence-based models that can capture temporal dynamics, but varying patterns of missingness could affect their learning process and predictive accuracy. The experience of imputing missing values in this study reveals that when dealing with time-series data that exhibit various patterns of missing values, random forest, LSTM, and GRU methods each present unique challenges in terms of computation time and hyperparameter tuning. Random forest, while powerful, can become computationally intensive and slow with high-dimensional multivariate time series and large datasets, requiring substantial computational resources. Hyperparameter tuning in random forest, such as determining the optimal number of trees and depth, can also be a time-consuming task. LSTM and GRU models, given their complex nature, often require significant training time, especially with larger datasets, making them less suitable for scenarios needing frequent updates. These models also have many hyperparameters (like the number of hidden layers, hidden units, learning rate, etc.) that need tuning, and this process can be quite time consuming and computationally demanding, adding to the overall complexity. Further, the risk of overfitting in LSTM and GRU models necessitates the use of regularization techniques, which introduce additional hyperparameters to tune, thereby adding another layer of complexity and resource demand. Although both LSTM and GRU are specialized types of RNNs that were developed to address some of the limitations of basic RNNs, the regular RNN showed good performance in handling the last two missing value patterns (completely missing value and hybrid missing value). Additionally, the regular RNN exhibits less complexity in terms of tuning hyperparameters compared with LSTM and GRU.

In relation to the VAR analysis and investigation of EpiEnv variables, the results indicate that pedestrian volumes on all traffic signals in Utah are influenced by EpiEnv variables in at least two ADs. Among all the traffic signals, 63% of them are influenced by average temperature in some ADs. The results of the VAR analysis indicate that pedestrian volume exhibits more changes during colder months, particularly in December and January, as well as in months with higher precipitation, such as March and April. The duration of VAR lag, which represents the number of previous data points used as input variables to model the current data point, is longer during colder months with lower average temperatures. Furthermore, the number of ADs influenced by average temperature tends to increase with a higher number of schools and percentage of residential land use in the vicinity of the traffic signals, while it decreases with an increasing number of transit stops within a quarter-mile buffer from the traffic signals.

AQI is another EpiEnv variable that has an effect on pedestrian volume data for 41% of the traffic signals. Most of these traffic signals are in the northern regions of Utah, specifically Weber and Cache counties. The VAR analysis results indicate that the ADs with pedestrian volume influenced by the AQI are more prevalent in July and August, with relatively low lag periods ranging from one to five days. Additionally, the t-test results ($t(865) = 7.2, p < 0.05$) demonstrate a significant difference in the mean percentage of density in commercial areas within a quarter-mile buffer around traffic signals that are affected by the AQI, compared with those that are not. In other words, the findings suggest that in areas with a higher density of commercial land use, pedestrian activity is more significantly affected by the AQI in Utah.

COVID-19, as an epidemiological variable investigated in this study, demonstrates a significant effect on pedestrian volume across all traffic signals. The majority of the analyzed time periods influenced by COVID-19 were observed during Phase 3, from June 2021 to February 2022. Within this phase, various lags ranging from four to 13 days were identified, indicating the variability in the impact of this variable on pedestrian volume. In relation to the built environment variables, the results of VAR analysis reveal that an increase in population density, the percentage of commercial and residential land uses, and the

number of schools and places of worship corresponds to a greater number of ADs, where pedestrian volume is influenced by COVID-19. Conversely, an increase in vehicle ownership within a quarter-mile buffer around traffic signals is associated with a decrease in the number of time periods where COVID-19 case rates influence pedestrian volume. Another interesting aspect of this EpiEnv variable is the variation in VAR lags in relation to the built environment variables. The results indicate that as the percentage of density of residential land use and the number of schools around the traffic signals increase, the number of lags in the VAR analysis also increases. However, there is no clear relationship observed with respect to commercial land use.

Our investigation affirms that ML and DL techniques provide a substantial leap forward from traditional pedestrian volume data analysis methods. The random forest algorithm, for example, captures complex, nonlinear relationships within the data that manual methods cannot, leading to more nuanced insights and a higher quality of data imputation. Similarly, DL models like LSTM and GRU offer advanced temporal analysis capabilities, enabling us to discern patterns over time with greater precision than is possible with basic statistical methods. While the improvements brought about by these sophisticated methods might be viewed as incremental when compared with their simpler counterparts, they are nonetheless transformative in certain contexts. The precise nature of ML and DL predictions is particularly valuable in urban planning, where the granularity of data can inform critical safety assessments and infrastructure decisions. Even if the quantitative leap in data quality is challenging to measure, the qualitative enhancements—such as improved model responsiveness to dynamic conditions and the ability to handle large, diverse datasets—justify the additional complexity for many applications. In considering the implementation of these methods, it is essential to weigh the benefits of increased accuracy and predictive power against the investment in computational resources and expertise required to operationalize them effectively.

In highlighting the practical applications of push-button data, Singleton, Mekker, and Islam (2021) leveraged such data from traffic signals to estimate pedestrian exposure and examine its relationship with pedestrian safety at signalized intersections. Their approach utilized high-resolution traffic signal controller logs, providing a robust measure of pedestrian exposure. The authors developed regression models to predict pedestrian crossing volumes as a function of push-button data, and then used these volumes as explanatory variables in crash frequency and severity models. This method enabled a nuanced analysis of pedestrian safety and supported the "safety in numbers" hypothesis—the idea that pedestrian crash rates decline with an increase in pedestrian volumes—at signalized intersections. Their study highlights the potential of detailed push-button data for critical safety analyses, indicating broader implications for traffic safety management and policymaking.

4.8 Conclusion and Future Work

Generally, anomalies can have a significant impact on the accuracy of predictive models and analyses that rely on pedestrian volume data. Therefore, it is crucial to detect and handle anomalies (and impute missing/anomalous values) effectively to avoid misleading conclusions and ensure accurate urban planning and pedestrian safety. In this study, we examined various anomaly detection methods, including statistical, ML, and DL approaches, in conjunction with EpiEnv variables. Additionally, we applied ML and DL-based imputation methods to address missing values.

The evaluation results demonstrated that DBSCAN, KNN, and iForest performed well in detecting anomalies, while random forest, LSTM, and GRU showed promising results for imputation across different missing value patterns. Moreover, the investigation of the relationship between EpiEnv variables and significant changes in pedestrian volume, as discovered by VAR analysis, revealed a self-organized pattern between the impact of EpiEnv variables and built environment variables on pedestrian activity. Therefore, for future research, we propose conducting a more detailed examination of these patterns by

incorporating self-organizing maps (SOMs). This method is capable of recognizing walking behavior patterns and, by identifying normal patterns, it provides actionable insights for urban planning. These insights can inform decisions related to strategic infrastructure placement and maintenance scheduling.

In an effort to make our research accessible to a broad audience, including those without a background in advanced computational methods, we have distilled the key aspects of our study into a more digestible format. Our work utilized advanced statistical, ML, and DL techniques to enhance the accuracy of pedestrian volume data, crucial for urban planning and transportation engineering. We essentially employed sophisticated computer algorithms to detect and correct inconsistencies in pedestrian traffic data, thereby providing more reliable information for urban infrastructure decision-making. These advanced methods, while complex and resource-intensive, offer substantial improvements in data precision over traditional methods. By providing more detailed and accurate insights into pedestrian behavior, our approach paves the way for creating safer, more efficient urban spaces. This summary aims to encapsulate the essence of our research, highlighting its practical implications and the benefits it brings to urban planning, making it accessible to a diverse range of stakeholders.

To incorporate the ML methods demonstrated in this study into everyday practice, agencies would require not only the appropriate computational infrastructure but also a level of expertise that may currently be lacking. The automation of these methods presents a viable pathway toward operationalization, particularly as data collection systems become more sophisticated. However, this evolution must be matched by investment in skill development and resources to overcome the initial barriers to implementation. The adoption of these advanced techniques could then represent a significant step forward in the maturation of quality assurance and quality control processes for nonmotorized traffic data.

While our study significantly enhances the process of pedestrian volume estimation at signalized intersections by detecting anomalies and accurately imputing missing data, it opens a pathway for exploring broader applicability and inherent limitations within the transportation network. The potential of our methods to enhance traffic safety and management at these critical junctions is clear, yet the scope of signalized intersections, despite their strategic importance, is limited within the overall network (i.e., there are many more intersections without traffic signals and pedestrian push-buttons). This reality underscores the necessity of extending our analytical frameworks to encompass the full spectrum of urban and suburban traffic environments, including both signalized and non-signalized intersections.

Additionally, as we venture into the domain of hourly data, the challenges increase. Transitioning from daily to hourly volume estimation not only entails a substantial increase in database size but also introduces complexities in data processing and model training. The need for a more robust computational infrastructure and sophisticated data management strategies becomes imperative to handle the surge in data volume efficiently. Moreover, the heightened variability in pedestrian traffic observed on an hourly basis demands the development of complex models capable of accommodating these fluctuations without succumbing to overfitting. This level of detail is crucial for models intended for real-time traffic management applications, where accuracy and timely processing are paramount.

In anticipation of these challenges, further research will need to delve into scalable ML architectures and efficient data streaming processes. This exploration might include leveraging cloud computing resources or distributed computing frameworks for more effective data handling, and adopting incremental learning approaches to update models in real time as new data become available. Additionally, to address the unique challenges presented by unsignalized intersections, we will explore the integration of advanced sensor technologies, alternative data sources such as mobile device signals and video analytics, and geospatial analysis techniques. These approaches aim to capture and predict pedestrian movements more accurately in areas lacking structured pedestrian data. Our perspective on both extending to hourly data and broadening the applicability to unsignalized intersections is one of cautious optimism. Despite the

clear technical hurdles, ongoing advancements in ML and computational power enhance the feasibility of these ambitious goals. Future iterations of this work will investigate these multifaceted challenges, aiming to extend the applicability of our methods to more granular time scales and a broader range of urban traffic contexts, ensuring comprehensive traffic safety and management strategies.

4.9 References

Akaike, H. (1974). "A new look at the statistical model identification." *IEEE Transactions on Automatic Control, 19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Attaset, V., Schneider, R. J., Arnold, L. S., & Ragland, D. R. (2010). "Effects of weather variables on pedestrian volumes in Alameda County, California." Presented at the 89th Annual Meeting of the Transportation Research Board. https://escholarship.org/uc/item/3zn9f4cr

Bai, J. (1995). "Least absolute deviation estimation of a shift." *Econometric Theory, 11*(3), 403-436. https://doi.org/10.1017/S026646660000935X

Bai, S., Kolter, J. Z., & Koltun, V. (2018). "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." https://doi.org/10.48550/arXiv.1803.01271

Banifakhr, M., & Sadeghi, M. T. (2021). "Anomaly detection in traffic trajectories using a combination of fuzzy, deep convolutional and autoencoder networks." *Computer and Knowledge Engineering, 4*(2), 1-10. https://doi.org/10.22067/cke.2021.71379.1018

Beck, M. J., & Hensher, D. A. (2020). "Insights into the impact of COVID-19 on household travel and activities in Australia–The early days under restrictions." *Transport Policy, 96*, 76-93. https://doi.org/10.1016/j.tranpol.2020.07.001

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*(1), 69-79. https://doi.org/10.3141/2527-08

Chung, J., Kim, S. N., & Kim, H. (2019). "The impact of PM10 levels on pedestrian volume: Findings from streets in Seoul, South Korea." *International Journal of Environmental Research and Public Health, 16*(23), 4833. https://doi.org/10.3390/ijerph16234833

Day, C. M., Premachandra, H., & Bullock, D. M. (2011). "Rate of pedestrian signal phase actuation as a proxy measurement of pedestrian demand." Presented at 90th Annual Meeting of the Transportation Research Board, Washington, DC. https://docs.lib.purdue.edu/civeng/24/

de Montigny, L., Ling, R., & Zacharias, J. (2012). "The effects of weather on walking rates in nine cities." *Environment and Behavior, 44*(6), 821-840. https://doi.org/10.1177/0013916511409033

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). "A density-based algorithm for discovering clusters in large spatial databases with noise." *KDD-96 Proceedings, 96*(34), 226-231.

Hanbanchong, A., & Piromsopa, K. (2012). "SARIMA based network bandwidth anomaly detection." JCSSE 2012 - 9th International Joint Conference on Computer Science and Software Engineering, 104–108. https://doi.org/10.1109/JCSSE.2012.6261934

Hautamaki, V., Karkkainen, I., & Franti, P. (2004). "Outlier detection using k-nearest neighbour graph." Proceedings of the 17th International Conference on Pattern Recognition, 3, 430-433. https://doi.org/10.1109/ICPR.2004.1334558

Holmes, A. M., Lindsey, G., & Qiu, C. (2009). "Ambient air conditions and variation in urban trail use." *Journal of Urban Health, 86*, 839-849. https://doi.org/10.1007/s11524-009-9398-8

Hunter, R. F., Garcia, L., de Sa, T. H., Zapata-Diomedi, B., Millett, C., Woodcock, J., ... & Moro, E. (2021). "Effect of COVID-19 response policies on walking behavior in US cities." *Nature Communications, 12*(1), 3652. https://doi.org/10.1038/s41467-021-23937-9

Iowa State University (ISU). "Iowa Environmental Mesonet (IEM) Reanalysis (IEMRE)" (accessed 12 July 2023). https://mesonet.agron.iastate.edu/iemre/

Jackson, K. N., O'Brien, S. W., Searcy, S. E., & Warchol, S. E. (2017). "Quality assurance and quality control processes for a large-scale bicycle and pedestrian volume data program." *Transportation Research Record: Journal of the Transportation Research Board, 2644*(1), 19-29. https://doi.org/10.3141/2644-03

Kothuri, S., Broach, J., McNeil, N., Hyun, K., Mattingly, S., Miah, M. M., ... & Proulx, F. (2022). "Exploring data fusion techniques to estimate network-wide bicycle volumes" (NITC-RR-1269). National Institute for Transportation and Communities. https://doi.org/10.15760/trec.273

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*(1), 11-18. https://doi.org/10.3141/2644-02

Lam, P., Wang, L., Ngan, H. Y., Yung, N. H., & Yeh, A. G. (2017). "Outlier detection in large-scale traffic data by naïve Bayes method and gaussian mixture model method." *Electronic Imaging, 29*, 73-78. https://doi.org/10.2352/ISSN.2470-1173.2017.9.IRIACV-272

Li, D., Chen, D., Jin, B., Shi, L., Goh, J., & Ng, S. K. (2019, September). "MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks." International Conference on Artificial Neural Networks, 703-716. https://doi.org/10.1007/978-3-030-30490-4_56

Lindsey, G., Coll, S., & Stewart, G. (2024). "Quality assurance methods for hourly nonmotorized traffic counts." *Transportation Research Record: Journal of the Transportation Research Board, 2678*(2), 723-742. https://doi.org/10.1177/03611981231175898

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). "Isolation forest." 2008 Eighth IEEE International Conference on Data Mining, 413-422. https://doi.org/10.1109/ICDM.2008.17

Ma, J., & Perkins, S. (2003). "Time-series novelty detection using one-class support vector machines." Proceedings of the International Joint Conference on Neural Networks, 3, 1741-1745. https://doi.org/10.1109/IJCNN.2003.1223670

Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., & Shroff, G. (2016). "LSTM-based encoder-decoder for multi-sensor anomaly detection." https://doi.org/10.48550/arXiv.1607.00148

Nordback, K., Kothuri, S., Petritsch, T., McLeod, P., Rose, E., & Twaddell, H. (2016). "Exploring Pedestrian Counting Procedures" (FHWA-HPL-16-026). Federal Highway Administration. https://rosap.ntl.bts.gov/view/dot/64905

Nordback, K., Tufte, K. A., Harvey, M., McNeil, N., Stolz, E., & Liu, J. (2015). "Creating a national nonmotorized traffic count archive: process and progress." *Transportation Research Record: Journal of the Transportation Research Board, 2527*(1), 90-98. https://doi.org/10.3141/2527-10

Omohundro, S. M. (1989). "Five Balltree Construction Algorithms." UC Berkeley. https://www1.icsi.berkeley.edu/ftp/pub/techreports/1989/tr-89-063.pdf

Park, K., Singleton, P. A., Brewer, S., & Zuban, J. (2023). "Pedestrians and the built environment during the COVID-19 pandemic: changing relationships by the pandemic phases in Salt Lake County, Utah, USA." *Transportation Research Record: Journal of the Transportation Research Board, 2677*(4), 448-462. https://doi.org/10.1177/03611981221083606

Rader, B., Gertz, A., Iuliano, A. D., Gilmer, M., Wronski, L., Astley, C. M., … & Brownstein, J. S. (2022). "Use of at-home COVID-19 tests — United States, August 23, 2021–March 12, 2022." *MMWR. Morbidity and Mortality Weekly Report, 71*(13), 489-494. https://doi.org/10.15585/mmwr.mm7113e1

Rafe, A., & Singleton, P. A. (2023). *PedImpute* (accessed 13 July 2023). https://github.com/pozapas/PedImpute

Runa, F., & Singleton, P. A. (2021). "Assessing the impacts of weather on pedestrian signal activity at 49 signalized intersections in Northern Utah." *Transportation Research Record: Journal of the Transportation Research Board, 2675*(6), 406-419. https://doi.org/10.1177/0361198121994111

Runa, F., & Singleton, P. A. (2023). “Impacts of the COVID-19 pandemic on pedestrian push-button utilization and pedestrian volume model accuracy in Utah.” *Transportation Research Record: Journal of the Transportation Research Board, 2677*(4), 494-502. https://doi.org/10.1177/03611981221089935

Ryus, P., Ferguson, E. M., Laustsen, K. M., Schneider, R. J., Proulx, F. R., Hull, T., & Miranda-Moreno, L. (2014). *Guidebook on Pedestrian and Bicycle Volume Data Collection* (NCHRP Research Report 797). Transportation Research Board. https://doi.org/10.17226/22223

Ryus, P., Musunuru, A., Bonneson, J., Kothuri, S., Monsere, C., McNeil, N., ... & Currin, S. (2022). *Guide to Pedestrian Analysis* (NCHRP Research Report 992). Transportation Research Board. https://doi.org/10.17226/26518

Schneider, R. J., Arnold, L. S., & Ragland, D. R. (2009). “Methodology for counting pedestrians at intersections: Use of automated counters to extrapolate weekly volumes from short manual counts.” *Transportation Research Record: Journal of the Transportation Research Board, 2140*(1), 1-12. https://doi.org/10.3141/2140-01

Schneider, R. J., Henry, T., Mitman, M. F., Stonehill, L., & Koehler, J. (2012). “Development and application of volume model for pedestrian intersections in San Francisco, California.” *Transportation Research Record: Journal of the Transportation Research Board, 2299*(1), 65-78. https://doi.org/10.3141/2299-08

Shaaban, K., & Muley, D. (2016). “Investigation of weather impacts on pedestrian volumes.” *Transportation Research Procedia, 14*, 115-122. https://doi.org/10.1016/j.trpro.2016.05.047

Singleton, P. A., Mekker, M., & Islam, A. (2021). “Safety in Numbers? Developing Improved Safety Predictive Methods for Pedestrian Crashes at Signalized Intersections in Utah Using Push Button-Based Measures of Exposure” (Report UT-21.08). Utah Department of Transportation. https://rosap.ntl.bts.gov/view/dot/56362

Singleton, P. A., Park, K., & Lee, D. H. (2021). “Varying influences of the built environment on daily and hourly pedestrian crossing volumes at signalized intersections estimated from traffic signal controller event data.” *Journal of Transport Geography, 93*, 103067. https://doi.org/10.1016/j.jtrangeo.2021.103067

Singleton, P. A., & Runa, F. (2021). “Pedestrian traffic signal data accurately estimates pedestrian crossing volumes.” *Transportation Research Record: Journal of the Transportation Research Board, 2675*(6), 429-440. https://doi.org/10.1177/0361198121994126

Sobreira, L. T. P., & Hellinga, B. (2023). “Comparing direct demand models for estimating pedestrian volumes at intersections and their spatial transferability to other jurisdictions.” *Transportation Research Record: Journal of the Transportation Research Board, 2677*(10), 260-271. https://doi.org/10.1177/03611981231161061

Tribby, C. P., Miller, H. J., Song, Y., & Smith, K. R. (2013). “Do air quality alerts reduce traffic? An analysis of traffic data from the Salt Lake City metropolitan area, Utah, USA.” *Transport Policy, 30*, 173-185. https://doi.org/10.1016/j.tranpol.2013.09.012

Turner, S., & Lasley, P. (2013). “Quality counts for pedestrians and bicyclists: Quality assurance procedures for nonmotorized traffic count data.” *Transportation Research Record: Journal of the Transportation Research Board, 2339*(1), 57-67. https://doi.org/10.3141/2339-07

US Environmental Protection Agency (US EPA). (2023). *AirNow.gov* (accessed 12 July 2023). https://www.airnow.gov/

Utah Department of Health & Human Services (Utah DHHS). (2023). “Case counts: Coronavirus” (accessed 11 July 2023). https://coronavirus.utah.gov/case-counts/

Utah Department of Transportation (UDOT). (2023). *Automated Traffic Signal Performance Measures* (accessed 4 July 2023). https://udottraffic.utah.gov/atspm/

Wang, X., Lindsey, G., Hankey, S., & Hoff, K. (2014). “Estimating mixed-mode urban trail traffic using negative binomial regression models.” *Journal of Urban Planning and Development, 140*(1), 04013006. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000157

Wang, Y., Li, D., Du, Y., & Pan, Z. (2015). “Anomaly detection in traffic using L1-norm minimization extreme learning machine.” *Neurocomputing, 149*, 415-425. https://doi.org/10.1016/j.neucom.2014.04.073

Xu, L., Taylor, J. E., & Tien, I. (2022). “Assessing the impacts of air quality alerts on micromobility transportation usage behaviors.” *Sustainable Cities and Society, 84*, 104025. https://doi.org/10.1016/j.scs.2022.104025

Yu, H., & Zhang, H. (2023). “Impact of ambient air pollution on physical activity and sedentary behavior in children.” *BMC Public Health, 23*(1), 357. https://doi.org/10.1186/s12889-023-15269-8

5. EVALUATING PEDESTRIAN "SAFETY IN NUMBERS" AT SIGNALIZED INTERSECTIONS IN UTAH WITH PEDESTRIAN EXPOSURE DATA FROM TRAFFIC SIGNALS

This chapter is the accepted manuscript of an article presented at the 100th Annual Meeting of the Transportation Research Board. It is reprinted here with permission from the authors. A revised version of this work (with additional analysis results) was later published in the *Journal of Transportation Engineering, Part A: Systems*. To cite, please use one of these references:

- Islam, A., Singleton, P. A., & Mekker. M. M. (2021). "Evaluating pedestrian 'safety in numbers' at signalized intersections in Utah with pedestrian exposure data from traffic signals." Presented at the 100th Annual Meeting of the Transportation Research Board, Washington, DC.
- Islam, A., Mekker, M., & Singleton, P. A. (2022). "Examining pedestrian crash frequency, severity, and safety in numbers using pedestrian exposure from Utah traffic signal data." *Journal of Transportation Engineering, Part A: Systems, 148*(10), 04022084. https://doi.org/10.1061/JTEPBS.0000737

5.1 Abstract

The focus of this study is twofold: (1) to estimate models of pedestrian crash frequency and severity at signalized intersections using pedestrian and traffic volumes and other predictor variables; and (2) to examine whether the "safety in numbers" effect applies to pedestrian safety in the United States using robust measures of pedestrian exposure. Specifically, the analysis used pedestrian crossing volumes estimated from one year of pedestrian push-button data, and 10 years of crash data at signalized intersections in Utah. Data from 1,606 signalized intersections were used to calibrate a zero-inflated negative binomial model for crash frequency analysis. The model results indicated that signals with longer crossing distances, no prohibitions against turning right on red, more nearby bus stops, and larger shares of vacant land uses saw more pedestrian crashes. To analyze injury severity in pedestrian crashes, an ordered logit model was fitted with 1,572 pedestrian crash observations. The model results indicated that vehicle size, vehicle maneuvering direction, crossing distance, and involvement of DUI/drowsy/distracted driving in crashes had significant effects on severity. The study also found a nonlinear relationship where pedestrian-vehicle crash rates decreased with an increase in pedestrian volumes, supporting the safety in numbers effect. The authors suggest potential countermeasures, policy alterations, and scope of future research for improving pedestrian safety at signalized intersections.

5.2 Introduction

Pedestrian safety is a growing health concern and a critical transportation issue. People face four times higher risk of injury while walking than while driving a car per million kilometers of travel (Elvik, 2009). In the U.S., according to the National Highway Traffic Safety Administration, there were nearly 6,300 pedestrian fatalities, representing about 18% of all traffic fatalities, in 2018 (NHTSA, 2020a). This number was an increase from 4,100 and 12% in 2009 and was higher than in any other year since 1990. On average, 17 pedestrians were killed daily in the U.S., despite an overall decrease in fatal crashes nationwide. Note that around 25% of the total pedestrians killed in the year were at intersections (NHTSA, 2020b).

Given these troubling trends, there is a need for improved pedestrian crash prediction models to better understand factors associated with pedestrian safety and to optimize selection of countermeasures to improve pedestrian safety at signalized intersections. Pedestrian exposure data are vital in the development of such models, as the frequency of pedestrian crashes varies with pedestrian volumes (Tulu

et al., 2015; Harwood et al., 2008). Considerable past research estimated pedestrian volumes based on assumptions about pedestrian travel and incorporated them into safety analyses (Lam et al., 2014; Raford & Ragland, 2006; Tulu et al., 2015). These studies explored physical, social, and environmental characteristics related to pedestrian safety but were limited by the unavailability of accurate pedestrian volume estimation (Raford & Ragland, 2006). Typically, the biggest barrier to overcome for pedestrian safety analysis is the lack of more robust data on pedestrian exposure.

The safety in numbers hypothesis for walking has been examined over the last three decades. This concept suggests that pedestrian (and bicycle) crash rates decrease with increasing volumes of people walking and bicycling. Although research has yet to clearly identify the specific causes of this observed relationship, it is assumed that the more often drivers see pedestrians and bicyclists, the more likely they are to anticipate them and have more experience driving safely around them. As with safety predictive methods, the challenge with studying the safety in numbers concept is the lack of pedestrian exposure data. Most research on the topic was conducted with surrogate measures of pedestrian exposure. For example, for the estimation of pedestrian volumes, researchers have taken a "space syntax" modeling approach (Geyer et al., 2006; Raford & Ragland, 2006), used travel characteristics survey data (Jacobsen, 2015), and generated random numbers (Elvik, 2013). An authentic dataset on pedestrian exposure would be more reliable for understanding whether the safety in numbers concept applies to pedestrian safety, knowledge that could promote more walking and bicycling through policy and planning.

The primary objective of this study was to calibrate models with actual pedestrian and traffic volumes and other predictor variables (including road network characteristics) to investigate their relationships with the frequency and severity of pedestrian crashes at signalized intersections. A second objective was to examine whether the safety in numbers phenomenon is observed after the inclusion of actual pedestrian exposure data. For both of these objectives, the authors used a novel data source to measure pedestrian exposure: annual average pedestrian crossing volumes as estimated using push-button-based pedestrian data from traffic signals.

5.3 Literature Review

5.3.1 Factors Affecting Pedestrian Crash Frequency

For the improvement of pedestrian safety at intersections, a detailed exploration of crash-related factors is required to develop effective countermeasures (Lee & Abdel-Aty, 2005; Stutts et al., 1996). Factors studied in the past regarding pedestrian crashes include traffic exposure, built environment characteristics, socio-demographic characteristics, site specific characteristics, and other spatial variables.

Exposure is typically operationalized using average volumes of traffic. Several studies found positive associations between vehicle volume and pedestrian crashes (Cotrill & Thankuriah, 2010; El-Basyouny & Sayed, 2013; Harwood et al., 2008). But only a few studies explored the link with pedestrian volumes due to difficulty in obtaining such data. When included, the volume of pedestrians was the single most important variable to explain variations in pedestrian crashes (Brüde & Larsson, 1993; Lyon & Persaud, 2002; Zegeer et al., 1985). Overall, both pedestrian and vehicular volumes show positive associations with pedestrian-vehicle crashes (Harwood et al., 2008; Xu et al., 2019; Yasmin & Eluru, 2016).

Built environment characteristics include population, job density, and local land use types. Confoundingly, population density showed both a positive (Dumbaugh & Li, 2010; Gladhill & Monsere, 2012) and negative (Loukaitou-Sideris et al., 2007; Graham & Glaister, 2003) association with pedestrian crash occurrence in different studies. Job or employee density was found to be positively associated with pedestrian crashes (Loukaitou-Sideris et al., 2007). Increased proportions of land used for commercial,

mixed use, park, retail, or community use has been associated with increased vehicle-pedestrian collisions (Loukaitou-Sideris et al., 2007; Wier et al., 2009).

Examples of socio-demographic characteristics are household income, population by age, race/ethnicity, and number of children. One study found a relationship between pedestrian crashes and population demographics such as income and the presence of children in households (Cotrill & Thankuriah, 2010). Children and the elderly are more at risk as they take a longer time to cross the road, increasing their exposure to motor vehicle traffic (Demetriades et al., 2004).

Different road and intersection characteristics—including number of lanes, signal types, and lighting conditions—have been investigated. A greater number of lanes was related to higher pedestrian crash frequency, whereas speed limit, crosswalk marking conditions, and crosswalk marking types had no significant effect on pedestrian crash rates (Zegeer et al., 2005). Pedestrian crash risk was observed to be reduced by improved lighting conditions (Lee & Abdel-Aty, 2005).

5.3.2 Factors Affecting Pedestrian Crash Severity

As pedestrians are more likely (than other road users) to be injured or killed when involved in crashes, identifying factors contributing to pedestrian crash severity is essential for selection of appropriate countermeasures (Haleem et al., 2015). Demographic characteristics of pedestrians or drivers showed significant associations to crash risk in several studies, with age standing out as a particularly important predictor of the crash severity. Lee and Abdel-Aty (2005) suggested that elderly or alcohol-impaired pedestrians risked higher injury severity when involved in crashes. Haleem et. al. (2015) observed that the involvement of elderly and pedestrians younger than 15 years of age increased the likelihood of fatal crashes. Sarkar et. al. (2011) found that male and elderly pedestrians were more likely to have severe injuries than other population groups when involved in crashes.

Vehicle characteristics and conditions, including vehicle size, speed, and trajectory/action, have also been related to pedestrian crash severity. Pedestrians involved in crashes with vehicles larger than passenger cars experienced higher injury severity (Lee & Abdel-Aty, 2005). Oh et. al. (2005) identified collision speed as the most significant factor, where higher speed was associated with increased likelihood of pedestrian fatality. Roudsari et. al. (2006) found that a straight-moving vehicle hitting a pedestrian increased the injury severity and the chance of fatality.

A few studies included roadway geometry, traffic volume, and environmental conditions for investigation of pedestrian crash severity. Haleem et. al. (2015) included all of these factors when investigating pedestrian crash severity at intersections. At signalized intersections, they found that higher average annual daily traffic (AADT), rain, and dark conditions were significant predictors of pedestrian crash severity. Zajac and Ivan (2003) found that rural, downtown fringe, and low-density residential areas experienced more severe pedestrian crashes than downtown, compact residential, and medium- to low-density commercial areas. Mohammed et al. (2013) demonstrated that the prevalence of mixed land use increased the probability of fatal pedestrian crashes.

5.3.3 Safety in Numbers

Although a positive relationship was found between pedestrian/bicycle crash frequency and measures of exposure, researchers have argued that it is a nonlinear relationship. Specifically, they suggest that while there is a steeper increase in crash rate at lower levels of pedestrian/bicycle traffic, the crash rate does not increase proportionately and actually becomes smaller with higher levels of pedestrian/bicycle traffic. This phenomenon is popularly known as the safety in numbers concept (Carlson et al., 2018; Elvik et al., 2013; Jacobsen, 2015). Elvik and Goel (2019) stated that the risk of injury to each pedestrian and cyclist

becomes lower with a greater number of pedestrians and cyclists. Initially, Elvik (2013) had proposed an opposing concept called "hazard in numbers," which suggests that the number of crashes doubles when traffic volume is doubled. He proved that hazard in numbers can co-exist with safety in numbers in a dataset of pedestrians, bicyclists, or motorists. Later, in a meta-analysis of estimates, Elvik and Goel (2019) reported that although there is considerable variation in estimates, nearly all studies support safety in numbers. It was also found that the safety in numbers effect for pedestrians is stronger than for cyclists or motorists, and newer investigations support safety in numbers more than earlier studies.

5.3.4 Summary of Literature Review

Most research on pedestrian safety has been limited by the unavailability of pedestrian exposure data. The few studies that included pedestrian exposure used a surrogate measure. Additionally, the studies that examined the effect of explanatory variables on pedestrian crashes mostly ignored the characteristics of different facilities used by pedestrians in the analysis. Also, studies on the safety in numbers concept are mostly limited to European settings. This study addresses several of these limitations by:

- Incorporating stronger measures of pedestrian exposure
- Including key intersection variables
- Examining whether the "safety in numbers" concept applies to pedestrian safety in the U.S.

5.4 Data

This study investigated pedestrian crashes that occurred over 10 years, from 2010 through 2019, at signalized intersections in Utah. The three different datasets used for the analysis—traffic signals and intersection data, pedestrian crash data, and pedestrian exposure data—are briefly described in this section.

5.4.1 Traffic Signals and Intersection Data

At the time of this study, there were 2,214 traffic signals in use across Utah. Among those, 1,606 signals—excluding pedestrian-activated flashers, pedestrian hybrid beacons, signals without pedestrian push-buttons, and signals not connected to the central network—were included in this study.

As one of the objectives of this study was to identify intersection and road characteristics that are directly related to pedestrian crash frequency and severity at signalized intersections, detailed data regarding different features at selected sites were gathered using Google Earth and Street View. In addition to measuring crosswalk distances, crosswalk presence and marking types (standard, continental, ladder, or zebra [Harkey & Zegeer, 2004]) were also recorded. The presence of inbound and outbound bike lanes at the intersections were identified, as well as whether a near-side or far-side bus stop was located within 300 feet of the intersection.

The study dataset also included vehicle exposure data, land use and built environment characteristics, and socioeconomic characteristics of the surrounding area. These independent variables were calculated for within a quarter-mile of intersections using data from a variety of state and national sources, including the Utah Geospatial Resource Center and the US Census Bureau.

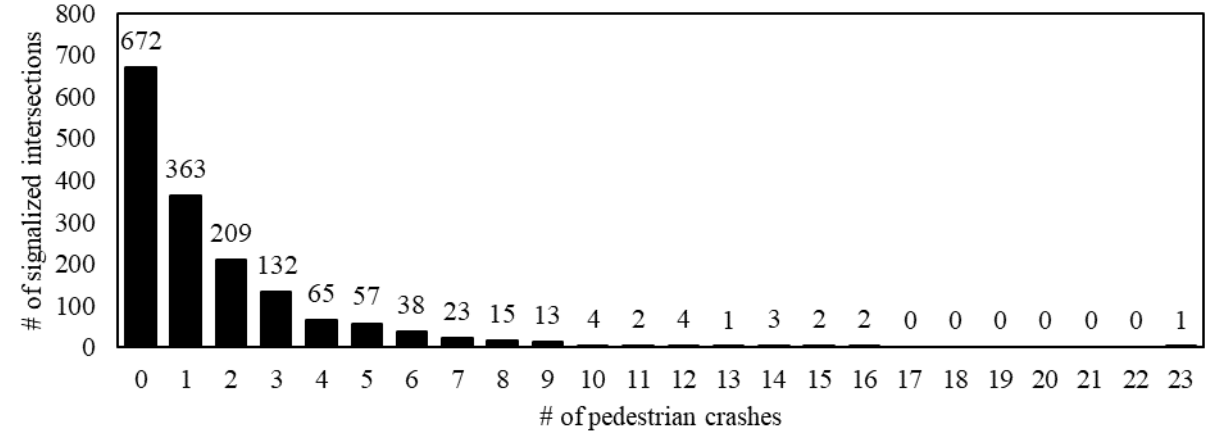
5.4.2 Pedestrian Crash Data

Crash data were obtained from the Utah Department of Transportation (UDOT). Each crash record includes information on temporal characteristics, spatial characteristics, contributing factors, crash

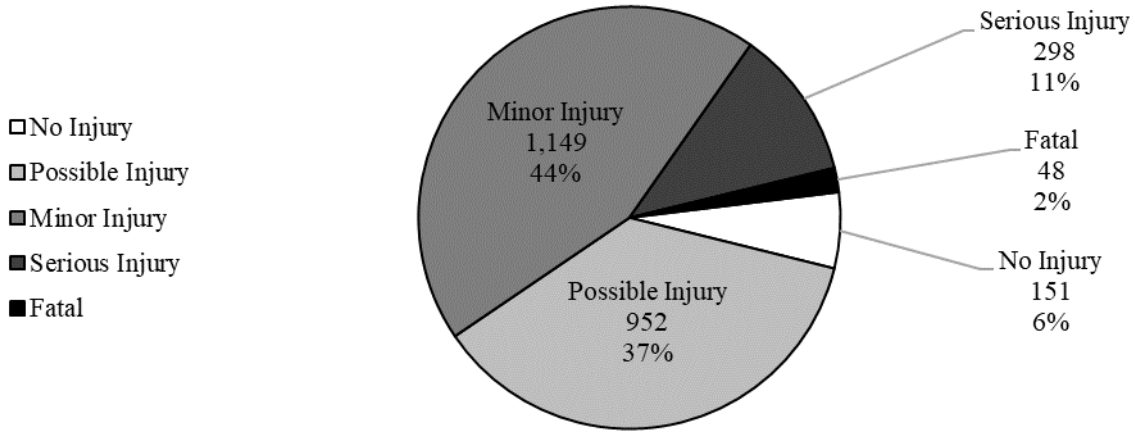
severity, weather conditions, and crash participants. This information was extracted from police crash reports, and no personally identifying information was included. There were 2,939 observed pedestrian-involved crashes at (or related to) signalized intersections from 2010 through 2019.

For analyzing pedestrian crash frequency, pedestrian crashes that occurred during the study period were assigned to the nearest signal location based on the longitude/latitude data of the crash location. Of the 1,606 study intersections, a plurality (42%) had zero pedestrian crashes during the study period. Figure 5.1a shows the distribution of pedestrian crash frequency.

For pedestrian crash severity analysis, additional information, such as environmental characteristics, crash characteristics, vehicle characteristics, and driver characteristics, were available in the crash database. Each pedestrian crash was designated by one of five injury severity levels: no injury, possible injury, minor injury, serious injury, and fatal. Figure 5.1b shows the distribution of pedestrian crash severities.



(a) Pedestrian crash frequencies



(b) Pedestrian crash severities

Figure 5.1 Distributions of dependent variables

5.4.3 Pedestrian Exposure Data

The pedestrian exposure data came from high-resolution traffic signal controller logs (Sturdevant et al., 2012). UDOT archives every controller event at almost all traffic signals in the state through the Automated Traffic Signal Performance Measures (ATSPM) system. If a traffic signal included walk indications and pedestrian detection (usually push-buttons), data regarding push-button presses and walk phases were available. Although pedestrian traffic signal data are not perfect measures of pedestrian volumes, recent work by Singleton et al. (2020) has demonstrated that such data can be used to predict pedestrian crossing volumes at signalized intersections with relative accuracy. They developed simple regression models predicting hourly pedestrian crossing volumes as a function of pedestrian signal data (detailed model results are available from the authors). Over more than 22,500 hours of data, the correlation between observed and model-predicted hourly pedestrian crossing volumes was 0.84, with a mean absolute error of only 3.0 (Singleton et al., 2020).

For this study, one year (July 2017 through June 2018) of pedestrian data was obtained from traffic signals in Utah. After cleaning the data, five regression models developed by Singleton et al. (2020) were applied to the pedestrian signal data to estimate the annual average daily pedestrian (AADP) crossing volume at each signal.

5.4.4 Descriptive Statistics

Descriptive statistics of the dependent and all independent variables for the pedestrian crash frequency analysis are presented in Table 5.1. Descriptive statistics for continuous and categorical independent variables for the pedestrian crash severity analysis are shown in Table 5.2.

Table 5.1 Descriptive statistics of variables in the frequency analysis

Variable	*Min.*	*Max.*	*Mean*	*Std. Dev.*
Dependent variable, frequency model				
# of pedestrian-involved crashes	0	23	1.62	2.32
Measures of exposure				
Annual average daily pedestrian volume (AADP)	0.16	6,737	269.95	572.78
Average daily traffic in major direction ($AADT_{MAJ}$)	450	186,000	23,312.09	12,900.82
Average daily traffic in minor direction ($AADT_{MIN}$)	0	57,000	8565.02	7,789.45
Transportation characteristics				
Presence of overhead street lighting	0	1	0.97	0.16
Intersection type				
2-leg (mid-block)	0	1	0.00	0.06
3-leg	0	1	0.09	0.29
4-leg	0	1	0.87	0.33
5-leg	0	1	0.00	0.04
Diverging diamond interchange (DDI)	0	1	0.00	0.07
Single point urban interchange (SPUI)	0	1	0.02	0.14
# crosswalks, total	0	4	3.45	0.96
# crosswalks with standard markings	0	4	3.14	1.17
# crosswalks with continental markings	0	4	0.27	0.71
# crosswalks with ladder, zebra, or other markings	0	3	0.01	0.11
# crosswalks with continental, ladder, or zebra markings	0	4	0.29	0.72
Crosswalk length (mean, ft)	20	185	81.83	19.89
# approaches with no pedestrian crossing	0	4	0.44	0.83
# approaches with no right-turn-on-red	0	1	0.01	0.12
# approaches with channelized right turns	0	4	0.20	0.69
# approaches with bike lanes	0	4	0.59	1.03
# of bus stops within 300 ft of intersection	0	6	0.93	1.18
# approaches with near-side bus stops	0	4	0.31	0.60
# approaches with far-side bus stops	0	4	0.62	0.89
Intersection density (# per mi^2) [a]	6.07	313.17	97.66	49.12
Land use and built environment characteristics [a]				
% land use residential	0	84	31	23.51
% land use commercial	0	92	28	20.75
% land use industrial	0	83	2.41	10.51
% land use vacant	0	100	4.54	8.74
Population density (1,000 per mi^2)	0.08	23.51	4.51	3.02
Employment density (1,000 per mi^2)	0.02	216.03	7.30	11.51
Park area (acre)	0	37.15	1.45	3.61
# of schools	0	5	0.31	0.61
# of places of worship	0	6	0.51	0.78
Sociodemographic characteristics [a]				
Household income (median, $1,000)	20.5	144.61	61.33	21.87
Vehicle ownership (mean)	0.55	3.00	1.81	0.45
Household size (mean)	1.41	13.72	3.11	0.85
% of the population with a disability	2.51	27.06	10.64	4.12
% of the population of Hispanic or non-white race/ethnicity	0.00	75.66	17.26	13.50

[a] These variables were measured using a quarter-mile network buffer.

Table 5.2 Descriptive statistics of independent variables in the severity analysis

Category	*Variable name*	*Summary statistics*	
Categorical independent variables		*Frequency*	*Percentage*
Lighting condition	Lighted	1,524	59%
	Poorly lighted	898	35%
	Unlighted	152	6%
Weather condition	Clear	1,952	76%
	Cloudy or foggy	347	14%
	Precipitation	261	10%
Vehicle classification by body type	Small (passenger cars)	1,311	54%
	Medium (van/SUV/pickup)	1,063	44%
	Large (bus/truck/tractor/RV)	46	2%
Roadway surface condition	Dry	2,212	87%
	Wet	345	13%
Crash involving…	More than 1 vehicle	125	5%
	Disregarding traffic control device	89	3%
	DUI, distraction, or drowsy driving	190	7%
	Improper/unrestrained driver	50	2%
	Older/teenage driver	492	19%
Vehicle movement	Turning left	816	34%
	Turning right	953	39%
Functional class of road	Arterial	1,665	65%
	Collector	261	10%
	Local	666	25%
Road alignment	Horizontal alignment: curve	25	1%
	Vertical alignment: grade	163	6%
Continuous independent variables		*Mean*	*Std. Dev.*
Measures of exposure	Annual average daily pedestrian volume (AADP)	493.00	726.80
	Average daily traffic in major direction ($AADT_{MAJ}$)	27,408.36	11,661.46
	Average daily traffic in minor direction ($AADT_{MIN}$)	12,015.11	9,206.43
Intersection Characteristics	Crosswalk length (mean, ft)	86.96	19.08
	# approaches with marked crosswalk	3.78	0.60
	# approaches with no pedestrian crossing	0.16	0.49
	# approaches with no right-turn-on-red	0.01	0.10
	# approaches with channelized right turns	0.12	0.53
	# approaches with bike lanes	0.59	1.04
	# of bus stops within 300 ft of intersection	1.55	1.36
	# approaches with near-side bus stops	0.47	0.72
	# approaches with far-side bus stops	1.08	1.10
	Intersection density (# per mi^2) [a]	105.56	44.11
Land use and built environment characteristics [a]	% land use residential	31.06	21.57
	% land use commercial	34.32	19.36
	% land use industrial	1.40	6.10
	% land use vacant	3.41	5.56
	Population density (1,000 per mi^2)	5.63	2.74
	Employment density (1,000 per mi^2)	8.60	11.27
	Park area (acre)	1.60	3.65
	# of schools	0.34	0.63
	# of places of worship	0.56	0.83
Sociodemographic characteristics [a]	Household income (median, $1,000)	53.80	17.67
	Vehicle ownership (mean)	1.59	0.40
	Household size (mean)	2.94	0.90
	% of the population with a disability	11.58	4.09
	% of the population of Hispanic or non-white race/ethnicity	20.87	13.76

[a] These variables were measured using a quarter-mile network buffer.

5.5 Methods

This study used two different modeling approaches to evaluate the effect of various explanatory variables on the frequency and severity of pedestrian crashes at signalized intersections. For pedestrian crash frequency, a set of count data models were developed and compared, culminating in a zero-inflated negative binomial (ZINB) model. For pedestrian crash severity, an ordered logit model was fitted.

5.5.1 Pedestrian Crash Frequency Modeling

Like most crash frequency data, the pedestrian crash frequency data used in this study were discrete, random, and non-negative. The modeling framework of generalized linear models (GLMs) are more suited to such count data than ordinary linear regressions, which can predict negative, non-integer values of the dependent variable. The Poisson regression model has been widely used as a starting point to model count data (Lord & Mannering, 2010), but it requires the variance of the count data to equal the mean. When count data used are over-dispersed (i.e., the mean is less than the variance), a negative binomial (NB) regression model is usually more appropriate for the dataset. An additional error term allows the variance to be different from the mean of the dataset. Although this model yielded a more accurate hypothesis test, it did not account for the excess zeros in the dataset.

As mentioned earlier, there were no pedestrian crashes during the study period at a plurality of the signalized intersections. Hence, the adoption of ZINB was plausible as it can accommodate overdispersion arising from both unobserved heterogeneity and excess zeros (Miranda-Moreno & Fu, 2006). The probability density function for the ZINB model is as follows:

$$P(Y = y_{it}) = \begin{cases} P_{it} + (1 - P_{it})\frac{1}{(1+\alpha\mu_{it})^{\frac{1}{\alpha}}} & y_{it} = 0 \\ (1 - P_{it})\frac{\Gamma\left(y_{it}+\left(\frac{1}{\alpha}\right)\right)}{\Gamma(y_{it}+1)\Gamma\left(\frac{1}{\alpha}\right)}\frac{(\alpha\mu_{it})^{y_{it}}}{(1+\alpha\mu_{it})^{y_{it}+\left(\frac{1}{\alpha}\right)}} & y_{it} > 0 \end{cases} \quad (4)$$

where α is dispersion parameter and Γ is gamma function for the ZINB model.

Since the criteria to compare and select appropriate models depends on the presence and the source of overdispersion in the crash data, the likelihood ratio test can be used to check for the existence of overdispersion (Isgin et al., 2008). The Poisson and zero-inflated Poisson (ZIP) models are nested within the NB and ZINB models, respectively, while performing the test. The Vuong test can be used to examine the contribution of excess zeros in overdispersion (Vuong, 1989). The test also compares the zero inflated models with single count models (Poisson and NB). When the value of the test is significant for the Poisson-based models, it indicates that only zero counts contribute to overdispersion, and hence, ZIP is more appropriate than the single Poisson model (Hosseinpour et al., 2013). When the value of the Vuong test is significant in the case of the NB-based model, it indicates that both excess zero and heterogeneity account for overdispersion.

5.5.2 Pedestrian Crash Severity Modeling

The study also aimed to identify the factors that contribute to injury severity in pedestrian crashes. While the intersection is the unit of analysis in crash frequency models, each crash is typically analyzed for crash severity models. In this case, the dependent variable is categorical and ordered (i.e., from no injury to fatal injury). An appropriate technique to model these data is the ordered probit or ordered logit

models, which assume that there is some underlying continuous version of the ordinal/categorical dependent variable. In light of this, an ordered logit model was used in this study.

The ordered logit can be estimated using several open-source software packages. The specification of an ordered logit model is as follows:

$$y_i^* = \beta' x_i + \varepsilon_i \quad (5)$$

where y_i^* is the predicted level of injury severity by a pedestrian i, β' is a vector of unknown parameters, x_i is a vector of explanatory variables, and ε_i is the random error term that follows a standard logistic distribution. The classification of observed injury severity is done based on the predicted injury using the following criteria:

$$y_i = \begin{cases} 0 \; if \; y_i^* \leq 0 \; (\text{no} - \text{injury}) \\ 1 \text{ if } 0 \leq y_i^* \leq \mu_1 \; (\text{possible injury}) \\ 2 \text{ if } \mu_1 \leq y_i^* \leq \mu_2 \; (\text{minor injury}) \\ 3 \text{ if } \mu_2 \leq y_i^* \leq \mu_3 \; (\text{major injury}) \\ 4 \text{ if } \mu_3 \leq y_i^* \leq \mu_4 \; (\text{fatal}) \end{cases} \quad (6)$$

where μ_1, μ_2, and μ_3 are the thresholds estimated by the model.

5.6 Results

5.6.1 Pedestrian Crash Frequency

The ZINB model was deemed most appropriate (as discussed above) and showed a decent fit (McFadden's pseudo-R^2 = 0.327) in the dataset of 1,038 signalized intersections. (Some signals were removed due to missing or no pedestrian exposure data or lack of intersection data.) Table 5.3 indicates the estimated parameters of the ZINB model, first from the count portion, followed by the zero portion. We will focus on interpreting the count model portion of the ZINB model. Variables with positive coefficients suggest higher crash frequencies while negative parameter values suggest lower crash frequencies relative to the base case.

Table 5.3 ZINB model results for pedestrian crash frequency (N = 1,038)

Variables	*B*	*SE*	*z*	*p*
Negative binomial portion				
(Intercept)	-6.8573	0.6995	-9.804	0.000
Measures of exposure				
Annual average daily pedestrian volume, estimated (AADP) [a]	0.4005	0.0387	10.352	0.000
Annual average daily traffic, major approaches ($AADT_{MAJ}$) [a]	0.4063	0.0722	5.624	0.000
Annual average daily traffic, minor approaches ($AADT_{MIN}$) [a]	0.0607	0.0212	2.866	0.004
Transportation system characteristics				
Intersection type (ref. = 4-leg)				
2-leg (mid-block)	-1.2396	0.7981	-1.553	0.120
3-leg	-0.2217	0.1507	-1.472	0.141
5-leg	-0.4915	0.5316	-0.925	0.355
Diverging diamond interchange (DDI)	-1.0314	1.0947	-0.942	0.346
Single point urban interchange (SPUI)	-0.5658	0.4457	-1.269	0.204
# crosswalks with continental, ladder, or zebra markings	0.1157	0.0360	3.219	0.001
Crosswalk length, mean (ft)	0.0041	0.0018	2.230	0.026
# approaches with no right-turn-on-red	-0.4995	0.2694	-1.854	0.064
# approaches with bike lanes	-0.0775	0.0288	-2.692	0.007
# of bus stops within 300 ft of intersection	0.1060	0.0237	4.472	0.000
Land use and built environment characteristics				
% land use vacant [b]	0.0099	0.0055	1.813	0.070
Employment density (1,000 per mi^2) [b]	-0.0099	0.0031	-3.176	0.002
Sociodemographic characteristics				
% of population with a disability [b]	0.0208	0.0079	2.648	0.008
% of population of Hispanic or non-white race/ethnicity [b]	0.0127	0.0025	5.007	0.000
Zero-inflated portion				
(Intercept)	4.0533	0.8469	4.786	0.000
Annual average daily pedestrian volume, estimated (AADP) [a]	-0.9666	0.2167	-4.462	0.000
Population density (1,000 per mi^2) [b]	-0.8187	0.1769	-4.627	0.000
% of population of Hispanic or non-white race/ethnicity [b]	0.0517	0.0169	3.062	0.002

[a] The natural log of these variables (+1) entered the model.
[b] These variables were measured using a quarter-mile network buffer.

The results suggested that pedestrian volume (AADP) and both major and minor leg traffic volume (AADT) were significantly associated with crashes. Pedestrian-vehicle collisions occurred more frequently at signalized intersections where the volumes of pedestrian and motor vehicle traffic were higher. The finding is consistent with the existing literature, which suggests that both the pedestrian and vehicular traffic exposure show positive associations with pedestrian-vehicle crashes *(5, 20)*. An increase in vehicle volumes by 10% on each of the major and minor legs would be expected to increase the number of pedestrian crashes by 4.0% and 0.6%, respectively. Among the road network characteristics examined, the mean crosswalk distance was found to be significantly associated with pedestrian crash frequency: i.e., intersections with greater crosswalk distances had slightly more pedestrian crashes. A 12-foot increase in mean crosswalk distance was associated with a 5% increase in crash frequency. Also, intersections with more nearby bus stops saw more pedestrian crashes. Notably, signalized intersections where right turns on red were prohibited had fewer pedestrian crashes than would otherwise be expected. Among the built environment characteristics, pedestrian crashes were more frequent in areas with larger shares of vacant land uses. While considering socio-demographic characteristics, there were more pedestrian crashes in neighborhoods with a greater share of people with disabilities and in areas with more people of Hispanic or non-White race/ethnicity. Specifically, neighborhoods with 1% more people with disabilities or Hispanic/non-White populations would be predicted to have 1% to 3% more pedestrian crashes.

5.6.2 Pedestrian Crash Severity

Table 5.4 lists the estimation results of the ordered logit model for pedestrian crash severity. The model was fitted with a dataset consisting of 1,573 pedestrian crashes—observations were removed due to missing data—and had a good fit overall (McFadden's pseudo-$R^2 = 0.38$).

The results indicated that involvement of large and medium size vehicles significantly increased severity. In comparison with crashes involving small vehicles, large vehicles were associated with a 156% increase in the odds of more severe injuries, while medium size vehicle increased the chances of a more severe injury by 36%. When left- and right-turning vehicles were involved in pedestrian-vehicle collisions, the odds of a more severe crash decreased by 44% and 64%, respectively, with respect to vehicles moving straight through the intersection. Results also indicated that involvement of an older or teenage driver in a crash was associated with more severe pedestrian crashes at signalized intersections (an increase of almost 22% compared with crashes involving drivers of other ages). Involvement of DUI, drowsy, or distracted driving was found to increase the probability of more severe crashes involving pedestrians by about 160%. Compared with crashes in good light conditions, crashes in poorly lighted or unlighted conditions were associated with a 34% increase in the odds of more severe injuries. None of the variables related to land use or built environment characteristics were associated with pedestrian crash severity. Among other variables, pedestrian crashes at locations with horizontal curves, more near-side bus stops, and in areas with more people of Hispanic or non-White race/ethnicity were generally less severe. Pedestrian crashes at intersections with more approaches having pedestrian crossings were more severe. Neither pedestrian volume nor traffic volume in major direction showed any association with pedestrian crash severity at signalized intersections.

Table 5.4 Ordered logit model results for pedestrian crash severity (N = 1,572)

Variable		*B*	*SE*	*t*	*P*
Vehicle and driver attributes					
Vehicle body type:	**Large (bus/truck/tractor/RV)**	**0.940**	**0.38**	**2.46**	**0.014**
	Medium (van/SUV/pickup)	**0.310**	**0.10**	**3.12**	**0.002**
Crash involving:	**DUI, distraction, or drowsy driving**	**0.954**	**0.20**	**4.83**	**0.000**
	Disregarding traffic control device	0.145	0.28	0.51	0.609
	Improper/unrestrained driver	-0.239	0.38	-0.63	0.531
	Older/teenage driver	**0.202**	**0.12**	**1.65**	**0.099**
Vehicle movement:	**Turning left**	**-0.580**	**0.13**	**-4.36**	**0.000**
	Turning right	**-1.022**	**0.13**	**-7.86**	**0.000**
Environmental characteristics					
Lighting condition:	**Poor or unlighted**	**0.291**	**0.10**	**2.78**	**0.005**
Weather condition:	Cloudy or foggy	-0.003	0.15	-0.02	0.983
	Precipitation	0.080	0.30	0.26	0.792
Surface condition:	Wet	-0.151	0.27	-0.57	0.569
Roadway characteristics					
Functional class:	Arterial	-0.002	0.11	-0.01	0.989
	Collector	-0.163	0.18	-0.90	0.369
Horizontal alignment: Curve		**-0.361**	**0.03**	**-12.82**	**0.000**
Vertical alignment:	Grade	-0.028	0.21	-0.13	0.896
Transportation characteristics					
Annual average daily pedestrian volume (AADP)		0.000	0.00	0.29	0.772
Average daily traffic, major direction ($AADT_{MAJ}$) (1,000s)		0.002	0.01	0.44	0.661
Average daily traffic, minor direction ($AADT_{MIN}$) (1,000s)		**-0.013**	**0.01**	**-1.86**	**0.062**
Crosswalk length (mean, ft)		**0.008**	**0.00**	**2.35**	**0.019**
Speed limit (mph)		0.004	0.01	0.40	0.686
# of bus stops		0.022	0.02	1.29	0.199
Intersection density (# per mi^2)		-0.001	0.00	-0.62	0.533
# approaches with pedestrian crossing		**0.147**	**0.07**	**2.06**	**0.040**
# approaches with no pedestrian crossing		0.073	0.21	0.35	0.730
# approaches with markings		-0.092	0.08	-1.19	0.233
# approaches with near-side bus stops		**-0.121**	**0.07**	**-1.78**	**0.076**
# approaches with far-side bus stops		-0.007	0.05	-0.14	0.886
# approaches with bike lanes (inbound)		-0.365	0.31	-1.19	0.235
# approaches with bike lanes (outbound)		0.430	0.31	1.40	0.162
# approaches with channelized right turn		0.077	0.12	0.64	0.520
Land use and built environment characteristics					
% land use residential		0.004	0.01	0.61	0.545
% land use commercial		-0.002	0.01	-0.33	0.741
% land use industrial		0.006	0.01	0.55	0.580
% land use vacant		-0.002	0.01	-0.22	0.828
Park area (acre)		-0.006	0.01	-0.43	0.667
# of schools		-0.105	0.08	-1.34	0.180
Sociodemographic characteristics					
Household income (median, $1,000)		-0.002	0.00	-0.50	0.614
Vehicle ownership (mean)		-0.028	0.16	-0.18	0.857
Household size (mean)		-0.061	0.07	-0.93	0.351
% of the population with a disability		0.001	0.02	0.07	0.943
% of the population of Hispanic or non-white race/ethnicity		**-0.009**	**0.00**	**-2.14**	**0.032**

Bold = $p < 0.10$, Regular = $p > 0.10$.

5.7 Discussion

The unique use of robust measures of pedestrian exposure estimated from traffic signal data, especially in the frequency model, allows our study to provide stronger insights into the safety in numbers concept for pedestrians at U.S. signalized intersections. Specifically, we find strong support of a safety in numbers effect for pedestrians: a doubling (10% increase) in pedestrian crossing volumes would be predicted to only increase crash frequency by around 4%. In other words, pedestrian crashes increase less than half as fast as pedestrian volumes, thus leading to reduced crash rates (on a per-person basis) as pedestrian volumes increase. Figure 5.2 depicts this relationship, where pedestrian crash rates (frequency/exposure) decline with increasing pedestrian volumes.

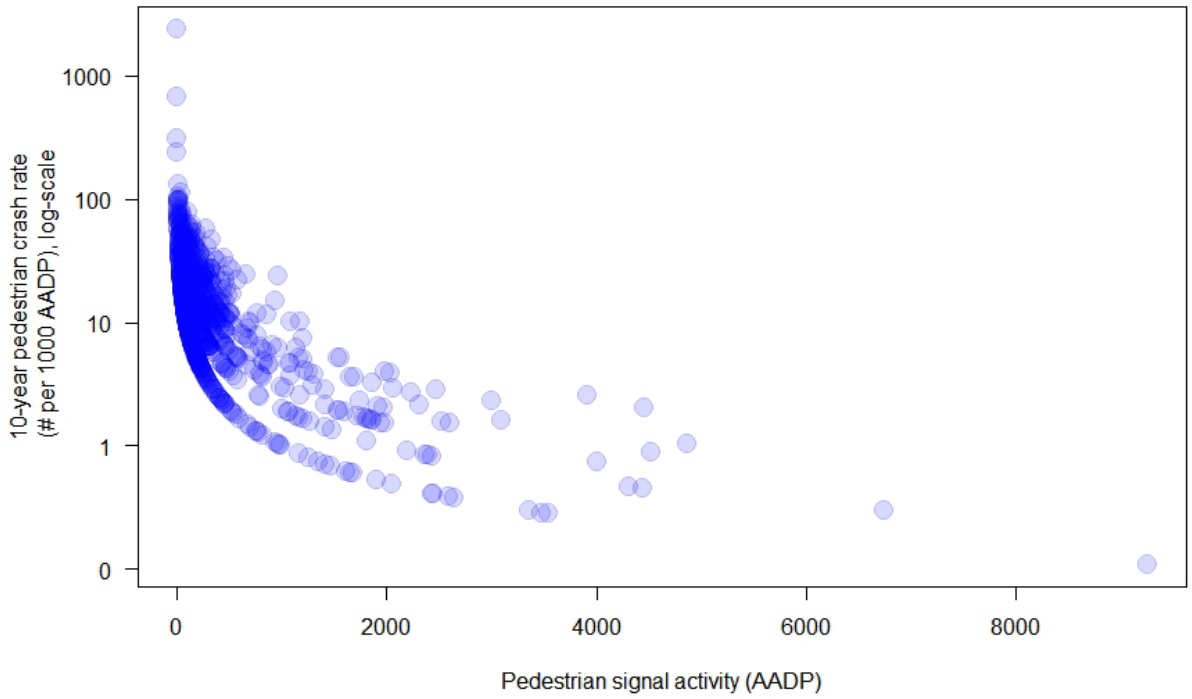


Figure 5.2 Demonstration of the "safety in numbers" effect for pedestrians at signals

Our study was not without limitations that could be addressed through future work. Due to their small sample size and different operations, we excluded pedestrian activated flasher and pedestrian hybrid beacon signals from the analysis. Separate (and larger-scale) analysis for these signals may yield useful insights related to pedestrian safety. Also, the built environment, sociodemographic, road characteristics, and pedestrian volume data were collected for a single time point or year, while models included crashes over a 10-year period. Factors such as household income, land use types, crosswalk marking/type/distance, the location of bus stops, or pedestrian volumes may have changed slightly (or even significantly) over the study period. Future work on the topic may consider using multiyear data of predictor variables for a more comprehensive analysis.

5.8 Conclusion

The objective of this study was to develop models of pedestrian crash frequency and severity based on traffic signal pedestrian push-button data and other factors, as well as to investigate the safety in numbers concept for pedestrians in the U.S. Data from 1,038 signalized intersections and 1,572 pedestrian-involved crashes at those intersections over 10 years were analyzed. A zero-inflated negative binomial model was used to estimate pedestrian crash frequency and an ordered logit model was used to predict pedestrian crash severity. Overall, model results agreed with past research findings. Key conclusions from this study include:

- The results indicated a strong safety in numbers effect on pedestrian crash occurrence, showing that the number of pedestrian crashes increased by only 40% when pedestrian volume doubled.
- At signalized intersections with longer crosswalks, no prohibitions against turning right on red, higher numbers of transit stops, and larger shares of vacant land uses, the frequency of pedestrian crashes was higher.
- Crashes involving large- or medium-size vehicles; involving DUI, drowsy, or distracted drivers; and at intersections with longer crossings had an increased probability of higher severity injury outcomes.

5.9 References

Brüde, U., & Larsson, J. (1993). "Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit?" *Accident Analysis & Prevention, 25*(5), 499-509. https://doi.org/10.1016/0001-4575(93)90001-D

Carlson, K., Murphy, B., Ermagun, A., Levinson, D., & Owen, A. (2018). "Safety in Numbers: Pedestrian and Bicyclist Activity and Safety in Minneapolis" (Report CTS 18-05). University of Minnesota. https://hdl.handle.net/11299/194707

Cottrill, C. D., & Thakuriah, P. V. (2010). "Evaluating pedestrian crashes in areas with high low income or minority populations." *Accident Analysis & Prevention, 42*(6), 1718-1728. https://doi.org/10.1016/j.aap.2010.04.012

Demetriades, D., Murray, J., Martin, M., Velmahos, G., Salim, A., Alo, K., & Rhee, P. (2004). "Pedestrians injured by automobiles: relationship of age to injury type and severity." *Journal of the American College of Surgeons, 199*(3), 382-387. https://doi.org/10.1016/j.jamcollsurg.2004.03.027

Dumbaugh, E., & Li, W. (2010). "Designing for the safety of pedestrians, cyclists, and motorists in urban environments." *Journal of the American Planning Association, 77*(1), 69-88. https://doi.org/10.1080/01944363.2011.536101

El-Basyouny, K., & Sayed, T. (2013). "Safety performance functions using traffic conflicts." *Safety Science, 51*(1), 160-164. https://doi.org/10.1016/j.ssci.2012.04.015

Elvik, R. (2009). "The non-linearity of risk and the promotion of environmentally sustainable transport." *Accident Analysis & Prevention, 41*(4), 849-855. https://doi.org/10.1016/j.aap.2009.04.009

Elvik, R. (2013). "Can a safety-in-numbers effect and a hazard-in-numbers effect co-exist in the same data?" *Accident Analysis & Prevention, 60*, 57-63. https://doi.org/10.1016/j.aap.2013.08.010

Elvik, R., & Goel, R. (2019). "Safety-in-numbers: An updated meta-analysis of estimates." *Accident Analysis & Prevention, 129*, 136-147. https://doi.org/10.1016/j.aap.2019.05.019

Elvik, R., Sørensen, M. W., & Nævestad, T. O. (2013). "Factors influencing safety in a sample of marked pedestrian crossings selected for safety inspections in the city of Oslo." *Accident Analysis & Prevention, 59*, 64-70. https://doi.org/10.1016/j.aap.2013.05.011

Geyer, J., Raford, N., Pham, T., & Ragland, D. R. (2006). "Safety in numbers: data from Oakland, California." *Transportation Research Record: Journal of the Transportation Research Board, 1982*(1), 150-154. https://doi.org/10.1177/0361198106198200119

Gladhill, K., & Monsere, C. M. (2012). "Exploring traffic safety and urban form in Portland, Oregon." *Transportation Research Record: Journal of the Transportation Research Board, 2318*(1), 63-74. https://doi.org/10.3141/2318-08

Graham, D. J., & Glaister, S. (2003). "Spatial variation in road pedestrian casualties: the role of urban scale, density and land-use mix." *Urban Studies, 40*(8), 1591-1607. https://doi.org/10.1080/0042098032000094441

Haleem, K., Alluri, P., & Gan, A. (2015). "Analyzing pedestrian crash injury severity at signalized and non-signalized locations." *Accident Analysis & Prevention, 81*, 14-23. https://doi.org/10.1016/j.aap.2015.04.025

Harkey, D. L., & Zegeer, C. V. (2004). *PEDSAFE: Pedestrian Safety Guide and Countermeasure Selection System* (FHWA-SA-04-003). Federal Highway Administration.

Harwood, D. W., Bauer, K. M., Richard, K. R., Gilmore, D. K., Graham, J. L., Potts, I. B., ... & Hauer, E. (2008). *Pedestrian Safety Prediction Methodology* (NCHRP Web-Only Document 129). Transportation Research Board. https://doi.org/10.17226/23083

Hosseinpour, M., Prasetijo, J., Yahaya, A. S., & Ghadiri, S. M. R. (2013). "A comparative study of count models: Application to pedestrian-vehicle crashes along Malaysia federal roads." *Traffic Injury Prevention, 14*(6), 630-638. https://doi.org/10.1080/15389588.2012.736649

Isgin, T., Bilgic, A., Forster, D. L., & Batte, M. T. (2008). "Using count data models to determine the factors affecting farmers' quantity decisions of precision farming technology adoption." *Computers and Electronics in Agriculture, 62*(2), 231-242. https://doi.org/10.1016/j.compag.2008.01.004

Jacobsen, P. L. (2015). "Safety in numbers: more walkers and bicyclists, safer walking and bicycling." *Injury Prevention, 21*(4), 271-275. https://doi.org/10.1136/ip.9.3.205rep

Lam, W. W., Yao, S., & Loo, B. P. (2014). "Pedestrian exposure measures: A time-space framework." *Travel Behaviour and Society, 1*(1), 22-30. https://doi.org/10.1016/j.tbs.2013.10.004

Lee, C., & Abdel-Aty, M. (2005). "Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida." *Accident Analysis & Prevention, 37*(4), 775-786. https://doi.org/10.1016/j.aap.2005.03.019

Lord, D., & Mannering, F. (2010). "The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives." *Transportation Research Part A: Policy and Practice, 44*(5), 291-305. https://doi.org/10.1016/j.tra.2010.02.001

Loukaitou-Sideris, A., Liggett, R., & Sung, H. G. (2007). "Death on the crosswalk: A study of pedestrian-automobile collisions in Los Angeles." *Journal of Planning Education and Research, 26*(3), 338-351. https://doi.org/10.1177/0739456X06297008

Lyon, C., & Persaud, B. (2002). "Pedestrian collision prediction models for urban intersections." *Transportation Research Record: Journal of the Transportation Research Board, 1818*(1), 102-107. https://doi.org/10.3141/1818-16

Miranda-Moreno, L. F., & Fu, L. (2006). "A comparative study of alternative model structures and criteria for ranking locations for safety improvements." *Networks and Spatial Economics, 6*(2), 97-110. https://doi.org/10.1007/s11067-006-7695-2

National Highway Traffic Safety Administration (NHTSA). (2020a). "2018 Fatal Motor Vehicle Crashes: Overview" (DOT HS 812 826). https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812826

NHTSA. (2020b). "Fatality Analysis Reporting System (FARS)" (accessed 28 July 2020). https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

Oh, C., Kang, Y. S., Kim, B., & Kim, W. (2005). "Analysis of pedestrian-vehicle crashes in Korea." Presented at the 84th Annual Meeting of the Transportation Research Board, Washington, DC.

Raford, N., & Ragland, D. (2006). "Pedestrian volume modeling for traffic safety and exposure analysis: The case of Boston, Massachusetts." Presented at the 85th Annual Meeting of the Transportation Research Board, Washington, DC. https://escholarship.org/uc/item/61n3s4zr

Roudsari, B., Kaufman, R., & Koepsell, T. (2006). "Turning at intersections and pedestrian injuries." *Traffic Injury Prevention, 7*(3), 283-289. https://doi.org/10.1080/15389580600660153

Sarkar, S., Tay, R., & Hunt, J. D. (2011). “Logistic regression model of risk of fatality in vehicle–pedestrian crashes on national highways in Bangladesh.” *Transportation Research Record: Journal of the Transportation Research Board, 2264*(1), 128-137. https://doi.org/10.3141/2264-15

Singleton, P. A., Runa, F., & Humagain, P. (2020). “Utilizing Archived Traffic Signal Performance Measures for Pedestrian Planning and Analysis” (Report UT-20.17*)* (Report UT-20.17). Utah Department of Transportation. https://rosap.ntl.bts.gov/view/dot/54924

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.” Purdue University. https://doi.org/10.4231/K4RN35SH

Stutts, J. C., Hunter, W. W., & Pein, W. E. (1996). “Pedestrian crash types: 1990s update.” *Transportation Research Record: Journal of the Transportation Research Board, 1538*(1), 68-74. https://doi.org/10.1177/0361198196153800109

Tarko, A., & Azam, M. S. (2011). “Pedestrian injury analysis with consideration of the selectivity bias in linked police-hospital data.” *Accident Analysis & Prevention, 43*(5), 1689-1695. https://doi.org/10.1016/j.aap.2011.03.027

Tulu, G. S., Washington, S., Haque, M. M., & King, M. J. (2015). “Investigation of pedestrian crashes on two-way two-lane rural roads in Ethiopia.” *Accident Analysis & Prevention, 78*, 118-126. https://doi.org/10.1016/j.aap.2015.02.011

Vuong, Q. H. (1989). “Likelihood ratio tests for model selection and non-nested hypotheses.” *Econometrica: Journal of the Econometric Society, 57*(2), 307-333. https://doi.org/10.2307/1912557

Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., & Bhatia, R. (2009). “An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning.” *Accident Analysis & Prevention, 41*(1), 137-145. https://doi.org/10.1016/j.aap.2008.10.001

Xu, P., Xie, S., Dong, N., Wong, S. C., & Huang, H. (2019). “Rethinking safety in numbers: are intersections with more crossing pedestrians really safer?” *Injury Prevention, 25*(1), 20-25. https://doi.org/10.1136/injuryprev-2017-042469

Yasmin, S., & Eluru, N. (2016). “Latent segmentation-based count models: analysis of bicycle safety in Montreal and Toronto.” *Accident Analysis & Prevention, 95*, 157-171. https://doi.org/10.1016/j.aap.2016.07.015

Zajac, S. S., & Ivan, J. N. (2003). “Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural Connecticut.” *Accident Analysis & Prevention, 35*(3), 369-379. https://doi.org/10.1016/S0001-4575(02)00013-1

Zegeer, C. V., Opiela, K. S., Cynecki, M. J., & Fegan, J. C. (1985). “Pedestrian Signalization Alternatives.” (FHWA-RD-83-102). Federal Highway Administration. https://rosap.ntl.bts.gov/view/dot/41946

Zegeer, C. V., Stewart, J. R., Huang, H. H., Lagerwey, P. A., Feaganes, J. R., & Campbell, B. J. (2005). “Safety Effects of Marked versus Unmarked Crosswalks at Uncontrolled Locations - Final Report and Recommended Guidelines” (FHWA-HRT-04-100). Federal Highway Administration. https://rosap.ntl.bts.gov/view/dot/40068

6. EXPLORING FACTORS AFFECTING PEDESTRIAN CRASH SEVERITY USING TABNET: A DEEP LEARNING APPROACH

This chapter is the accepted manuscript of an article presented at the 103rd Annual Meeting of the Transportation Research Board. It is reprinted here with permission from the authors. To cite, please use this reference:

- Rafe, A., & Singleton, P. A. (2024). "Exploring factors affecting pedestrian crash severity using TabNet: A deep learning approach." Presented at the 103rd Annual Meeting of the Transportation Research Board, Washington, DC. https://doi.org/10.48550/arXiv.2312.00066

6.1 Abstract

This study presents the first investigation of pedestrian crash severity using the TabNet model, a novel tabular deep learning method exceptionally suited for analyzing the tabular data inherent in transportation safety research. Through the application of TabNet to a comprehensive dataset from Utah covering the years 2010 to 2022, we uncover intricate factors contributing to pedestrian crash severity. The TabNet model, capitalizing on its compatibility with structured data, demonstrates remarkable predictive accuracy, eclipsing that of traditional models. It identifies critical variables—such as pedestrian age, involvement in left or right turns, lighting conditions, and alcohol consumption—which significantly influence crash outcomes. The utilization of SHapley Additive exPlanations (SHAP) enhances our ability to interpret the TabNet model's predictions, ensuring transparency and understandability in our deep learning approach. The insights derived from our analysis provide a valuable compass for transportation safety engineers and policymakers, enabling the identification of pivotal factors that affect pedestrian crash severity. Such knowledge is instrumental in formulating precise, data-driven interventions aimed at bolstering pedestrian safety across diverse urban and rural settings.

6.2 Introduction

Pedestrian safety remains a critical challenge in traffic systems worldwide, with pedestrians often bearing the highest risk of traffic crashes. In 2021 alone, the National Highway Traffic Safety Administration (NHTSA, 2023) reported 7,388 pedestrian fatalities in the United States, underscoring the need for improved safety measures. Various factors contribute to the severity of pedestrian crashes, with urban settings, intersections, and low-light conditions being predominant risk factors.

Data-driven analysis of crash reports is a key strategy for identifying factors that influence pedestrian crash severity. Recently, deep learning techniques have shown promise in this domain due to their ability to capture complex patterns from large volumes of data. This study harnesses the potential of TabNet, a state-of-the-art deep learning model designed for tabular data, which is prevalent in the field of transportation safety. TabNet's innovative architecture enables it to focus on the most relevant factors for crash severity prediction, thereby offering a powerful tool for traffic safety analysis.

Utilizing pedestrian crash data from Utah spanning 2010 to 2021, this study is the first to apply TabNet to pedestrian crash severity analysis. In conjunction with SHAP, we interpret the model's predictions, providing insights into the significance of various contributing factors. This novel approach not only enhances model interpretability but also aids in developing targeted strategies to improve pedestrian safety. The ensuing sections will detail the methodology, present the findings, and discuss the implications of employing TabNet within this vital area of public safety.

6.3 Literature Review

The severity of traffic incidents involving pedestrians is contingent upon a myriad of factors. A comprehensive review of relevant academic literature (Shrinivas et al., 2023) reveals several key variables. These include the demographic characteristics of the pedestrian, with a particular emphasis on age and gender; the speed and type of the implicated vehicle; the details of the accident location and the timing of the incident; the presence of intoxicating substances in the pedestrian or driver; and the use of safety equipment such as helmets or high-visibility clothing. These elements collectively contribute to the understanding and assessment of pedestrian-related traffic incidents.

Numerous prediction models have been employed to investigate the impact of various factors on pedestrian crash severity. These models encompass statistical techniques, such as negative binomial models (Rahman et al., 2022), logistic regression models (Nasri et al., 2022), ordered probit models (Fountas & Anastasopolous, 2018; Yang et al., 2019), and structural equation modeling (Kashani et al., 2021). Machine learning (ML) models, including random forest, AdaBoost (Al-Mistarehi et al., 2022), XGBoost (Goswamy et al., 2023), decision trees, k-nearest neighbor, and ensemble models (Yang et al., 2022), have also been utilized. Additionally, deep learning (DL) models, like deep neural networks (DNN) (Kang & Khattak, 2022), have been explored for pedestrian crash severity analysis. To better understand the application of these techniques, Table 6.1 presents an overview of the advantages and limitations of these methods used in pedestrian crash severity analysis.

While TabNet (Arik & Pfister, 2021) (a DL technique designed for tabular data analysis, capable of handling both numerical and categorical variables) has been used in crash severity analysis before, our study is novel in terms of applying TabNet specifically to pedestrian crash severity analysis. Prior work by Sattar et al. (2023) utilized TabNet for modeling injury severity in motor vehicle crashes using different ML approaches. However, their study did not focus on pedestrian-related crashes, and they did not propose the TabNet interpretation results. Therefore, our study also contributes by introducing the interpretation of TabNet results using SHAP, a framework previously employed for interpreting DNN models in crash injury severity analysis by Kang and Khattak (2022), and for XGBoost models in similar studies by Chang et al. (2022) and Li (2022). By incorporating SHAP, we aim to provide deeper insights into the factors influencing pedestrian crash severity predictions using the TabNet model.

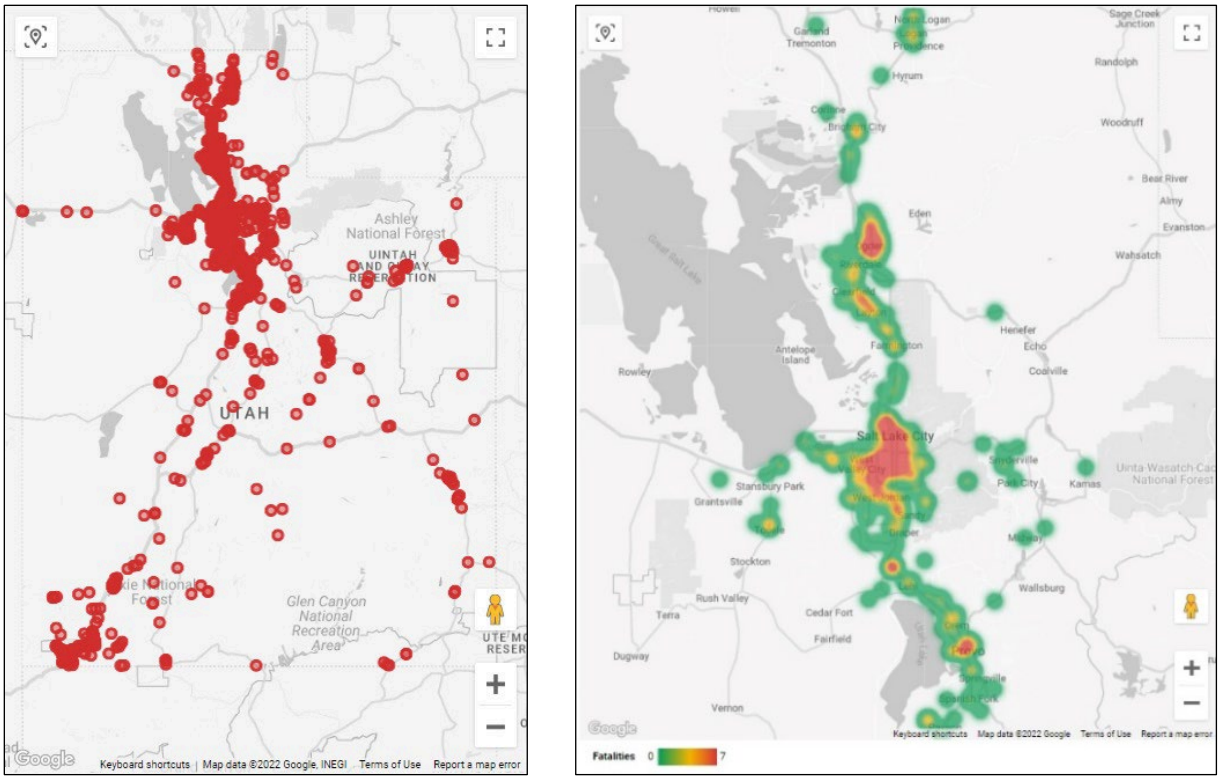
Table 6.1 Summary of benefits and limitations of various techniques for pedestrian crash severity analysis

Techniques	*Benefits*	*Limitations*
Statistical methods	- *Interpretability:* Statistical models, such as logistic regression and negative binomial models, offer greater interpretability and understanding compared with ML and DL models (Kashani et al., 2021; Infante et al., 2022). - *Simplicity:* Statistical models are generally simpler and require fewer computational resources than ML and DL models (Infante et al., 2022). - *Well-established techniques:* Statistical methods have a long history of use and research, making them reliable and well-established for analyzing crash severity (Kashani et al., 2021).	- *Linearity assumptions:* Some statistical models, like logistic regression, may assume a linear relationship between predictors and the outcome, which could be limited in capturing more complex real-world scenarios (Infante et al., 2022). - *Limited predictive power:* Statistical models might have lower predictive accuracy compared with ML and DL models, especially when handling intricate and nonlinear relationships between variables (Infante et al., 2022).
ML and DL methods	- *Higher predictive accuracy:* ML and DL methods can achieve superior predictive accuracy compared with statistical models, particularly when handling complex and nonlinear relationships between variables or when dealing with large and complex datasets (Kang & Khattak, 2022; Komol et al., 2021). - *Feature importance:* ML models can effectively identify significant features (explanatory variables) and their relationships with crash severity, providing valuable insights that might be more challenging to extract from statistical models (Komol et al., 2021).	- *Interpretability:* ML and DL models can be more challenging to interpret and comprehend than statistical models, which may hinder the ability to explain the relationships between variables and crash severity (Infante et al., 2022). - *Overfitting:* ML and DL models may be susceptible to overfitting, particularly when dealing with many features or a small dataset. This can lead to reduced generalizability and accuracy on unseen data (Kang & Khattak, 2022; Komol et al., 2021). - *Computational resources:* DL models typically demand more computational resources and longer training times in comparison to statistical and ML models (Kang & Khattak, 2022).

6.4 Data and Method

6.4.1 Data and Variables

In our research, we leveraged crash data (UDPS, 2023) to explore the determinants of pedestrian crash severities in Utah from 2010 to 2021. The severity of pedestrian crashes in our study was gauged using the KABCO scale. This scale classifies crashes into several categories: fatal, suspected serious injury, suspected minor injury, possible injury, and no injury or property-damage-only (PDO). To visually represent these data, Figure 6.1 showcases the spatial distribution of these crashes. Additionally, it includes a heatmap that accentuates the locations of fatal crashes within the dataset.



(a) Pedestrian crashes dispersion (b) Heatmap of pedestrian fatalities

Figure 6.1 The spatial configuration of pedestrian crashes

In this study, we examined 8,812 pedestrian crash incidents, analyzing the impact of 29 different variables, as detailed in Table 6.2. The breakdown of crash severities was as follows: fatal crashes comprised 5%, serious injuries 15%, minor injuries 44%, possible injuries 30%, and no injuries or property-damage-only (PDO) accounted for 6%. Notable insights from the data include a higher incidence of injury among male pedestrians and an increased rate of fatalities in the 30 to 59 age group. Factors like DUI (driving under the influence) and crashes involving older drivers contributed to 13% and 11% of pedestrian fatalities, respectively. A significant majority of these crashes occurred on arterial roads (52%) and predominantly in urban areas (97%). Intersections emerged as common sites for pedestrian crashes, accounting for 61% of the total, with nearly 3% of these being fatal. The study also found that left-turn and right-turn crashes occurred at similar rates. Regarding lighting conditions, 60% of crashes happened in daylight, while dark conditions without lighting were present in 37% of fatal crashes.

Table 6.2 Descriptive statistics of the variables

Characteristics	*Class*	*Total*	*Fatal*	*Serious injury*	*Minor injury*	*Possible injury*	*No injury / PDO*
Pedestrian crashes		8812 (0%)	476 (5%)	1363 (15%)	3856 (44%)	2624 (30%)	493 (6%)
Sex	Male	5282 (60%)	309 (65%)	849 (62%)	2261 (59%)	1514 (58%)	349 (71%)
	Female	3530 (40%)	167 (35%)	514 (38%)	1595 (41%)	1110 (42%)	144 (29%)
Age group	0 to 9	826 (9%)	33 (7%)	118 (9%)	391 (10%)	240 (9%)	44 (9%)
	10 to 29	4039 (46%)	119 (25%)	556 (41%)	1840 (48%)	1280 (49%)	244 (49%)
	30 to 59	3009 (34%)	202 (42%)	509 (37%)	1265 (33%)	866 (33%)	167 (34%)
	> 59	938 (11%)	122 (26%)	180 (13%)	360 (9%)	238 (9%)	38 (8%)
Aggressive driving	No	8703 (99%)	472 (99%)	1332 (98%)	3812 (99%)	2601 (99%)	486 (99%)
	Yes	109 (1%)	4 (1%)	31 (2%)	44 (1%)	23 (1%)	7 (1%)
Alcohol-drug test result	Both-Positive	11 (0%)	11 (2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Drug-Positive	34 (0%)	34 (7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Alcohol-Positive	15 (0%)	13 (3%)	2 (0%)	0 (0%)	0 (0%)	0 (0%)
	Negative	9 (0%)	9 (2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Not related	8743 (99%)	409 (86%)	1361 (100%)	3856 (100%)	2624 (100%)	493 (100%)
DUI	No	8586 (97%)	413 (87%)	1305 (96%)	3786 (98%)	2598 (99%)	484 (98%)
	Yes	226 (3%)	63 (13%)	58 (4%)	70 (2%)	26 (1%)	9 (2%)
Distracted driving	No	8105 (92%)	430 (90%)	1218 (89%)	3554 (92%)	2449 (93%)	454 (92%)
	Yes	707 (8%)	46 (10%)	145 (11%)	302 (8%)	175 (7%)	39 (8%)
Drowsy driving	No	8773 (100%)	463 (97%)	1354 (99%)	3846 (100%)	2620 (100%)	490 (99%)
	Yes	39 (0%)	13 (3%)	9 (1%)	10 (0%)	4 (0%)	3 (1%)
Older driver involved	No	7896 (90%)	423 (89%)	1223 (90%)	3446 (89%)	2366 (90%)	438 (89%)
	Yes	916 (10%)	53 (11%)	140 (10%)	410 (11%)	258 (10%)	55 (11%)
Teenage driver involved	No	7966 (90%)	432 (91%)	1205 (88%)	3496 (91%)	2390 (91%)	443 (90%)
	Yes	846 (10%)	44 (9%)	158 (12%)	360 (9%)	234 (9%)	50 (10%)
Holiday	No	7745 (88%)	397 (83%)	1185 (87%)	3402 (88%)	2324 (89%)	437 (89%)
	Yes	1067 (12%)	79 (17%)	178 (13%)	454 (12%)	300 (11%)	56 (11%)
Right-turn involved	No	7118 (81%)	464 (97%)	1254 (92%)	3113 (81%)	1912 (73%)	375 (76%)
	Yes	1694 (19%)	12 (3%)	109 (8%)	743 (19%)	712 (27%)	118 (24%)
Intersection involved	Yes	5361 (61%)	136 (29%)	718 (53%)	2471 (64%)	1766 (67%)	270 (55%)
	No	3451 (39%)	340 (71%)	645 (47%)	1385 (36%)	858 (33%)	223 (45%)
Left-turn involved	No	7079 (80%)	441 (93%)	1144 (84%)	3016 (78%)	2051 (78%)	427 (87%)
	Yes	1733 (20%)	35 (7%)	219 (16%)	840 (22%)	573 (22%)	66 (13%)
Overturn rollover	No	8785 (100%)	474 (100%)	1358 (100%)	3842 (100%)	2620 (100%)	491 (100%)
	Yes	27 (0%)	2 (0%)	5 (0%)	14 (0%)	4 (0%)	2 (0%)
Domestic animal involved	No	8793 (100%)	469 (99%)	1363 (100%)	3847 (100%)	2622 (100%)	492 (100%)
	Yes	19 (0%)	7 (1%)	0 (0%)	9 (0%)	2 (0%)	1 (0%)
Commercial vehicle involved	No	8559 (97%)	440 (92%)	1302 (96%)	3772 (98%)	2581 (98%)	464 (94%)
	Yes	253 (3%)	36 (8%)	61 (4%)	84 (2%)	43 (2%)	29 (6%)

Heavy truck involved	No	8534 (97%)	440 (92%)	1297 (95%)	3760 (98%)	2577 (98%)	460 (93%)
	Yes	278 (3%)	36 (8%)	66 (5%)	96 (2%)	47 (2%)	33 (7%)
Transit vehicle involved	No	8732 (99%)	470 (99%)	1349 (99%)	3823 (99%)	2606 (99%)	484 (98%)
	Yes	80 (1%)	6 (1%)	14 (1%)	33 (1%)	18 (1%)	9 (2%)
Work zone involved	No	8425 (96%)	447 (94%)	1298 (95%)	3705 (96%)	2503 (95%)	472 (96%)
	Yes	387 (4%)	29 (6%)	65 (5%)	151 (4%)	121 (5%)	21 (4%)
Wrong way driving	No	8784 (100%)	473 (99%)	1359 (100%)	3841 (100%)	2620 (100%)	491 (100%)
	Yes	28 (0%)	3 (1%)	4 (0%)	15 (0%)	4 (0%)	2 (0%)
Road type	Urban	8548 (97%)	419 (88%)	1296 (95%)	3784 (98%)	2583 (98%)	466 (95%)
	Rural	264 (3%)	57 (12%)	67 (5%)	72 (2%)	41 (2%)	27 (5%)
Functional class	Local	2651 (30%)	71 (15%)	352 (26%)	1256 (33%)	847 (32%)	125 (25%)
	Collector	1578 (18%)	71 (15%)	232 (17%)	703 (18%)	487 (19%)	85 (17%)
	Arterial	4583 (52%)	334 (70%)	779 (57%)	1897 (49%)	1290 (49%)	283 (57%)
Roadway surface is dry	Yes	7607 (86%)	409 (86%)	1181 (87%)	3312 (86%)	2273 (87%)	432 (88%)
	No	1205 (14%)	67 (14%)	182 (13%)	544 (14%)	351 (13%)	61 (12%)
Lighting condition	Dark-Not lighted	1167 (13%)	176 (37%)	285 (21%)	401 (10%)	255 (10%)	50 (10%)
	Dark-Lighted	1912 (22%)	138 (29%)	332 (24%)	824 (21%)	542 (21%)	76 (15%)
	Daylight	5292 (60%)	141 (30%)	678 (50%)	2409 (62%)	1725 (66%)	339 (69%)
	Dusk	244 (3%)	10 (2%)	40 (3%)	115 (3%)	59 (2%)	20 (4%)
	Dawn	197 (2%)	11 (2%)	28 (2%)	107 (3%)	43 (2%)	8 (2%)
Weather condition	Clear	6758 (77%)	355 (75%)	1068 (78%)	2925 (76%)	2030 (77%)	380 (77%)
	Cloudy	1214 (14%)	69 (14%)	176 (13%)	543 (14%)	353 (13%)	73 (15%)
	Rain	509 (6%)	31 (7%)	80 (6%)	220 (6%)	153 (6%)	25 (5%)
	Fog, Smog	25 (0%)	3 (1%)	4 (0%)	9 (0%)	8 (0%)	1 (0%)
	Snowing	213 (2%)	10 (2%)	26 (2%)	112 (3%)	53 (2%)	12 (2%)
	Others	93 (1%)	8 (2%)	9 (1%)	47 (1%)	27 (1%)	2 (0%)
Vertical alignment	Level	6891 (78%)	360 (76%)	1108 (81%)	3005 (78%)	2042 (78%)	376 (76%)
	Uphill	61 (1%)	3 (1%)	7 (1%)	33 (1%)	14 (1%)	4 (1%)
	Downhill	50 (1%)	2 (0%)	12 (1%)	28 (1%)	8 (0%)	0 (0%)
	Others	1810 (21%)	111 (23%)	236 (17%)	790 (20%)	560 (21%)	113 (23%)

In the development of our TabNet models, we adhered to the categorization outlined in Table 6.2. To ensure a uniform encoding of the dataset, we assigned numerical values to categorical data. For instance, we designated “Yes” as 1 and “No” as 0; “Male” received a value of 1, while “Female” was assigned 0; similarly, “Rural” was encoded as 1 and “Urban” as 0. Other categories were numerically encoded following their sequential arrangement in Table 6.2, starting from 1 and increasing. Furthermore, we treated age as a continuous variable, rather than categorizing it into different age groups.

6.4.2 Method

In this study, we utilized the TabNet methodology to delve into the effects of various explanatory variables on pedestrian crash injury outcomes. TabNet, a model tailored for tabular data, is celebrated for its robust performance and interpretability, initially developed by the team at Google Cloud AI (Arik & Pfister, 2021). It ingeniously merges the capabilities of deep learning models with feature selection techniques, adept at processing both numerical and categorical data. TabNet’s core functionality lies in its use of sequential attention. This feature allows the model to selectively focus on different explanatory variables (EVs) at each decision-making step, thereby enhancing its interpretability. Figure 6.2 in our study depicts the specific structure of the TabNet model as applied here, highlighting its architecture over two steps. Within this framework, the EV transformer plays a crucial role in refining input data, which helps in better understanding the interplay between EVs and crash severity levels. Concurrently, the attentive transformer assesses the significance of each EV during each decision step. It creates a mask to emphasize the most influential predictors, enabling the model to dynamically concentrate on pertinent factors such as weather conditions and alcohol involvement, among other EVs. This approach not only bolsters the model’s focus but also significantly augments the accuracy of its predictions.

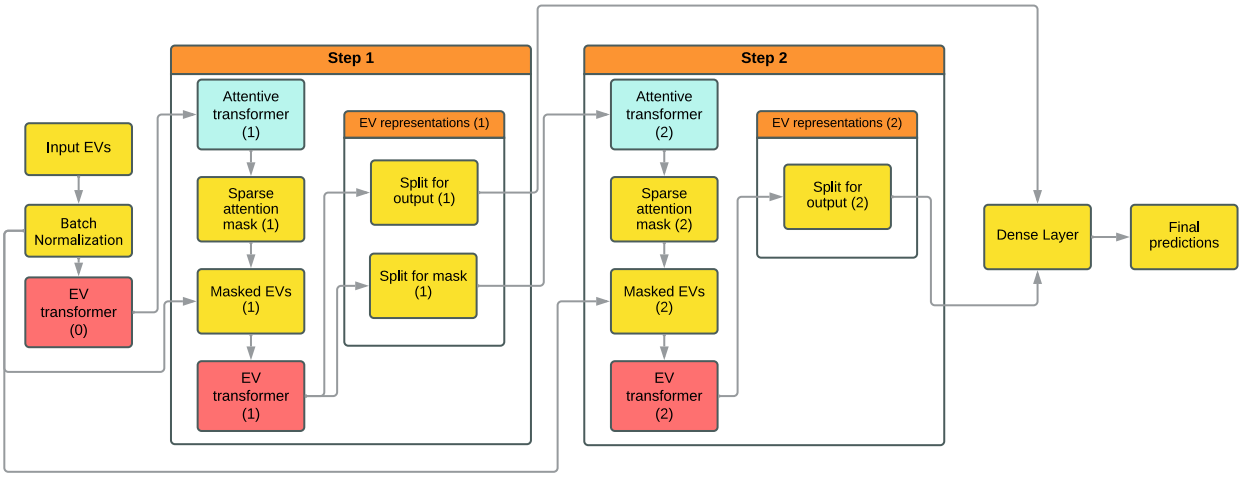


Figure 6.2 The structure of the TabNet model for predicting crash severity levels using various EVs

When applying the TabNet model to predict pedestrian crash severity, we implemented several steps to optimize its performance and accuracy. To counter the class imbalance in our dataset, we used the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002), enhancing the model’s proficiency in predicting less-represented classes. The model’s hyperparameters were fine-tuned with the help of Optuna (Akiba et al., 2019), a framework specialized in hyperparameter optimization, to achieve the best possible configuration tailored to our specific dataset. Additionally, to prevent overfitting and improve the model’s ability to generalize, we conducted multiple training iterations on varied subsets of data through bootstrapping. We evaluated the model’s effectiveness using a range of metrics, including accuracy, precision, recall, and the F1-score, to ensure a comprehensive assessment of its performance. The calculations for these metrics are represented by Eq. 1 to 4:

$$accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (1)$$

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F1\ score = \frac{2 \times (precision \times recall)}{precision + recall} \quad (4)$$

For interpreting the results of our TabNet model, we employed SHAP (Lundberg et al., 2018; Lundberg & Lee, 2017). SHAP assigns an importance value to each EV for a given prediction, making the model's output more understandable in terms of the input EVs. Drawing from cooperative game theory, SHAP values distribute the prediction output (crash severity) among the EVs based on their contribution. If we denote f(x) as the prediction for a specific instance x and $E[f(X)]$ as the expected prediction for the model, which is calculated as the average prediction over the training dataset, the additive EV attribution can be calculated as follows:

$$\text{f(x)} - E[f(X)] = \sum_{i=1}^{N} \varphi_i \quad (5)$$

Furthermore, the importance value assigned to each EV or the Shapley value for the i-th EV φ_i can be calculated as follows:

$$\varphi_i = \sum_{S \subseteq N\{i\}} \left[\frac{|S|!(|N|-|S|-1)!}{|N|!} \right] (f_i(S \cup \{i\}) - f_i(s)) \quad (6)$$

where N is the set of all EVs, S is a subset of N that includes the i-th EV, |S| is the size of S, and f_i is a version of where only the EVs in S and i (if it is included) are used. For this analysis, the SHAP python package (SHAP, 2023) was utilized to determine the importance of EVs in the TabNet model.

6.5 Model Results

For evaluating the various models, we partitioned our data, dedicating 80% for training purposes and reserving the final 20% for testing and evaluation. With the TabNet model, we employed the SHAP method to discern the significance of each EV across different crash severity classes. This approach and its findings are illustrated in Figure 6.3, providing a clear visual representation of how each EV influences the model's predictions for each severity level.

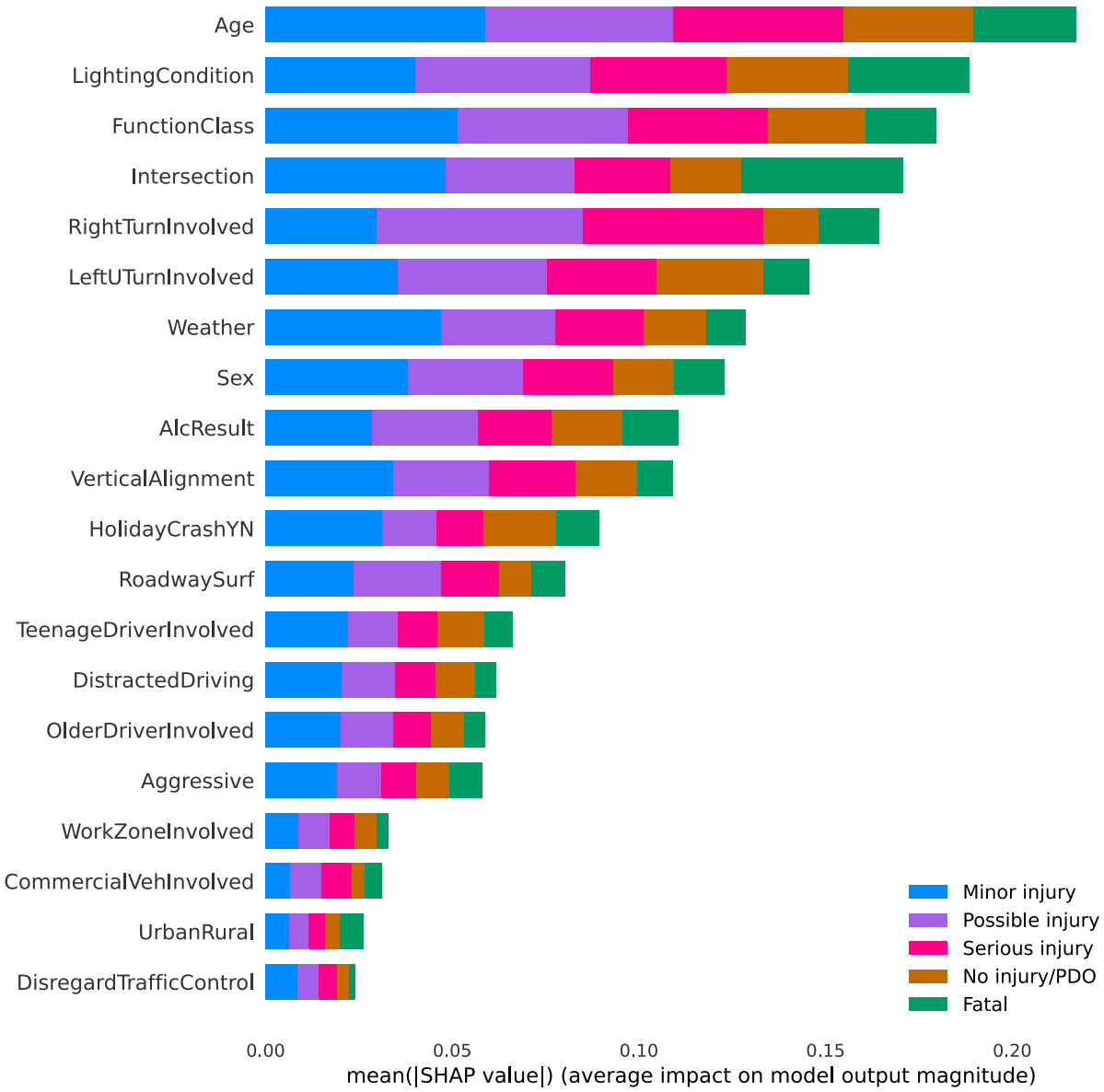


Figure 6.3 The importance of each EV for each crash severity class in TabNet model

To enhance the precision of the TabNet model, we meticulously adjusted its hyperparameters through a combination of GridSearchCV (Pedregosa et al., 2012) and the Optuna optimization technique. The specific values and methods utilized for this fine-tuning process are comprehensively listed in Table 6.3.

Table 6.3 Optimum hyperparameters of the TabNet models in this study

Model	*Hyperparameter*	*Value/Method*
TabNet	- Dimension of the prediction layer	53
	- Dimension of the attention layer	58
	- Number of decision steps	1
	- Sparsity regularization	0.023989318
	- Entmax* temperature** (gamma)	1.952667709
	- Number of independent GLU*** layers	8
	- Number of shared GLU layers across decision steps	6
	- Momentum in batch normalization	0.3
	- Gradient clipping for each parameter	2
	- Optimizer function	Adam
	- Learning rate (lr)	0.007566832
	- Mask type	Entmax

* It is a combination of "Maximum" and "Entropy," which signifies the objective of maximizing entropy while adhering to specific constraints. ** It is a hyperparameter that controls the sharpness of the probability distribution. *** Gated linear unit

For assessing the TabNet model's efficacy, we employed a suite of evaluation metrics, including precision, recall, F-1 score, and overall accuracy. The outcomes derived from these metrics, offering insights into the model's performance, are detailed in Table 6.4.

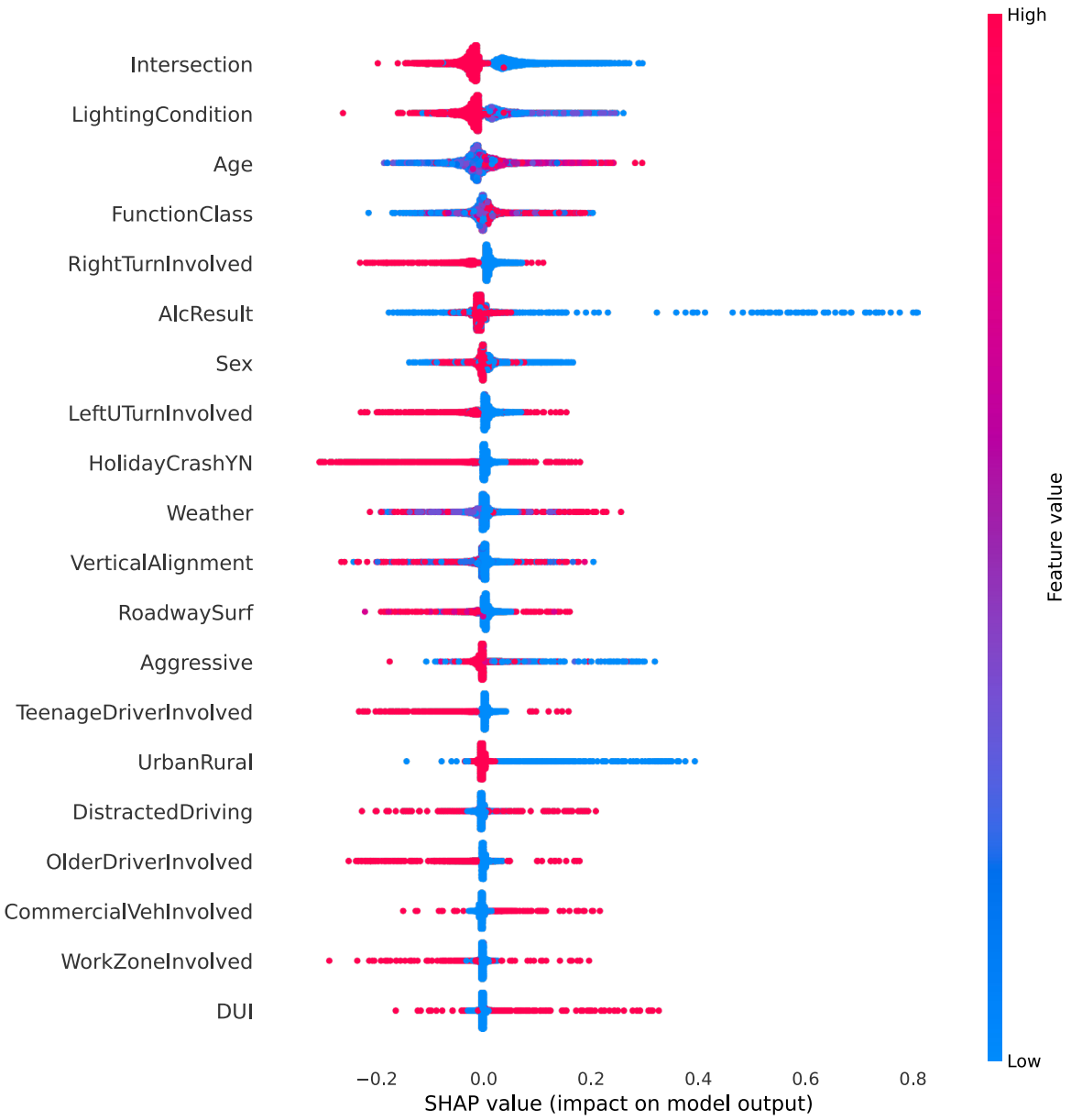
Table 6.4 Performance evaluation metrics

Crash severity class	*Evaluation metrics*		
	Precision	*Recall*	*F1 score*
Fatal	0.910	0.950	0.930
Serious injury	0.860	0.860	0.860
Minor injury	0.927	0.980	0.950
Possible injury	0.960	0.970	0.959
No injury/PDO	0.948	0.916	0.927
Accuracy		0.959	

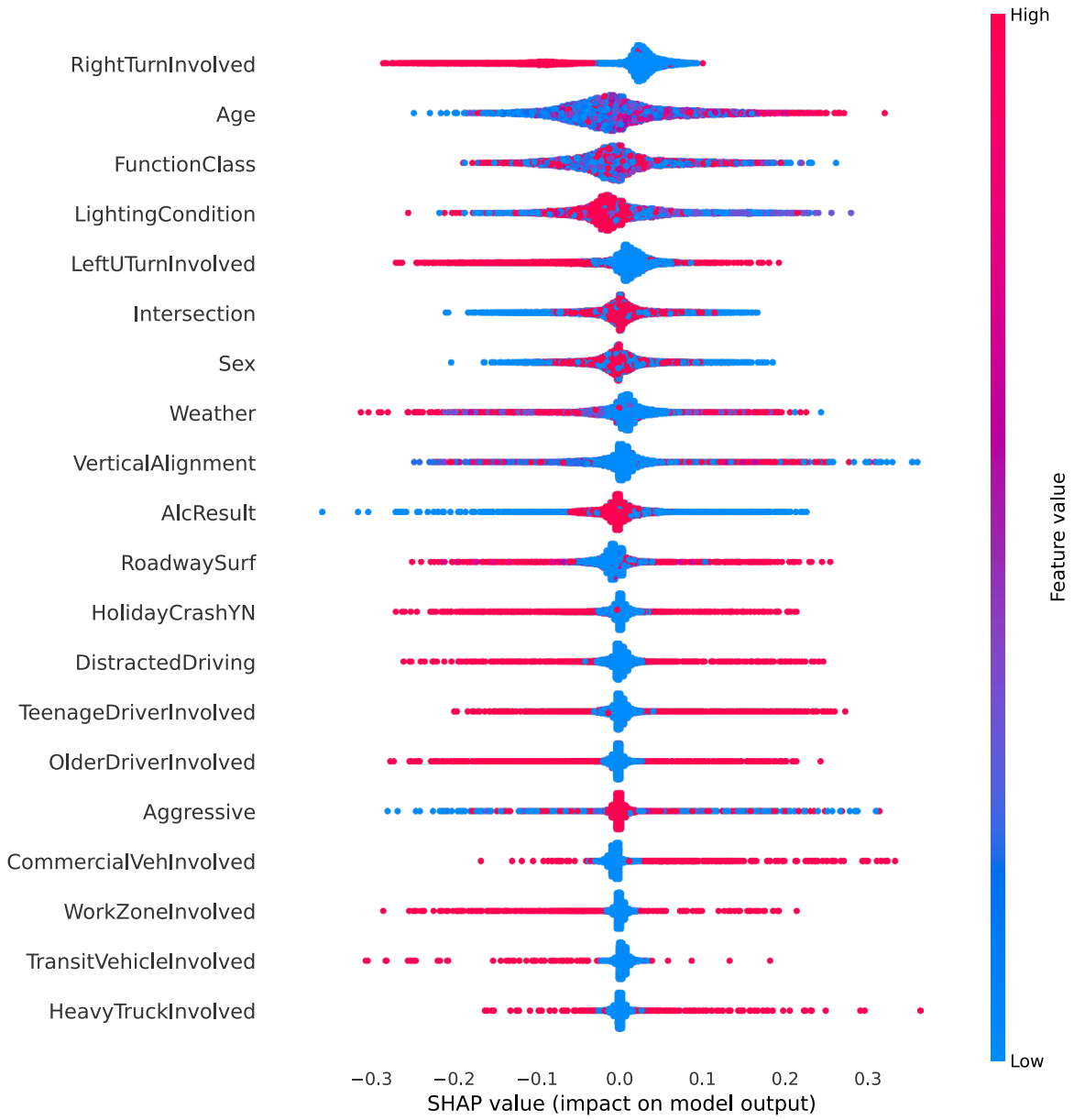
6.6 Model Interpretation and Discussion

When we examined the performance metrics, as detailed in Table 6.4, the TabNet model distinguished itself with its precise predictions of pedestrian crash severity. It demonstrated particular strength in predicting minor and possible injury outcomes, as evidenced by its F1-score in these categories. To maintain the integrity of the TabNet model and to address the risk of overfitting due to its notable accuracy, we employed a range of methods. These included cross-validation, regularization parameters, an early stopping mechanism, the use of SMOTE, and training with diverse bootstrap samples. These strategies collectively improved both the performance and dependability of the model in our analysis.

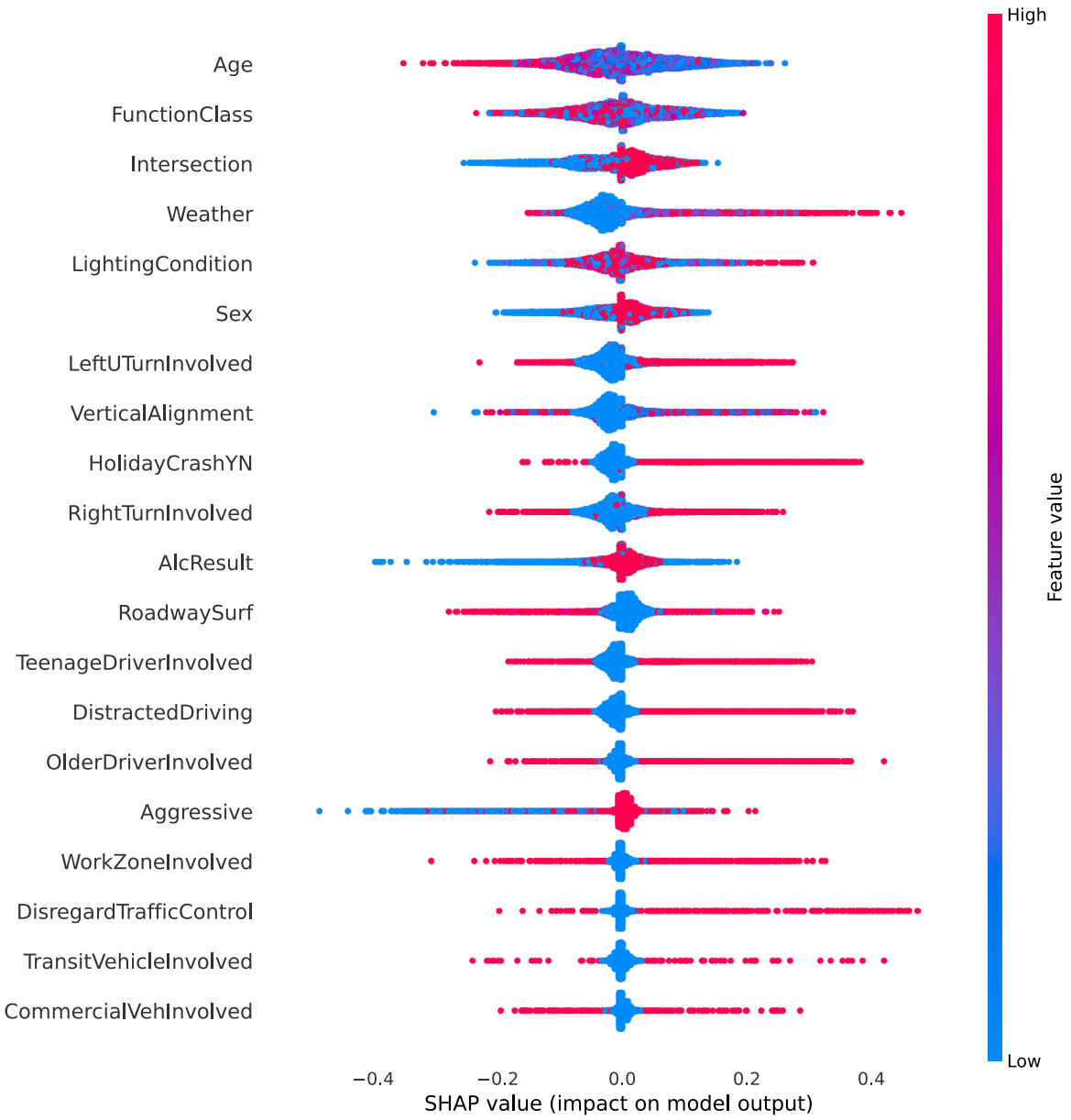
The TabNet model's findings highlight pedestrian age, lighting conditions, and road functional class as key EVs in predicting crash severity. Figure 6.3 elucidates these influential EVs. To further understand the model, Figure 6.4 offers a SHAP summary plot that correlates EV features with crash severity classes. In this plot, each row represents an EV, with the color of the dots indicating the EV's value (red for high, blue for low), and their horizontal position indicating how the EV influences the probability of a higher severity outcome. The clustering of dots indicates a strong correlation between the feature and the prediction, with the spread showing the EV's impact and dot dispersion highlighting variation due to interactions with other EVs.


Intersection
LightingCondition
Age
FunctionClass
RightTurnInvolved
AlcResult
Sex
LeftUTurnInvolved
HolidayCrashYN
Weather
VerticalAlignment
RoadwaySurf
Aggressive
TeenageDriverInvolved
UrbanRural
DistractedDriving
OlderDriverInvolved
CommercialVehInvolved
WorkZoneInvolved
DUI
−0.2
0.0
0.2
0.4
0.6
0.8
SHAP value (impact on model output)
High
Feature value
Low

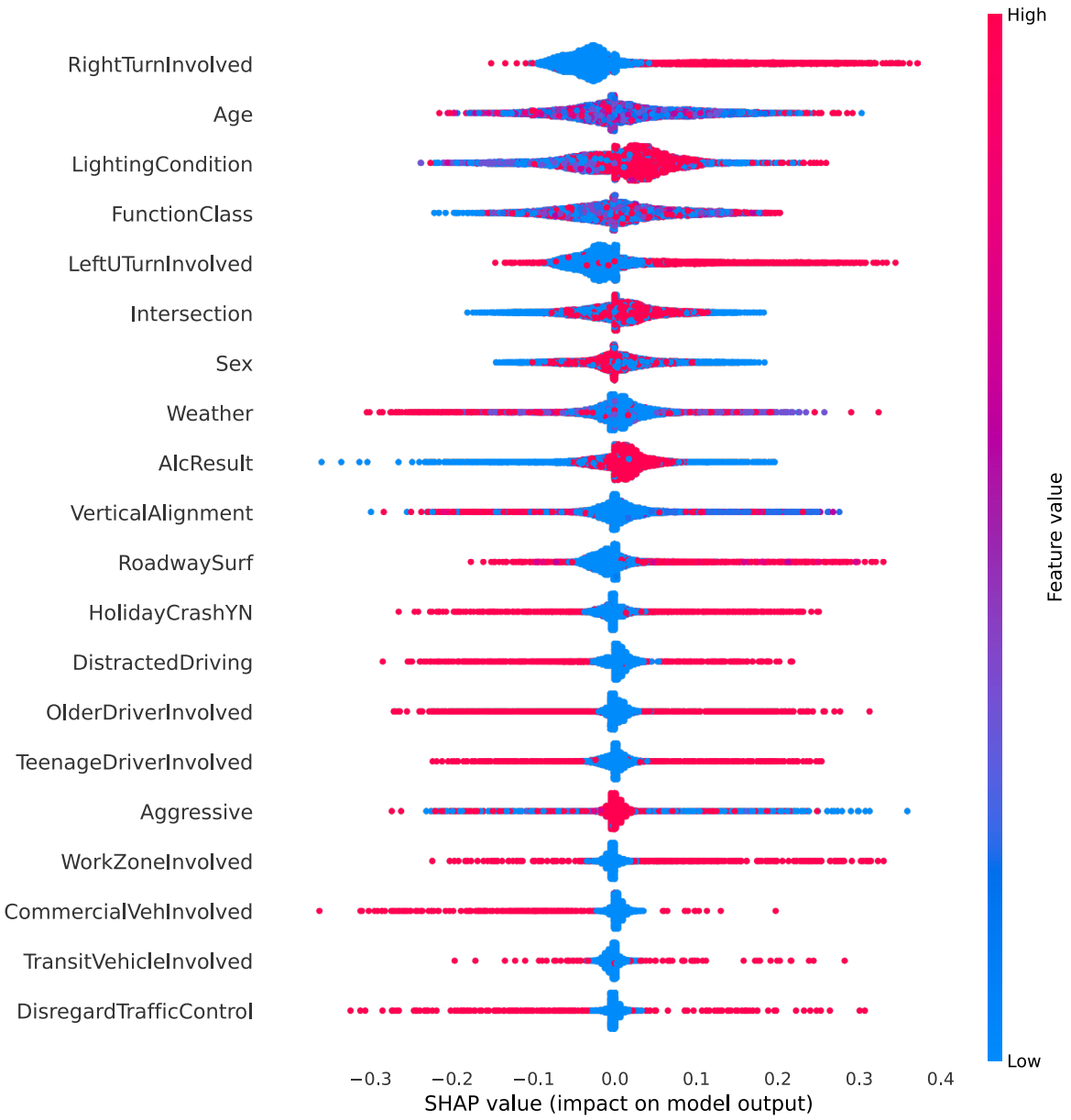
(a) Fatal



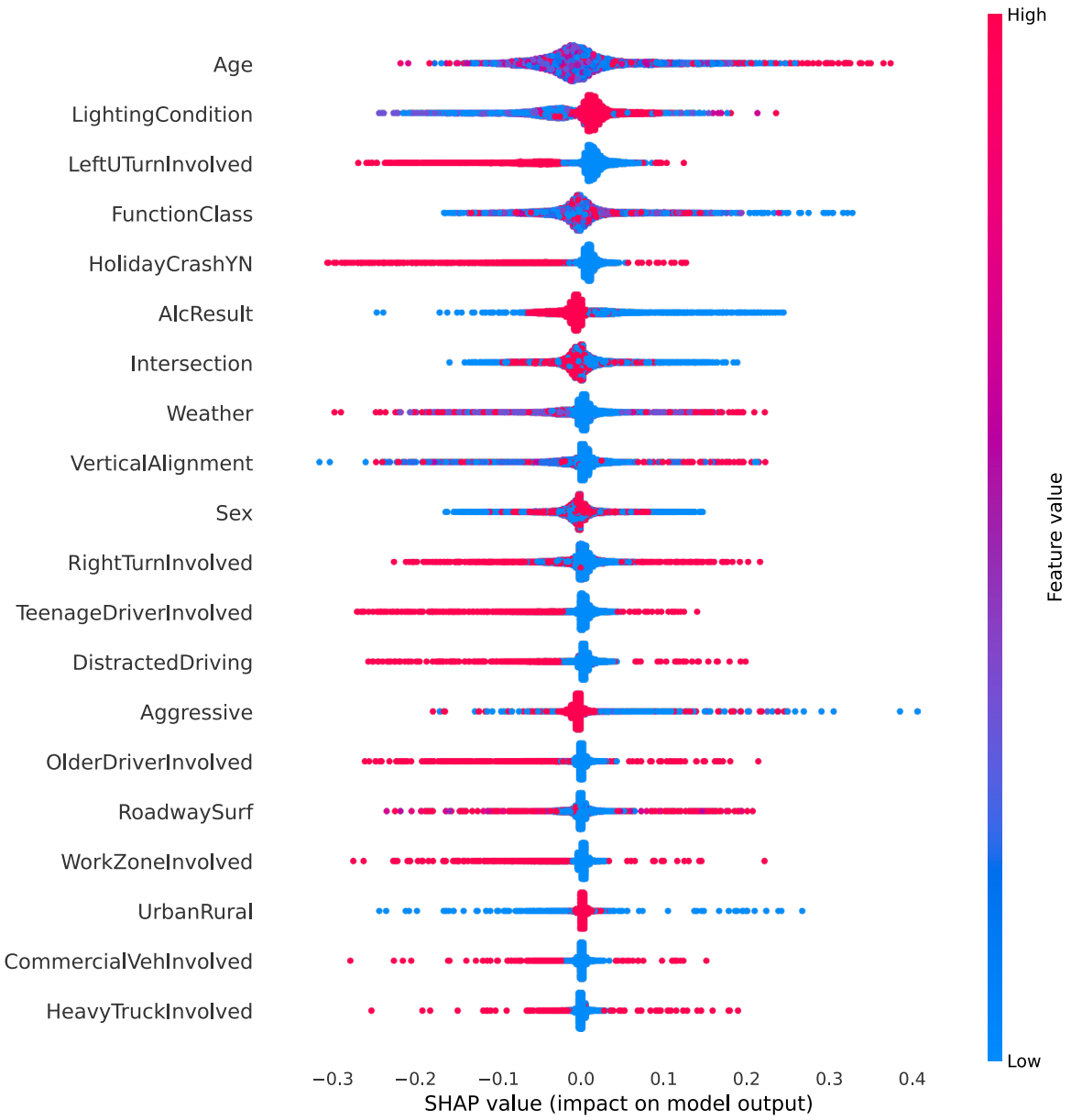
(b) Serious injury



(c) Minor injury



(d) Possible injury



(e) No inury/PDO

Figure 6.4 The SHAP summary plot for each crash severity class in TabNet model

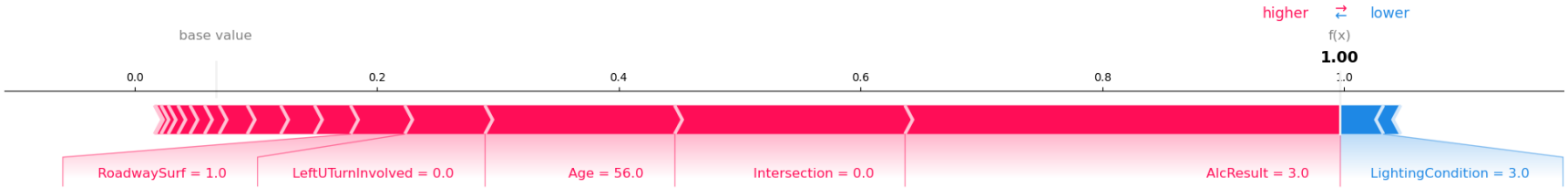
From the dot plot, certain features stand out for increasing the likelihood of fatal outcomes. Figure 6.4(a) shows that alcohol or drug consumption by pedestrians, and crashes in urban areas, are linked with higher chances of fatal severity. In serious injury cases [Figure 6.4(b)], the involvement of commercial vehicles and heavy trucks is a critical factor, though other EVs show variable impacts, suggesting complex interplays within the model. In minor injury cases, factors like disregard for traffic control, distracted driving, and holiday-period crashes are predictors, while right and left turns, work zones, and dry road conditions are more associated with possible injuries. Contrarily, adverse weather and commercial vehicle involvement reduce the odds of possible injuries. For the no injury/PDO category, holidays, left turns, and the involvement of older or teenage drivers are inversely related to severity. These insights, provided by

the SHAP analysis in Figure 6.4, highlight the nuanced interplay of various factors in pedestrian crash severity outcomes, as captured by the TabNet model.

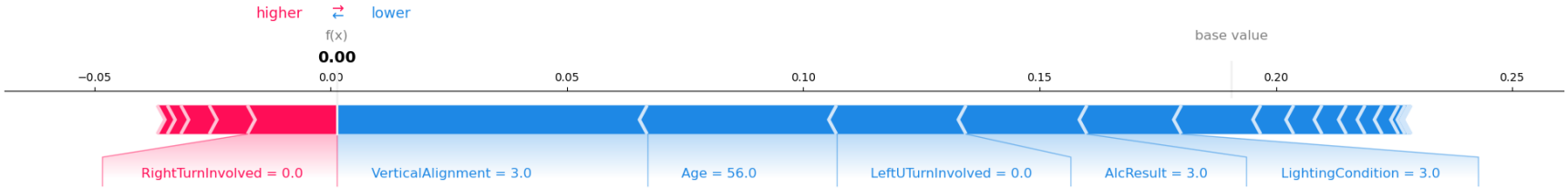
To navigate the complexity of the dot summary plot, especially for intricate categories like serious injury, and to delve deeper into how each EV contributes to the final prediction, we employed SHAP force plots, exemplified in Figure 6.5 using observation #631 from our dataset. These force plots visually depict the influence of each EV on the model's prediction, starting from the base value (the average prediction) and culminating in the specific outcome for an observation. Here, the impact of each EV is shown as a horizontal force, indicating its effect in either increasing or decreasing the prediction likelihood.

In Figure 6.5(a), focusing on observation #631, the TabNet model shows a tendency to classify this case as fatal (f(x)=1). Factors like the alcohol result, presence at an intersection, age, involvement in a left turn, and roadway surface type all point toward a fatal outcome. Conversely, the lighting condition applies a minor negative impact, but it is insufficient to outweigh the substantial positive influences from the other variables.

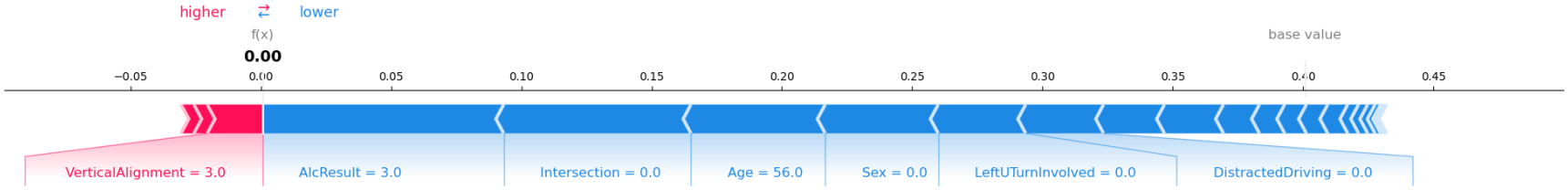
Additionally, in Figure 6.5(b), the model predicts a non-serious injury outcome (f(x)=0). The base value, ranging between 0.15 and 0.20, acts as the starting point in the absence of specific information about this observation. A significant blue arrow indicates that the vertical alignment variable heavily influences the prediction toward f(x) = 0.00. Factors like age, left-turn involvement, alcohol result, and lighting condition also contribute negatively, albeit to a lesser extent. In contrast, the right-turn involvement (indicated by a pink arrow) partly mitigates but does not fully offset the negative influences. This pattern is representative across other categories as well. From this analysis, it is clear that the model accurately classified observation #631 as fatal. Among the influencing features, the most impactful was the alcohol result, underscoring its significance in determining crash severity in this instance.



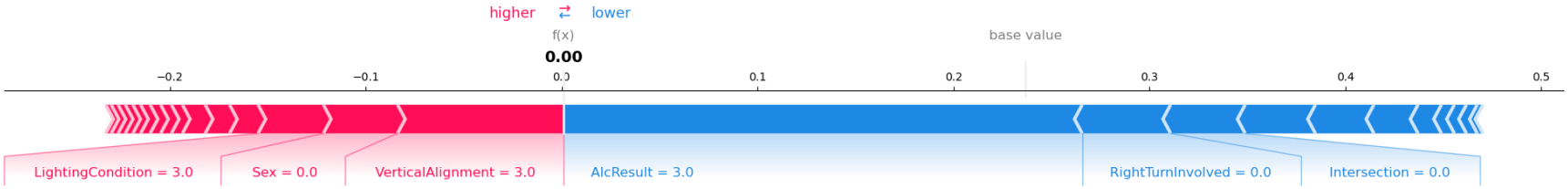
(a) Fatal



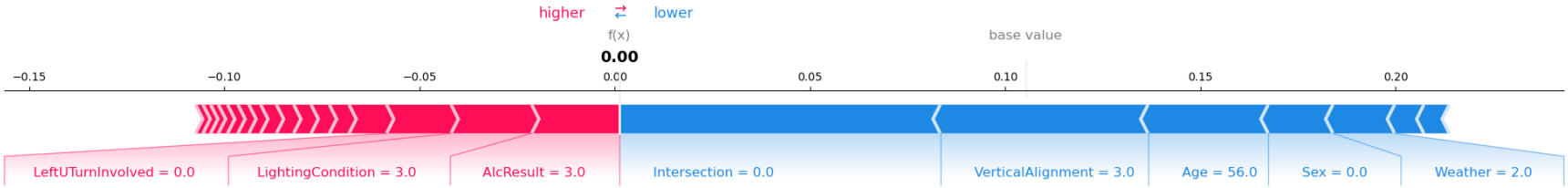
(b) Serious injury



(c) Minor injury



(d) Possible Injury



(e) No injury/PDO

Figure 6.5 The SHAP values, explaining the contribution of EVs to the raw TabNet model output for a specific observation

6.7 Conclusion

In the realm of transportation safety, understanding pedestrian crash severity is crucial, particularly due to the inherent vulnerability of pedestrians. This study focused on employing TabNet, an advanced deep learning (DL) method designed for tabular data. Our use of SHAP techniques for interpretation further enhanced our understanding of TabNet's application. The findings indicated that TabNet was exceptionally effective in analyzing pedestrian crash data from Utah. However, employing TabNet did pose challenges, particularly in hyperparameter tuning and model interpretation. For instance, tuning hyperparameters for TabNet required a considerable amount of time—20 hours and 16 minutes—on a general computer setup (Core i7- 9th generation with 32 GB RAM). Moreover, interpreting the TabNet results using SHAP was a time-intensive process, taking approximately 68 hours and 31 minutes. This highlights a crucial trade-off: the choice between achieving high accuracy with DL and ML models, which necessitates more time, versus opting for faster but potentially less accurate results from statistical methods.

In summary, our study provides valuable insights for transportation engineers in choosing appropriate methods for analyzing pedestrian crash severity. The methodologies and approaches we employed, especially focusing on TabNet, offer a framework that can be adapted for broader crash variable investigations in the field of transportation safety.

6.8 References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). "Optuna: A next-generation hyperparameter optimization framework." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623-2631. https://doi.org/10.1145/3292500.3330701

Al-Mistarehi, B. W., Alomari, A. H., Imam, R., & Mashaqba, M. (2022). "Using machine learning models to forecast severity level of traffic crashes by R Studio and ArcGIS." *Frontiers in Built Environment, 8*, 860805. https://doi.org/10.3389/fbuil.2022.860805

Arik, S. Ö., & Pfister, T. (2021). "TabNet: Attentive interpretable tabular learning." Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6679-6687. https://doi.org/10.1609/aaai.v35i8.16826

Chang, I., Park, H., Hong, E., Lee, J., & Kwon, N. (2022). "Predicting effects of built environment on fatal pedestrian accidents at location-specific level: Application of XGBoost and SHAP." *Accident Analysis & Prevention, 166*, 106545. https://doi.org/10.1016/J.AAP.2021.106545

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: "Synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953

Fountas, G., & Anastasopoulos, P. C. (2018). "Analysis of accident injury-severity outcomes: The zero-inflated hierarchical ordered probit model with correlated disturbances." *Analytic Methods in Accident Research, 20*, 30-45. https://doi.org/10.1016/J.AMAR.2018.09.002

Goswamy, A., Abdel-Aty, M., & Islam, Z. (2023). "Factors affecting injury severity at pedestrian crossing locations with Rectangular RAPID Flashing Beacons (RRFB) using XGBoost and random parameters discrete outcome models." *Accident Analysis & Prevention, 181*, 106937. https://doi.org/10.1016/j.aap.2022.106937

Infante, P., Jacinto, G., Afonso, A., Rego, L., Nogueira, V., Quaresma, P., ... & Manuel, P. R. (2022). "Comparison of statistical and machine-learning models on road traffic accident severity classification." *Computers, 11*(5), 80. https://doi.org/10.3390/computers11050080

Kang, Y., & Khattak, A. J. (2022). "Deep Learning Model for Crash Injury Severity Analysis Using Shapley Additive Explanation Values." *Transportation Research Record: Journal of the Transportation Research Board, 2676*(12), 242-254. https://doi.org/10.1177/03611981221095087

Kashani, A. T., Jafari, M., & Bondarabadi, M. A. (2021). "A new approach in analyzing the accident severity of pedestrian crashes using structural equation modeling." *Journal of Injury and Violence Research, 13*(1), 23-30. https://doi.org/10.5249%2Fjivr.v13i1.1545

Komol, M. M. R., Hasan, M. M., Elhenawy, M., Yasmin, S., Masoud, M., & Rakotonirainy, A. (2021). "Crash severity analysis of vulnerable road users using machine learning." *PLoS One, 16*(8), e0255828. https://doi.org/10.1371/journal.pone.0255828

Li, Z. (2022). "Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost." *Computers, Environment and Urban Systems, 96*, 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). "Consistent Individualized Feature Attribution for Tree Ensembles." https://doi.org/10.48550/arXiv.1802.03888

Lundberg, S. M., & Lee, S. I. (2017). "A Unified Approach to Interpreting Model Predictions." https://doi.org/10.48550/arXiv.1705.07874

Nasri, M., Aghabayk, K., Esmaili, A., & Shiwakoti, N. (2022). "Using ordered and unordered logistic regressions to investigate risk factors associated with pedestrian crash injury severity in Victoria, Australia." *Journal of Safety Research, 81*, 78-90. https://doi.org/10.1016/J.JSR.2022.01.008

National Highway Traffic Safety Administration (NHTSA). (2023). "Pedestrian Safety: Prevent Pedestrian Crashes" (accessed 22 July 2023). https://www.nhtsa.gov/road-safety/pedestrian-safety

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). "Scikit-learn: Machine learning in Python." *The Journal of Machine Learning Research, 12*, 2825-2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Rahman, M., Kockelman, K. M., & Perrine, K. A. (2022). "Investigating risk factors associated with pedestrian crash occurrence and injury severity in Texas." *Traffic Injury Prevention, 23*(5), 283-289. https://doi.org/10.1080/15389588.2022.2059474

Sattar, K., Chikh Oughali, F., Assi, K., Ratrout, N., Jamal, A., & Masiur Rahman, S. (2023). "Transparent deep machine learning framework for predicting traffic crash severity." *Neural Computing and Applications, 35*(2), 1535-1547. https://doi.org/10.1007/s00521-022-07769-2

shap. (2023). "SHAP: A game theoretic approach to explain the output of any machine learning model" (accessed 19 July 2023). https://github.com/shap/shap

Shrinivas, V., Bastien, C., Davies, H., Daneshkhah, A., & Hardwicke, J. (2023). "Parameters influencing pedestrian injury and severity–a systematic review and meta-analysis." *Transportation Engineering, 11*, 100158. https://doi.org/10.1016/j.treng.2022.100158

Utah Department of Public Safety (UDPS). (2023). "Utah Crash Summary" (accessed 18 July 2023). https://udps.numetric.net/utah-crash-summary#/

Yang, L., Aghaabbasi, M., Ali, M., Jan, A., Bouallegue, B., Javed, M. F., & Salem, N. M. (2022). "Comparative analysis of the optimized KNN, SVM, and ensemble DT models using Bayesian optimization for predicting pedestrian fatalities: An advance towards realizing the sustainable safety of pedestrians." *Sustainability, 14*(17), 10467. https://doi.org/10.3390/SU141710467

Yang, Z., Chen, F., Ma, X., & Dong, B. (2019, July). "Injury severity of pedestrians at mid-blocks: A random parameter ordered probit approach." 2019 5th International Conference on Transportation Information and Safety *(ICTIS)*, 735-740. https://doi.org/10.1109/ICTIS.2019.8883531