

Estimating annual average daily pedestrian traffic volume at intersections

by

Lucas Tito Pereira Sobreira

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Civil Engineering

Waterloo, Ontario, Canada, 2025

© Lucas Tito Pereira Sobreira 2025

Examining Committee Membership

The following members served on the Examining Committee for this Thesis. The decision of the Examining Committee is by majority vote.

- External Examiner: **Dr. Patrick Singleton**
Associate Professor, Dept. of Civil and Environmental
Engineering, Utah State University
- Supervisor: **Dr. Bruce Hellinga**
Professor, Dept. of Civil and Environmental
Engineering, University of Waterloo
- Internal Member: **Dr. Liping Fu**
Professor, Dept. of Civil and Environmental
Engineering, University of Waterloo
- Internal-External Member: **Dr. Siby Samuel**
Assistant Professor, Dept. of Systems Design Engineering,
University of Waterloo
- Internal-External Member: **Dr. Lisa Aultman-Hall**
Professor, Dept. of Systems Design Engineering,
University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of contributions

Several of the chapters of this thesis have already been published. Details of the publications are provided at the beginning of each chapter.

Abstract

Pedestrian exposure at intersections is a critical input for jurisdictions developing pedestrian-centric strategies and is typically quantified as the annual average daily pedestrian traffic (AADPT). The ability to estimate AADPT across all intersections depends on the availability of pedestrian volume data – specifically, the number of sites with continuous count (CC) stations, sites with only short-term counts (STCs), and sites with no available data. At sites with CC stations, AADPT can be directly calculated. Where only STCs are available, expansion factors derived from CC stations are used to expand STCs to AADPT. For sites lacking pedestrian volume data, Direct-Demand (DD) models are commonly employed to estimate pedestrian exposure based on land use, socioeconomic, and transportation attributes. This paper-based PhD thesis addresses gaps in the existing literature by exploring methods for estimating pedestrian exposure at intersections under varying data availability scenarios.

Objectives 1 and 2 focus on estimating pedestrian exposure at sites with no available data in jurisdictions that lack sufficient sites to locally develop DD models. Objective 1 examines the direct (or naïve) spatial transferability of DD models developed in other jurisdictions. Performance was found to be inconsistent, varying based on the similarity between the characteristics of the jurisdiction where the model was developed and where it was applied. In Objective 2, it is assumed that a limited number of CC sites are available within the study jurisdiction, enabling the adjustment (or local calibration) of transferred DD models. This approach improved the transferability of DD models, achieving performance levels comparable to those of locally developed models.

Continuing on jurisdictions that lack sufficient data to develop their own DD model, Objective 3 proposes a novel method to estimate pedestrian exposure leveraging AI Large Language Models (such as ChatGPT), satellite imagery, and pedestrian volume data from a few key sites. The performance of this method rivaled – and in some cases surpassed –

existing conventional methods like DD models, all without requiring complex statistical models or extensive datasets.

Objective 4 shifts the focus to sites where STCs are available, providing a detailed application of the expansion factor method to expand 8-hour STCs. Specifically, it examines grouping sites with CC stations (where expansion factors are derived) into factor groups based on similar activity patterns. To achieve this, temporal indicators capturing different types of seasonality were developed, and models were built to associate STC sites with an appropriate factor group. The results indicated only marginal improvements when using the factor group approach compared to applying a single factor group (i.e., the average across all sites).

Objective 5 examines the scenario in which CC stations are unavailable within a jurisdiction, preventing the local calculation of expansion factors. The spatial transferability of expansion factors across jurisdictions is investigated. The findings indicate that transferring expansion factors across jurisdictions with similar characteristics in terms of weather and school holiday periods proved to be of practical value, with only minor performance degradation compared to expansion factors developed within the local jurisdiction.

Objective 6 evaluates various methods for estimating pedestrian exposure under different data availability scenarios. The performance of these methods was analyzed based on the number of sites with available data, resulting in practical guidelines for practitioners. Specifically, it provides recommendations on the minimum number of sites with pedestrian volume data needed for the local development of DD models, compared to alternative approaches such as spatial transferability and local calibration of DD models.

Objective 7 introduces a novel approach to disaggregating pedestrian volumes from the intersection level to the crosswalk level. Land use models were developed and compared to methods that incorporate crosswalk-specific volume data from STCs. Overall, methods using STC data outperformed the land use models.

Acknowledgements

I am profoundly grateful to my wife, Maiara, for her unwavering support, encouragement, and partnership throughout this journey. Moving to Canada in the middle of a pandemic was a difficult and uncertain decision, but sharing this experience with you has made it all the more worthwhile.

I am deeply thankful to my brilliant supervisor, Prof. Bruce Hellinga, for his invaluable guidance and constant encouragement throughout my research. His supervision made my PhD journey pleasant, rewarding and enriching. I will always be grateful for this opportunity.

I would like to thank my family – Cla, Davi, Deusito, Dudu, Glauco, and Rafinha – who, although physically distant, have always made themselves present and supportive.

A heartfelt thank you to my lab friends, fellow grad colleagues, members of the UW ITE Student Chapter, and non-research-related friends for their support, stimulating discussions, and for making this journey more enjoyable.

I extend my sincere appreciation to my former supervisors, Prof. Flavio Cunto and Marcelo Luna, for their guidance before I came to Canada and for encouraging me to embrace this opportunity.

I am thankful to the committee members – Prof. Liping Fu, Prof. Samuel Siby, Prof. Lisa Aultman-Hall, and Prof. Patrick Singleton – for their time and willingness to serve on my committee and contribute to my thesis. I also thank Prof. Chris Bachmann for his guidance and contributions as a member of my comprehensive exam and throughout my journey.

Table of Contents

List of Figures	xiii
List of Tables	xvi
Chapter 1 Introduction	1
1.1 Scenarios of data availability	3
1.2 Research objectives	6
1.3 Thesis structure	8
Chapter 2 Literature Review	9
2.1 Expansion factor method.....	9
2.1.1 Factor groups	13
2.2 Alternative methods to estimate pedestrian exposure	15
2.3 Performance metrics: a brief discussion.....	17
2.4 Thesis roadmap: research gaps and objectives.....	19
Chapter 3 Comparing Direct Demand Models for Estimating Pedestrian Volumes at Intersections and Their Spatial Transferability to Other Jurisdictions	23
3.1 Overall context	23
3.2 Existing Direct Demand models	23
3.3 Spatial transferability of Direct Demand models.....	29
3.3.1 Jurisdictions chosen for the spatial transferability assessment.....	30
3.3.2 Summary of the applications of the DD models.....	34
3.4 Discussion and conclusions.....	37
Chapter 4 Enhancing the Spatial Transferability of Direct Demand Models for Estimating Pedestrian Volumes at Intersections	43

4.1 Overall context	43
4.2 Proposed local calibration models.....	44
4.3 Evaluation methodology	47
4.4 Generating synthetic AADPT data.....	48
4.5 Results and discussion.....	51
4.5.1 Local calibration using synthetic AADPTs	53
4.5.2 Local calibration using real AADPTs	61
4.6 Conclusions	63
Chapter 5 Estimating Pedestrian Volumes at Intersections using Artificial Intelligence: A ChatGPT Vision Approach	67
5.1 Overall context	67
5.2 Datasets	69
5.3 Can GPT-4V visually assess satellite images?.....	71
5.4 Ranking intersections: GPT-4V and satellite images.....	74
5.4.1 Method.....	74
5.4.2 Results	76
5.4.3 Exploring GPT-4V responses	79
5.5 Estimating pedestrian volume from rankings.....	86
5.5.1 Method.....	87
5.5.2 Results	90
5.5.3 Simulation approaches.....	94
5.6 Conclusions and practical application.....	97

Chapter 6 Expanding Short-Term Pedestrian Volume Counts to Average Annual Daily Counts at Intersections: Temporal Pattern Indices and a Comparison of Statistical Land Use and Simple Threshold Methods	100
6.1 Overall context	100
6.2 Data sources	101
6.3 Assessing types of expansion factors	103
6.4 Identifying factor groups	104
6.4.1 Visual inspection	105
6.4.2 Proposed temporal indicators for identifying factor groups	107
6.4.3 Clustering sites	109
6.4.4 Temporal indicator thresholds for identifying factor groups	112
6.5 Modeling factor groups	114
6.5.1 Multinomial logistic regression	117
6.5.2 Logistic regression and school-based approaches	120
6.6 Assessing the accuracy of expanding STCs to AADPT	122
6.7 Conclusions and practical recommendations	130
Chapter 7 Spatial Transferability of Expansion Factors for Estimating Pedestrian Volume at Intersections	133
7.1 Overall context	133
7.2 Study sites	134
7.3 Framework	136
7.4 Steps 1, 2, and 3: overview	137
7.5 Step 4: expanding STCs to AADPT in the target jurisdiction	139

7.5.1 Method.....	139
7.5.2 Results	141
7.6 Discussion and conclusions.....	142
7.6.1 Question 1: Does the application of expansion factors spatially transferred from another jurisdiction provide sufficiently accurate AADPT estimates to be of practical value?.....	145
7.6.2 Question 2: Does transferring models for associating STC sites with factor groups provide practical benefits over using a single average factor group when transferring expansion factors?	149
7.6.3 Question 3: How does the accuracy of spatially transferring expansion factors compare to that of alternative methods when continuous monitoring is not available?	151
7.6.4 Additional findings	153
7.6.5 Conclusions and practical recommendations	154
Chapter 8 Estimating Pedestrian Exposure at Intersections for Different Data Availability Conditions.....	157
8.1 Overall context	157
8.2 Study sites	158
8.3 Framework and method.....	159
8.3.1 Naïve transferability and local calibration of DD models.....	160
8.3.2 Local development of DD models.....	161
8.3.3 Simulating scenarios of data availability.....	164
8.4 Results and discussion.....	166
8.4.1 Observed AADPT for all sites.....	167
8.4.2 Simulation of data availability scenarios.....	170

8.5 Guidelines for practitioners	181
8.6 Conclusions	185
Chapter 9 Estimating Pedestrian Volumes at Each Crosswalk of Intersections: Comparison of Land Use Models and Short-Term Count Methods	188
9.1 Overall context	188
9.2 Method	190
9.2.1 Modeling the percentage share of pedestrian volume per crosswalk	191
9.2.2 Estimating the percentage share of pedestrian volume per crosswalk from STCs	198
9.2.3 Evaluation method	200
9.3 Results and discussion	202
9.3.1 Optimal land use model	205
9.3.2 Evaluation: LU model and STC method	207
9.4 Conclusions and recommendations	213
Chapter 10 Conclusions and Recommendations	216
10.1 Contributions	216
10.2 Recommendations for future research	218
References	220
Appendix A Steps 1, 2, and 3 from Chapter 7	228

List of Figures

Figure 1-1 Scenarios of data availability and method employed for each type of site	4
Figure 2-1 Scenarios of data availability and research objectives	20
Figure 3-1 Site location: (a) Milton, (b) Pima County, and (c) Downtown Toronto.....	32
Figure 3-2 Boxplots of selected explanatory variables for selected DD models.....	34
Figure 3-3 Estimated and observed AADPT for sites in Milton for the two most accurate DD models	38
Figure 3-4 Toronto: Singleton et al. model error as a function of observed AADPT	40
Figure 4-1 Method for estimating AADPT using local calibration	44
Figure 4-2 Results: local calibration using synthetic AADPTs	54
Figure 4-3 Scatterplots: estimated AADPT estimate vs synthetic AADPT	56
Figure 4-4 Performance of Models B and C when compared to the Base Model (negative values indicate improved accuracy vs Base Model)	60
Figure 5-1 Framework	69
Figure 5-2 Satellite images used for the preliminary investigation (29): (a) Erb at Caroline, depicting a scenario with mixed land use; (b) River at Manchester, depicting a scenario with mostly residential land use.....	72
Figure 5-3 Scatterplot: true rank versus GPT-4V rank.....	77
Figure 5-4 MADPT and true rank.....	78
Figure 5-5 Exploring GPT-4V's consistency in ranking the same six images	81
Figure 5-6 Exploring the bias in GPT-4V's responses	85
Figure 5-7 Method to estimate pedestrian volume from rankings.....	87
Figure 5-8 Pedestrian volume distribution across different jurisdictions	89
Figure 5-9 Determining if the Logarithmic model or linear interpolation approach should be adopted on the basis of field volumes.....	90
Figure 5-10 Performance assessment: estimating pedestrian volume from ranks	91
Figure 5-11 Predicted versus true MADPT using four categories and the logarithmic model approach: (a) all sites, (b) zoomed at sites with MADPT lower than 1,000.....	93

Figure 5-12 Performance assessment of original and simulated approaches	97
Figure 6-1 Proposed framework	101
Figure 6-2 Different temporal pedestrian activity patterns for three different sites. (a) Fairway Road at Wilson Avenue; (b) King Street at Mt. Hope Street; (c) Charles Street at Cameron Street	106
Figure 6-3 DOWOM expansion factors for two sites showing the effect of seasonality on pedestrian activity patterns. (a) Fairway Road at Wilson Avenue; (b) King Street at Highway 85 Ramp	107
Figure 6-4 DOWOM expansion factor plots for factor groups identified through K-means clustering	111
Figure 6-5 Decision tree: temporal indicator thresholds for identifying factor groups	113
Figure 6-6 DOWOM expansion factors per factor groups	128
Figure 7-1 Framework	137
Figure 7-2 Single Factor Method expansion factors	148
Figure 7-3 Single Factor Method versus Modeling Methods	150
Figure 8-1 Cumulative relative frequency of intersection AADPT	159
Figure 8-2 Method: Naïve transferability and local calibration of DD models	161
Figure 8-3 Example of scenario of data availability	164
Figure 8-4 Simulation process: summary of model development	172
Figure 8-5 Simulation process: performance across all sites. DD model transferred: Griswold et al.	174
Figure 8-6 Simulation process: performance across all sites. DD model transferred: Schneider et al.	175
Figure 8-7 Simulation process: performance across sites with no information. DD model transferred: Schneider et al.	176
Figure 8-8 Guidelines for estimating AADPT at intersections	182
Figure 9-1 Quadrant representation	192
Figure 9-2 Example of desire line between transit stops and schools	196
Figure 9-3 Comparison of the performance of each method	207

Figure 9-4 Likelihood that STC method estimate is more accurate than naïve method estimate	209
Figure 9-5 Scatterplot: observed and predicted shares per crosswalk	210

Figures in the appendix:

Figure A1 Resulting factor groups in each jurisdiction	230
Figure A2 WWI*, SHTI and factor groups in Waterloo	232

List of Tables

Table 3-1 DD models selected for transferability	25
Table 3-2 Variables included in each DD model.....	28
Table 3-3 Summary of DD model applications	36
Table 3-4 Summary of DD model performance	41
Table 4-1 Descriptive summary: sample size and AADPT	52
Table 4-2 Summary of the synthetic AADPT approach using 10% of sites for local calibration	58
Table 4-3 Summary of the real AADPT approach	62
Table 5-1 Confusion matrix: true rank versus GPT-4V rank	79
Table 5-2 Additional jurisdictions: descriptive statistics.....	88
Table 5-3 GPT-4V versus StreetLight	92
Table 6-1 Characteristics of each factor group based on their pedestrian volume	110
Table 6-2 Site characteristic variables and descriptive statistics for radius of 400m	115
Table 6-3 MLR coefficients and odds ratio	117
Table 6-4 Confusion matrix for MLR approach	120
Table 6-5 Logistic regression coefficients and odds ratios for factor group #3 (Primarily School Trips).....	121
Table 6-6 Confusion matrices for Logistic regression and school-based approaches	122
Table 6-7 Expanding STC to AADPT: summary of results	125
Table 7-1 Descriptive statistics of AADPT	135
Table 7-2 Accuracy of estimated AADPT (MAPE): STCs from candidate days only	143
Table 7-3 Accuracy of estimated AADPT (MAPE): STCs from all days	144
Table 7-4 Summary of AADPT estimation accuracy using the Single Factor Method – candidate days.....	146
Table 7-5 Change in AADPT estimation accuracy (MAPE) when using all days versus candidate days.....	148
Table 7-6 Single Factor Method compared to alternative methods	152

Table 7-7 Summary of AADPT estimation accuracy when base and target jurisdictions are the same – candidate days.....	154
Table 7-8 Impact of Modeling Methods on AADPT estimation accuracy when base and target jurisdictions are the same – candidate days	154
Table 7-9 DOWOM expansion factors: Single Factor Method.....	156
Table 8-1 Summary of DD models transferred from other jurisdictions.....	168
Table 8-2 Summary of naïve transferability and local calibration of DD models.....	168
Table 8-3 Summary of DD model development: statistics and model coefficients	170
Table 8-4 Summary: simulation process.....	179
Table 9-1 Candidate attributes to represent each category of radius-based variables	194
Table 9-2 Candidate interactions between categories.....	195
Table 9-3 Descriptive statistics.....	203
Table 9-4 Significant explanatory variables	205
Table 9-5 Summary of the performance of each method	208
Table 9-6 Average estimation error of each method considering upper and lower bounds on STC method accuracy	213

Tables in the appendix

Table A1 Summary of MLR and LR models.....	234
Table A2 Summary of school-based models	235
Table A3 Confusion matrices: transferring models to assign factor groups.....	236

Chapter 1

Introduction

According to the World Health Organization (WHO), road crashes have caused approximately 1.3 million deaths and 50 million injuries annually over the past 20 years (WHO, 2021). While vehicle occupants account for most road crash casualties, increasing investments in active transportation (Government of Canada, 2021; Martin et al., 2012), combined with the vulnerability of pedestrians in urban environments, have led to a concerning rise in pedestrian-involved crashes. Studies indicate that the risk of injury while walking is about four times higher than while driving (Elvik, 2009).

For instance, pedestrians made up 13% of total road fatalities in Canada between 2005 and 2009, a figure that rose to 17% between 2015 and 2019 (Transport Canada, 2009, 2019). Similarly, in the United States, 4,092 pedestrians were killed on roadways in 2009, increasing to 6,205 by 2019 (NHTSA, 2010, 2020). Consequently, the proportion of pedestrian fatalities among total roadway deaths grew from 12% to 17%.

A jurisdiction has a wide range of opportunities to improve pedestrian safety, from long-term educational initiatives to relatively simple infrastructure changes, such as installing raised crosswalks. At intersections – the focus of this study – the selection of optimal pedestrian countermeasures is typically preceded by a network screening process, in which sites are ranked based on a measure of risk (AASHTO, 2010). Developing such rankings depends on the availability of pedestrian exposure data at the study intersections (Islam et al., 2022; Pulugurtha & Sambhara, 2011; Stipancic et al., 2020).

The ability of jurisdictions to obtain pedestrian exposure data for the entire road network depends on the type and quantity of available data. Intersections can be classified into three cases based on data availability.

The first case is the presence of continuous count (CC) stations at the intersection. CC stations enable continuous long-term monitoring (i.e., 24/7 over months or years), providing a comprehensive understanding of traffic patterns and allowing for the direct calculation of

Annual Average Daily Pedestrian Traffic (AADPT) when at least one year of data is available.

The second case is the availability of short-term counts (STCs) at the intersection. STCs are typically collected for traffic signal design studies and often include turning movements for vehicles and cyclists, as well as the number of pedestrians crossing specific crosswalks. Data collection usually takes place on one or a few non-holiday weekdays during designated hours. STCs provide punctual estimates of pedestrian volume and may be influenced by hourly, daily, monthly, or seasonal patterns, potentially leading to inaccuracies when predicting long-term volumes, such as AADPT (Hankey et al., 2014; Humagain & Singleton, 2021; Medury et al., 2019). To minimize this issue, expansion factors derived from sites with CCs are used to convert STCs into AADPT. This is referred to as the expansion factor method.

The third case is the lack of relevant pedestrian volume data at the intersection and, therefore, includes situations where there are no counts available. Several methods have been employed to estimate pedestrian volumes at sites without relevant pedestrian volume data, including using traffic signal push-buttons as a proxy for pedestrian volume (Singleton & Runa, 2021), leveraging third-party crowd-sourced data (StreetLight, 2023), disaggregating regional origin-destinations volumes to the intersection level (Kuzmyak et al., 2014), or employing Direct-Demand (DD) models, the most commonly used method.

DD models are used to predict pedestrian volume as a function of attributes of the intersection's surroundings, such as land use, socioeconomic factors, and transportation characteristics (Griswold et al., 2019; Schneider et al., 2021; Singleton et al., 2021). Traditionally, DD models are developed using data from sites within the study jurisdiction where pedestrian volume data is available. For jurisdictions lacking sufficient data to develop their own models, applying DD models calibrated in other jurisdictions may be a practical alternative; however, this approach has not yet been investigated in the literature.

In efforts to characterize the availability of traffic data in U.S. jurisdictions, Grossman et al. (2019) surveyed 133 state, regional, and local transportation agencies and found that: a) 84% of the agencies did not have any CC stations monitoring pedestrian activity; and b) 32% of the agencies do not collect any type of pedestrian counts. These statistics indicate that most jurisdictions lack CC stations – preventing them from applying the expansion factor method – and that some do not collect any pedestrian volume data, making it challenging to estimate pedestrian exposure across all intersections.

1.1 Scenarios of data availability

Figure 1-1 illustrates eight scenarios of data availability and the corresponding method for estimating AADPT for each site type. Each circle represents all sites within a jurisdiction, with colors indicating the proportion of sites belonging to each type. The objectives of this research are closely linked to these scenarios and their associated methods. Some terms related to the DD models are defined below:

- Local development: the DD model is developed (i.e., model coefficients are calibrated) using data from the study jurisdiction.
- Naïve transferability: a DD model developed in another jurisdiction is applied to the study jurisdiction without modification.
- Local calibration: pedestrian volume data from a small number of sites in the study jurisdiction is used to modify (i.e., locally calibrate) a DD model developed in another jurisdiction.

Scenarios 1-A and 3-D represent extreme cases that are unlikely to be observed in real situations. Scenario 1-A has CC stations at all sites, enabling direct calculation of AADPT, while Scenario 3-D lacks any information, requiring the naïve transferability of DD models or the use of alternative methods, such as purchasing third-party crowd-sourced data (StreetLight, 2023).

Scenarios 2-A and 3-A feature STCs for all sites, meaning the expansion factor method should be applied. Scenario 2-A has a sufficient number of CC stations (at least five stations, as suggested by Nordback et al. 2019) to enable accurate calculation of expansion factors, while Scenario 3-A lacks continuous monitoring and thus requires the use of expansion factors developed in another jurisdiction. The spatial transferability of expansion factors has not yet been investigated in the literature.

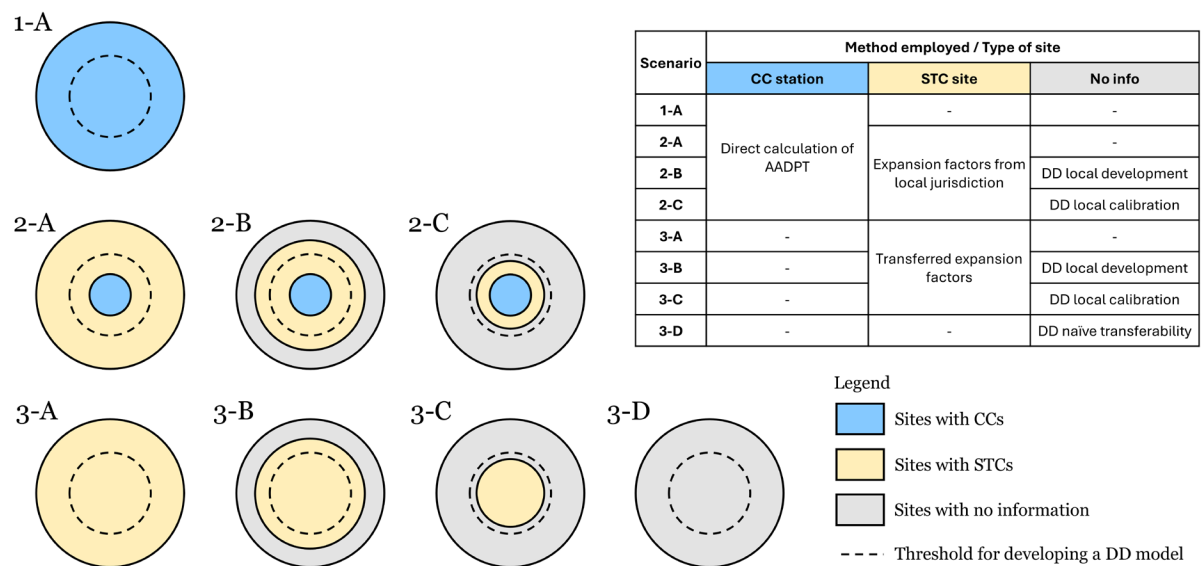


Figure 1-1 Scenarios of data availability and method employed for each type of site

One challenge associated with the expansion factor method is the existence of different pedestrian volume patterns across a jurisdiction (Hankey et al., 2014; Johnstone et al., 2018; Nordback et al., 2019) necessitates identifying sites with similar patterns, referred to as factor groups, to ensure the proper application of the expansion factor method. This approach enables more precise application of expansion factors, rather than simply averaging them across all sites with available CC stations. Identifying such patterns in sites with CC stations is straightforward due to continuous monitoring. The challenge arises when

associating STC sites – where temporal patterns are unknown – with factor groups. A common solution is to develop models linking land use, socioeconomic, and transportation attributes with factor groups to address this challenge (Humagain & Singleton, 2021; Medury et al., 2019).

Scenarios 2-B and 2-C both have sites without any relevant pedestrian count information. In these scenarios, the expansion factor method is applied to STC sites, and a DD model is used for sites without information. Scenario 2-B includes enough sites with either true or expanded AADPTs to enable local development of a DD model, whereas Scenario 2-C lacks sufficient sites with AADPT to permit local DD model development. This differentiation is represented by the dashed line. In the latter case, a DD model developed in another jurisdiction is transferred and locally calibrated using the available data. Literature indicates that DD models developed with 50-70 sites can achieve reasonable accuracy (i.e., $R^2 \sim 0.70$) for model calibration (Hankey et al., 2017; Munira et al., 2020; Sanders et al., 2017). Scenarios 3-B and 3-C follow the same logic, but with expansion factors transferred from another jurisdiction.

Some of the methods presented in Figure 1-1 have been extensively investigated in the literature, such as the expansion factor method and the development of DD models. However, certain methods require further documentation to strengthen the existing research, such as developing models to link STC sites to factor groups, while others have yet to be explored, including the spatial transferability of expansion factors and DD models. This research aims to investigate these less-explored methods.

Complementary, it is also in the scope of this research to propose a novel method for estimating pedestrian exposure at sites with no data and to evaluate alternatives for disaggregating pedestrian volumes from the entire intersection to the crosswalk level.

Throughout the thesis, the terms pedestrian activity, pedestrian exposure, and pedestrian volume are frequently used. Although they are sometimes used interchangeably, they are generally understood as follows: a) pedestrian activity refers to high-level or

macroscopic patterns, such as high, medium, or low activity; b) pedestrian exposure typically relates to annual averages, such as AADPT; and c) pedestrian volume refers to disaggregated measures, such as 15-minute or 8-hour counts.

1.2 Research objectives

This thesis has 7 research objectives as listed below. Much of the content in this thesis has already been peer-reviewed and published in journals or is currently under review. Each objective corresponds to a specific journal article, which is identified below.

1. Evaluate the spatial transferability of existing Direct-Demand models available in the literature for estimating pedestrian exposure at intersections.

Sobreira, LTP; Hellinga, B. 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, vol. 2677, issue 10, 2023. <https://doi.org/10.1177/03611981231161061>

2. Develop a method to improve the spatial transferability of Direct-Demand models by incorporating local pedestrian volume data and apply it to a sample of different jurisdictions.

Sobreira, LTP; Hellinga, B. Enhancing the Spatial Transferability of Direct Demand Models for Estimating Pedestrian Volumes at Intersections. *Transportation Research Record:*

Journal of the Transportation Research Board, vol. 2678, issue 7, 2023. <https://doi.org/10.1177/03611981231205881>

3. Leverage Artificial Intelligence Large Language Models as an alternative method for estimating pedestrian exposure at intersections.

Sobreira, LTP; Hellinga, B. Estimating Pedestrian Volumes at Intersections using Artificial Intelligence: A ChatGPT Vision Approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2024. <https://doi.org/10.1177/03611981241308873>

4. Employ the expansion factor method considering the development of land-use models to associate sites where only short-term counts are available with factor groups.

Sobreira, LTP; Hellinga, B. Expanding Short-Term Pedestrian Volume Counts to Average Annual Daily Counts at Intersections: Temporal Pattern Indices and a Comparison of Statistical Land Use and Simple Threshold Methods. *Transportation Research Record: Journal of the Transportation Research Board*, 2024. <https://doi.org/10.1177/03611981241281736>

5. Assess the spatial transferability of expansion factors for estimating pedestrian exposure at intersections.

Sobreira, LTP; Hellinga, B. Spatial Transferability of Expansion Factors for Estimating Pedestrian Volume at Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, under review.

6. Evaluate pedestrian exposure estimation under different data availability conditions to identify the most suitable method for each scenario and provide guidelines for practitioners.

Sobreira, LTP; Hellinga, B. Estimating Pedestrian Exposure at Intersections for Different Data Availability Conditions. *Transportation Research Record: Journal of the Transportation Research Board*, under review.

7. Develop and apply methods to disaggregate pedestrian volumes from the intersection level to the crosswalk level.

Sobreira, LTP; Hellinga, B. Estimating Pedestrian Volumes at Each Crosswalk of Intersections: Comparison of Land-Use Models and Short-Term Count Methods. *Journal of*

Transportation Engineering, Part A: Systems, vol. 150, issue 2, 2023.

<https://doi.org/10.1061/JTEPBS.TEENG-8141>

1.3 Thesis structure

The next Chapter of this thesis provides a literature review on existing methods for estimating pedestrian exposure at intersections and identifies current research gaps and how the seven research objectives of this thesis address those gaps. Chapters 3 to 9 are dedicated to addressing each of the research objectives, and Chapter 10 concludes the thesis with a summary of contributions and recommendations for future research.

Chapter 2

Literature Review

Chapter 2 presents a literature review on methods available for estimating pedestrian exposure at intersections. It begins with a discussion of the expansion factor method, followed by an overview of alternative methods. Next, performance metrics are examined. The chapter concludes with a summary of existing research gaps and their connection to the research objectives, serving as a roadmap for the Thesis.

2.1 Expansion factor method

Within the challenges of estimating AADPT for entire road networks, the conversion of STCs into AADPT using expansion factors is of great value. The U.S. Department of Transportation's Traffic Monitoring Guide (TMG) recommends a duration of 48 consecutive hours for STCs (FHWA, 2016). However, in practice, most readily available STCs come from traffic signal plan surveys, which include vehicular turning movement counts and pedestrian volumes per crosswalk (Allen, 2021). In Canadian jurisdictions, these surveys are typically collected over the course of eight hours on non-holiday weekdays covering the AM peak (7am-9am), mid-day period (11am-2pm), and PM peak (3pm-6pm) periods. To mitigate errors from using data from a single or few days to estimate long-term pedestrian volumes (i.e., AADPT), expansion factors derived from CC stations have been applied to account for hourly, daily, monthly, and seasonal patterns (Figliozi et al., 2014; L. Miranda-Moreno et al., 2013; Nosal et al., 2014; Orr, 2021).

The traditional method proposed by the TMG for estimating vehicular expansion factors – which is frequently used for non-motorized purposes – involves the calculation of 7 day-of-week (DOW) and 12 month-of-year (MOY) factors. Four steps are taken to convert short-term counts into AADPT estimations. First, AADPT is calculated for sites with CC stations (Equation 2.1). Then, DOW and MOY factors are calculated for each continuous

station (Equations 2.2 and 2.3). The resulting $DOW_{i,d}$ and $MOY_{i,m}$ factors can be averaged across all CC sites or grouped based on sites with similar pedestrian activity patterns; this is discussed in the next sub-section. Finally, AADPT is estimated for short-term count locations (Equation 2.4). This approach requires valid data for at least one day-of-week-of-month.

$$AADPT_i = \frac{1}{12} \sum_{m=1}^{12} \left[\frac{1}{7} \sum_{d=1}^7 \left(\frac{1}{n_{i,m,d}} \sum_{k=1}^{n_{i,m,d}} V_{i,k,m,d}^{24} \right) \right] \quad (2.1)$$

$$DOW_{i,d} = \frac{AADPT_i}{\frac{1}{n_{i,d}} \sum_{m=1}^{12} \left(\frac{1}{n_{i,m,d}} \sum_{k=1}^{n_{i,m,d}} V_{i,k,m,d}^x \right)} \quad (2.2)$$

$$MOY_{i,m} = \frac{AADPT_i}{\frac{1}{n_{i,m}} \sum_{d=1}^7 \left(\frac{1}{n_{i,m,d}} \sum_{k=1}^{n_{i,m,d}} V_{i,k,m,d}^x \right)} \quad (2.3)$$

$$AADPT'_i = V_{i,m,d}^x \times DOW_d \times MOY_m \quad (2.4)$$

Where:

$AADPT_i$ = observed (or true) AADPT of intersection i

m = months (1-12)

d = day of week (1-7)

$n_{i,m,d}$ = number of times a specific day of week occurs in a specified month

$V_{i,k,m,d}^{24}$ = daily (24 hour) pedestrian volume of site i on the k^{th} occurrence of month (m) and day of the week (d). For example, if there are three observations on Mondays in January, then k will range from 1 to 3

$DOW_{i,d}$ = day-of-week expansion factor of site i and day of the week (d). DOW_d represents the factor averaged across all sites or within groups with similar pedestrian activity patterns

$MOY_{i,m}$ = month-of-year expansion factor of site i and month (m). MOY_m represents the factor averaged across all sites or within groups with similar pedestrian activity patterns

$V_{i,k,m,d}^x$ = pedestrian volume of site i on the k^{th} occurrence of month (m) and day of the week (d) for the duration x of the STC ($x = 8$ -hour typically for STCs in Canadian jurisdictions)

$AADPT'_i$ = estimated AADPT at site i

Note that in cases where STCs contain less than 24 hours of observation, the term $V_{i,k,m,d}^x$ in either Equation 2.2 or 2.3 must correspond to the number of hours recorded in the STCs.

There are two alternative forms of expansion factors available that are more disaggregated than the DOW and MOY factors. The first consists of calculating factors across every day-of-week of every month-of-year, resulting in 84 day-of-week-of-month (DOWOM) factors (Equation 2.5). The second involves calculating individual factors for every day of the year (DOY – Equation 2.6).

$$DOWOM_{i,m,d} = \frac{AADPT_i}{\frac{1}{n_{i,m,d}} \sum_{k=1}^{n_{i,m,d}} V_{i,k,m,d}^x} \quad (2.5)$$

$$DOY_{i,t} = \frac{AADPT_i}{V_{i,t}^x} \quad (2.6)$$

Where:

$DOWOM_{i,m,d}$ = Day-of-week-of-month expansion factors of site i , month (m), and day of week (d)

$DOY_{i,t}$ = Day-of-year expansion factor of site i and day-of year t

DOY factors allow for the capture of volume oscillations on a daily basis, enabling the discernment of atypical days influenced by factors such as inclement weather or special events like concerts or sports events. Consequently, it has demonstrated better performance when compared to the DOW & MOY and DOWOM factors (Hankey et al., 2014; Nosal et al., 2014). Despite its advantages, practical application poses challenges including a) CC stations must operate continuously throughout the year; b) STCs must be collected during the same years as the long-term counts; c) the factors are likely not transferable across jurisdictions; and d) they introduce extra complexity when identifying similar pedestrian activity patterns across different sites (i.e., factor groups – discussed in the next sub-section).

The application of the expansion factor method for estimating pedestrian traffic volumes has been widely explored in the literature from various perspectives. Some established findings are outlined below:

- The application of the expansion factor method typically yields Mean Absolute Percent Error (MAPE) in the range of 15% to 30% (Hankey et al., 2014; Johnstone et al., 2018; Nordback et al., 2019).
- Collecting STCs in months where pedestrian activity is high (e.g., May through October in North American jurisdictions) reduces estimation errors (Hankey et al., 2014; Nordback et al., 2013).
- Longer STCs (e.g., 12-hour versus 2-hour or 7-day versus 1-day) reduce errors when performing the expansion (Hankey et al., 2014; Johnstone et al., 2018; Medury et al., 2019).

Despite the extensive literature available, one topic remains unexplored: the spatial transferability of expansion factors across different jurisdictions. This is particularly relevant for researchers and practitioners, as applying the expansion factor method with “local” expansion factors requires the presence of CC stations. Grossman et al. (2019) found that 84.2% of U.S. agencies lack CC stations for monitoring pedestrian activity. The author of the present thesis believes that a similar trend is likely observed worldwide. As a result,

jurisdictions cannot develop expansion factors unless they have or implement CC stations. The spatial transferability of expansion factors is examined as part of Objective 5 in Chapter 7.

An important consideration when applying the expansion factor method is the variation in pedestrian activity patterns across sites. For example, pedestrian activity at a site near a school is expected to decrease on weekends and during school holidays, whereas a site near a shopping mall may exhibit a more stable pattern year-round. These differences necessitate distinct expansion factors for each site, as averaging factors across both types of sites can degrade accuracy. This underscores the need to classify sites based on their pedestrian activity patterns.

2.1.1 Factor groups

The temporal pattern of pedestrian volume at an intersection, whether on an hourly, daily, or monthly basis, is influenced by factors such as land use, socioeconomic conditions, and operational features in its vicinity (Humagain & Singleton, 2021; Medury et al., 2019). To account for these variations, the expansion factor method requires classifying CC sites into factor groups, ensuring that sites within each group exhibit similar pedestrian activity patterns while patterns differ across groups.

Three approaches are commonly used in the literature to identify factor groups. The first is the visual classification of sites based on their temporal patterns (Medury et al., 2019; Turner et al., 2012). The second is the use of machine learning techniques to group similar temporal patterns (Hernández-Vega & Matamoros-Jiménez, 2021; Humagain & Singleton, 2021). The third – and most common – is the development of temporal indicators based on the CCs available (Hankey et al., 2014; Johnstone et al., 2018; Lindsey et al., 2015; L. Miranda-Moreno et al., 2013; Nordback et al., 2019). For instance, Miranda-Moreno et al. (2013) proposed two indicators to classify bicycle traffic patterns: WWI, relative index of weekend and weekday volume, and AMI, relative index of AM peak (7am-9am) and mid-day

period (11am-1pm) volume. Using different thresholds for these indicators, the authors categorized sites as Utilitarian, Mixed-Utilitarian, Mixed-Recreational, and Recreational. Although initially developed for bicycle volume patterns, studies on pedestrians have also used the same indicators. Johnstone et al. (2018) exclusively applied AMI, while Nordback et al. (2019) solely utilized WWI.

The primary limitation of existing temporal indicators to characterize different factor groups lies in their emphasis on measuring intra-day and weekday-to-weekend variations, without accounting for seasonal fluctuations throughout the year. For instance, consider two hypothetical sites where pedestrian traffic at one is influenced by the proximity of a high school, while the other is primarily determined by the presence of a manufacturing plant. It is expected that the AMI and WWI of both sites will be similar due to the higher pedestrian volume during the morning peak and on weekdays, as opposed to the mid-day period and weekends, respectively. However, it is likely that the site associated with the high school will experience reduced volume during school holidays, while the one linked to the manufacturing plant will likely maintain a relatively steady pattern throughout the year. Consequently, the expansion factors for both sites differ, requiring their classification into distinct factor groups. This distinction would not occur if only AMI and WWI were used.

Another challenge when dealing with factor groups is how to associate STC sites with a given factor group. Due to the absence of CCs, the long-term temporal pattern of pedestrian activity at STC sites is unknown. A common approach in previous studies involves using AMI, WWI, or both to link STC sites with a factor group. However, this approach has its limitations, namely: a) in cases where only brief STCs are available (e.g., for a few hours), AMI estimation is not feasible; b) when only one-day STCs are accessible, WWI cannot be estimated; and c) as discussed above, these indicators do not account for seasonal effects. To overcome these limitations, models that do not rely on empirical traffic data and are instead based on land use, socioeconomic, and operational attributes have been applied to associate STC sites with factor groups.

Medury et al. (2019) and Humagain & Singleton (2021) are two examples of pedestrian-centric studies that developed models to associate STC sites with factor groups. Methodologically, both studies are similar: they begin by identifying factor groups using sites with CCs. Then, they associate each factor group with land use (e.g., commercial use and presence of schools), socioeconomic (e.g., population density and household size), and operational (e.g., intersection density) attributes using multinomial logistic regressions (MLRs). Finally, they assess the accuracy of the expansion of STCs to long-term pedestrian volumes. The results from Medury et al. (2019) demonstrate benefits that range from 0.8% to 7.4% absolute reduction in MAPE when employing the MLR approach compared to a simple method using a single expansion factor averaged from all sites.

These two studies have made valuable contributions to the field. However, they have certain limitations and do not encompass all possible configurations of STCs. This creates an opportunity for further research on the topic. For instance, neither study used actual AADPT as their target measure. Medury et al. (2019) employed data from June to September, while Humagain & Singleton (2021) relied on pedestrian volume data extrapolated from traffic signal push buttons (Singleton & Runa, 2021). Furthermore, both studies focused on expanding hourly volume to weekly volume, with Humagain & Singleton (2021) also developing month-of-year factors. Consequently, there is uncertainty about the magnitude of error associated with expanding commonly available 8-hour STCs to estimate AADPT. This topic is explored as part of Objective 4 in Chapter 6.

2.2 Alternative methods to estimate pedestrian exposure

When CC stations or STCs are unavailable for specific sites within a jurisdiction, alternative methods can be used to estimate pedestrian exposure. While these methods are often less resource-intensive, they generally result in larger errors compared to the expansion factor method. Some of these alternative methods are described below.

One method involves hiring third-party companies that provide volume estimates based on crowd-sourced data. For example, the company StreetLight uses anonymized location data collected from various mobile apps to estimate pedestrian volume (StreetLight, 2023). From a jurisdictional perspective, using this type of data is convenient as it eliminates the need for direct data collection, though it may come with significant costs. In terms of performance, the method achieves MAPEs ranging between 71% and 151%, depending on the magnitude of pedestrian volume of the studied sites (sites with higher volume often exhibit lower error) (StreetLight, 2023).

Another method consists of employing pedestrian push-button activations as a proxy for pedestrian volume at signalized intersections. Singleton & Runa (2021) analyzed 10,000 hours of video from 90 signalized intersections in Utah, USA. The authors developed models to correlate the hourly frequency of push-button presses with hourly pedestrian volumes, achieving a linear correlation of 0.84. However, the authors did not assess the performance of using push-button activations to estimate AADPT.

A different method is based on the traditional urban planning process used to estimate trips across zones. Just as vehicle trips are assigned from zonal origins to the network, this approach allocates pedestrian trips to specific segments and intersections using trip assignment algorithms, such as shortest-path methods (Kuzmyak et al., 2014). This method is not commonly applied in practice.

A fourth and most commonly adopted method by practitioners is the use of Direct-Demand (DD) models to estimate pedestrian volume at sites with no available data. DD models are statistical models calibrated using sites with known pedestrian exposure, associating it with explanatory variables that are often readily available to jurisdictions, such as socioeconomic features (e.g., population and employment densities), land use attributes (e.g., the number of commercial and residential establishments), and geometric and operational factors (e.g., the presence of transit stops) (Griswold et al., 2019; Schneider et al., 2021; Singleton et al., 2021). Once developed, DD models allow practitioners to estimate pedestrian exposure at all sites within a given jurisdiction.

There is a substantial body of literature on DD models, with various models developed that consider different explanatory variables, calibration sample sizes, modeling techniques, and jurisdiction types. A detailed literature review on DD models is provided in Chapter 3.

One challenge associated with DD models is that their development requires jurisdictions to have pedestrian volume data (ideally in terms of AADPT) for a sufficiently large number of sites to support model development, including calibration and validation. For instance, researchers have developed DD models with an R-squared of 0.70 using calibration sample sizes ranging from 50 to 70 sites (Hankey et al., 2017; Munira et al., 2020; Sanders et al., 2017). Since some jurisdictions may lack the necessary data to develop their own DD models, a cost-effective alternative is to apply an existing model from the literature.

The spatial transferability of DD models is an area that has yet to be explored in the literature. The naïve transferability of DD models (i.e., direct application to the study jurisdiction) is examined as part of Objective 1 in Chapter 3. Additionally, the local calibration of DD models (i.e., using a small number of sites in the study jurisdiction to adjust transferred models) is explored in Objective 2 in Chapter 4.

It is also in the scope of this Thesis to propose and implement a cost-effective and novel alternative method for estimating pedestrian exposure at intersections where the expansion factor method cannot be applied. The proposed method leverages AI Large Language Models (such as ChatGPT), satellite imagery, and pedestrian volume counts at key sites (Objective 3, Chapter 5).

2.3 Performance metrics: a brief discussion

One subject often overlooked in estimating pedestrian exposure (and other engineering applications) is the adequacy of performance metrics and their comparability across studies conducted in different jurisdictions. Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are among the most commonly used performance

metrics in traffic engineering studies. In the context of estimating AADPT, Equations 2.7 and 2.8 present the calculation of MAE, while Equations 2.9 and 2.10 define MAPE.

$$AE_i = |AADPT_i - AADPT'_i| \quad (2.7)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n AE_i \quad (2.8)$$

$$APE_i = \frac{AE_i}{AADPT_i} \times 100 \quad (2.9)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n APE_i \quad (2.10)$$

Where:

AE_i = Absolute error of intersection i

n = Number of sites

MAE = Mean absolute error

APE_i = Absolute percent error of intersection i

$MAPE$ = Mean absolute percent error

MAE and MAPE are particularly valuable because they provide error measures that are easy to interpret. However, their magnitudes are highly sensitive to the level of pedestrian activity in each jurisdiction. For example, an MAE of 100 has little meaning without knowing the average magnitude of the studied variable. It may indicate good accuracy if the magnitude is around 1,000 but poor accuracy if it is around 100. A similar issue arises with MAPE: entities with low values of the variable (i.e., the denominator in Equation 2.9) can bias the calculation of the metric.

With that said, there are alternatives to make these metrics more meaningful and comparable across different studies and/or jurisdictions. One option is to present MAPE and MAE separately for volume categories (e.g., sites with low, medium, and high volumes) (StreetLight, 2023). Another is to calculate the ratio of MAE to the average observed AADPT in the jurisdiction (Metric E – Equation 2.11). A third option is to use metrics like the Geoffrey E. Havers (GEH) statistic, which incorporates both absolute and relative differences but produces results that are less intuitive to interpret.

$$E = \frac{MAE}{\bar{Y}} = \frac{MAE}{\frac{1}{n} \sum_{i=1}^n AADPT_i} \quad (2.11)$$

Where:

$\bar{Y}$ = Average of observed AADPTs

E = Indicator to represent the ratio between MAE and $\bar{Y}$

Throughout some Chapters of this Thesis, it will become evident that different performance metrics can lead to varying results and conclusions, particularly when assessed in their “raw” form. Whenever possible, metrics that enable comparisons across jurisdictions have been included.

2.4 Thesis roadmap: research gaps and objectives

Research gaps for Objectives 1 to 5 (out of seven objectives) were identified in the previous sub-sections. Referring back to the different scenarios of data availability presented in Figure 1-1, Figure 2-1 maps the objectives of this Thesis to these scenarios. Objectives 6 and 7 are not depicted in Figure 2-1 but are discussed at the end of this section.

Objective 1 (Chapter 3) focuses on Scenario 3-D, where no sites with pedestrian volume information are available. This objective assesses the spatial transferability of DD models – an area that has yet to be explored in the literature.

Objectives 2 and 3 (Chapter 4 and Chapter 5) address scenarios where certain sites lack pedestrian volume data (Scenarios 2-C and 3-C). In these cases, the number of available sites with pedestrian volume data (i.e., CC sites + STC sites) is insufficient for developing DD models. Therefore, alternative methods are explored to estimate pedestrian exposure specifically for sites with no data. Objective 2 focuses on using the limited number of available sites with pedestrian volume data to locally calibrate DD models. In Objective 3, a novel method is proposed to estimate pedestrian exposure at sites with no data by integrating Large Language Models, satellite imagery, and pedestrian volume data from key reference sites.

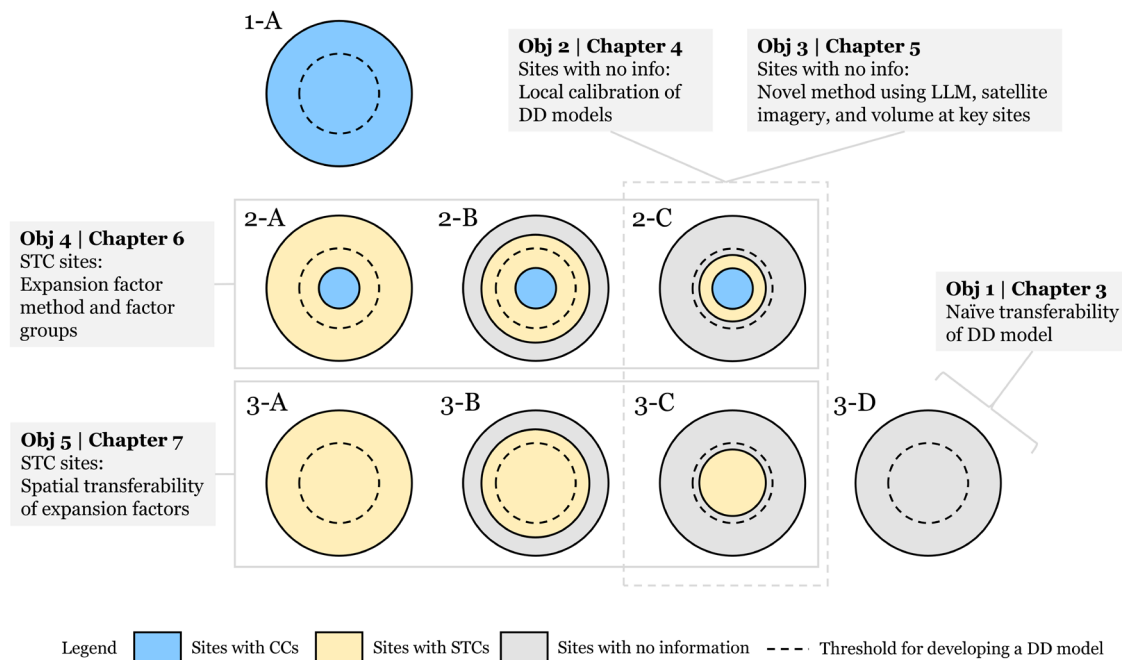


Figure 2-1 Scenarios of data availability and research objectives

Note: LLM = Large Language Model

Objective 4 (Chapter 6) examines the application of the expansion factor method using factor groups for STC sites in Scenarios 2-A, 2-B, and 2-C. Specifically, it targets the literature gap on developing land use models to assign STC sites to factor groups when 8-hour STCs are used.

Objective 5 (Chapter 7) also examines the expansion factor method, focusing on jurisdictions without CC stations to develop local expansion factors (Scenarios 3-A, 3-B, and 3-C). It addresses the gap in the literature regarding the spatial transferability of expansion factors across different jurisdictions for estimating pedestrian exposure.

Objective 6 (Chapter 8) emerged from the necessity of a more comprehensive assessment of the scenarios presented in Figure 2-1. The methods discussed in this section (i.e., the expansion factor method and the local development, local calibration, and naïve transferability of DD models) have been explored in the literature (and are analyzed in the previous objectives), but primarily as standalone methods. However, in Scenarios 2-B, 2-C, 3-B, and 3-C – scenarios that likely reflect the reality of most U.S. and Canadian jurisdictions – these methods are applied simultaneously, raising unresolved questions. For instance, how does the performance of each method vary with the number of sites with available data? What is the threshold number of sites with available data to determine whether to pursue DD model development or local calibration? Lastly, how do spatially transferred expansion factors influence DD model development and local calibration compared to those derived directly from the target jurisdiction? Objective 6 seeks to evaluate the performance of different methods for estimating pedestrian exposure at intersections under varying conditions of data availability.

Objective 7 (Chapter 9) arose from the need to disaggregate pedestrian volume at intersections to the crosswalk level for certain applications. For example, the 7th edition of the Highway Capacity Manual presents a methodology to estimate pedestrian level of service at signalized intersections, where one of the inputs is the pedestrian volume in each crosswalk (NASEM, 2022). Similarly, safety analyses often require disaggregation by

movement. For instance, to evaluate the conflicting movement between left-turn vehicles and pedestrians, pedestrian volumes for each crosswalk are necessary.

When CC stations are available at a given site, AADPT for each crosswalk can be directly calculated. However, when only STCs are available, only a point estimate of the percentage share per crosswalk can be observed. More critically, when no data is available and alternative methods (e.g., DD models) are employed, no information on the distribution across crosswalks exists. To the author's knowledge, no research has yet addressed these issues. Therefore, Objective 7 a) explores the reliability of using STCs to estimate pedestrian volume shares per crosswalk and b) proposes and applies models that estimate these shares based on land use, socioeconomic, and operational attributes.

Chapter 3

Comparing Direct Demand Models for Estimating Pedestrian Volumes at Intersections and Their Spatial Transferability to Other Jurisdictions

This Chapter is based on the following journal paper:

Sobreira, LTP; Hellinga, B. 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, vol. 2677, issue 10, 2023. <https://doi.org/10.1177/03611981231161061>

3.1 Overall context

This chapter focuses on assessing the spatial transferability of DD models across different jurisdictions. More specifically, its objectives are to a) identify and qualitatively assess existing DD models and b) evaluate their spatial transferability for estimating AADPT at signalized intersections.

This chapter is divided into two parts. The first part identifies and evaluates state-of-the-art DD models to select those for the spatial transferability task. The second part examines the spatial transferability of the selected models by applying them to sites in three different jurisdictions where AADPT is known.

3.2 Existing Direct Demand models

This section presents an assessment of DD models available in the literature, aiming at the selection of models calibrated in different contexts for the development of the spatial

transferability task. In total, twenty studies (models) published in the past two decades were reviewed. Each model from the literature was assessed with respect to the following three criteria: a) the number of sites in the calibration dataset; b) the model's complexity, represented by the number of explanatory variables; and c) the level of pedestrian activity observed in the original jurisdiction, represented by the calibration dataset's average AADPT. The application of these criteria resulted in six DD models that were then examined with respect to their spatial transferability.

Among the diversity of existing DD models, four studies stand out for having used more than a thousand sites for model calibration (Griswold et al., 2019; Kim et al., 2019; L. F. Miranda-Moreno & Fernandes, 2011; Singleton et al., 2021). The number of sites employed in these research projects is much higher than that of other studies, which typically use between 50 and 200 sites. Kim et al. (2019) developed a DD model using information from 10,000 sites in Seoul, South Korea. However, the authors employed a geographically weighted regression, where the model coefficients vary according to their location, making it impossible to transfer them to other jurisdictions. Miranda-Moreno & Fernandes (2011) calibrated a DD model for Montreal, Canada based on data from 1,018 signalized intersections. The authors used 8-hour pedestrian volume as the model outcome rather than the more common AADPT or annual volumes, making comparison with other models difficult. For those reasons, only Griswold et al. (2019) and Singleton et al. (2021) were chosen for the transferability task from the studies that applied the largest sample sizes.

Regarding the model's complexity, a common number of explanatory variables varies between five and eight. Four of the selected DD models are within this range, and two of them exemplify extreme situations where three and fourteen variables were employed. To represent jurisdictions with different levels of pedestrian activity, studies that exhibited an average AADPT of 186 to 2,433 were chosen for the transferability task. Table 3-1 details the information for each of the six selected studies, including the model coefficients.

Table 3-1 DD models selected for transferability

Study	Variable	Unit	Buffer radius	Model coefficient
Griswold et al. (2019) California State N = 1,270 Log-linear model Annual pedestrian volume Avg. True AADPT = 2,433 (Median = 438)	Number of employees ^{1, 2}	Count	0.25 mi	0.39
	Population	Count	0.50 mi	0.000142
	Street segments ^{1, 2}	Count	0.50 mi	0.302
	Walk commute mode share	% (decimal)	0.50 mi	2.84
	Schools ^{1, 2}	Count	0.50 mi	0.0444
	Int is on a major arterial	0, 1 or 2 ³	-	0.457
	Int is on a minor arterial	0, 1 or 2 ³	-	0.384
	Four-way intersection	0=No; 1=Yes	-	0.413
	Constant	-	-	5.58
Hankey et al. (2017) Blacksburg, USA N = 72 Log-linear model AADPT Avg. True AADPT = 192	Sidewalk length	Meters	750 m	0.000078
	Off-street trail	Meters	100 m	-0.004
	HH income	Dollars	1,750 m	-0.000016
	Residential address	Count	1,000 m	-0.00062
	Population density	Pop per sq-km	750 m	0.00017
	Transit stops	Count	250 m	0.13
	Constant	-	-	5.1
Munira et al. (2020) Austin, USA N = 44 Negative Binomial AADPT Avg. True AADPT = 605	Trail length	Feet	0.50 mi	0.0000637
	Commercial places	Count	0.10 mi	0.0239
	Population under 5 years	Count	0.50 mi	-0.00372
	Population working at home	Count	0.10 mi	0.061
	Transit stops	Count	1.00 mi	0.00896
	Constant	-	-	4.088
Sanders et al. (2017) Seattle, USA N = 50 Poisson model Annual pedestrian volume Avg. True AADPT = Not available	Residential address	Count divided by 10,000	0.25 mi	0.876
	Commercial places	Count	0.25 mi	0.0097
	Presence of university	0=No; 1=Yes	0.25 mi	0.4468
	Constant	-	-	12.9496
Schneider et al. (2021) Milwaukee, USA N = 260 Negative Binomial Annual pedestrian volume Avg. True AADPT = 186	Population density	Pop per sq-mile	400 m	0.00014
	Employment density	Emp per sq-mile	400 m	0.000021
	Transit stops	Count	100 m	0.336
	Retail places	Count	100 m	0.108
	Restaurants and bars	Count	100 m	0.116
	Presence of schools	0=No; 1=Yes	400 m	0.515
	Household with zero vehicle	% (decimal)	400 m	5.307
	Constant	-	-	8.334

Table 3-1 DD models selected for transferability (continued)

Study	Variable	Unit	Buffer radius	Model coefficient
Singleton et al. (2021) Utah, USA N = 1,494 Log-linear model AADPT Avg. True AADPT = 267	Population density ¹	Pop. per sq-mile divided by 1,000	0.50 mi	0.326
	Employment density ¹	Emp. per sq-mile divided by 1,000	0.25 mi	0.124
	Household size ¹	Person per HH	0.25 mi	0.418
	Household income	Annual income in \$1,000	0.50 mi	-0.01
	Vehicle ownership	Vehicles per HH	0.50 mi	-0.198
	Residential address	Land use % (integer)	0.25 mi	0.006
	Commercial places	Land use % (integer)	0.25 mi	0.019
	Intersection density	Intersection per sq-mile	0.25 mi	0.004
	4-way intersections	% (integer)	0.50 mi	0.006
	Schools	Count	0.25 mi	0.155
	Worship places	Count	0.50 mi	0.06
	Transit stops	Count	0.25 mi	0.068
	Park acreage	Area in acres	0.50 mi	0.022
	Int is on a major road	0=No; 1= Yes	-	0.242
	Constant	-	-	2.747

Note: ¹Ln-transformed variables. ²The authors of the original paper added a constant value of 0.001 to ln-transformed variables to avoid null values. ³Represents the number of major (or minor) arterials that cross the intersection.

Table 3-2 presents a summary of the variables that were significant, and therefore included, in each model, divided into three categories: census-based, land use, and geometry and operation variables. Regarding census-based variables, it is observed that population and employment are important features to represent the production and attraction of trips, respectively. Besides this, the variables that characterize income and vehicle ownership indicate that increases in the level of motorization are associated with a reduced generation of walking trips. Concerning land use attributes, the presence of commercial establishments, schools, and universities in the intersection's surroundings is linked to increases in pedestrian activity. The contribution of residential land use (e.g., sign of the coefficient) is not consistent among the models. In some models (e.g., Sanders et al. (2017) and Singleton et al.

(2021)) the sign of the coefficient is positive indicating that increases in the number of residential addresses results in larger pedestrian volumes. In Hankey et al. (2017) model, the coefficient of residential land use is negative indicating that if the land use around an intersection is predominantly residential, the production and attraction of walking trips may be reduced. Studies have suggested that incorporating variables that represent the land use mix could be beneficial for characterizing pedestrian activity (Ewing & Cervero, 2010). In terms of geometric and operational characteristics, the presence of transit stops is certainly a generator of walking trips. Furthermore, four-way intersections and those located on major or minor arterials are associated with more pedestrian trips than three-way intersections and those located on collector streets, respectively. Similar tables to Table 3-2 can be found in Singleton et al. (2021), Schneider et al. (2021), and Munira & Sener (2017).

In respect to the model's formulation, log-linear, Poisson, and Negative Binomial structures are observed in the six chosen models. All these formulations apply a link function (natural logarithm transformation) to the dependent variable. The general formulation of DD models is presented in Equation 3.1. To convert the ln-transformed variable to AADPT or annual volume, both sides of the equation are exponentiated (Equation 3.2). Note that Griswold et al. (2019) and Singleton et al. (2021) applied a ln-transformation to some of the independent variables.

Table 3-2 Variables included in each DD model

Variable	Hankey et al.	Sanders et al.	Griswold et al.	Munira et al.	Schneider et al.	Singleton et al.
Census based						
Population (count or density)	+		+		+	+
Population under 5 years (count)				-		
Population working at home (count)				+		
Employment (count or density)			+		+	+
Household size (avg)						+
Household income (\$)	-					-
Household with zero vehic (%)					+	
Vehicle ownership (avg)						-
Walk commute mode share (%)			+			
Land use						
Commercial (count, area or %)		+		+	+	+
Residential (count, area or %)	-	+				+
Restaurants and bars (count)					+	
Park area						+
Worship places (count)						+
Schools (count or presence)			+		+	+
University (presence)		+				
Geometry and operation						
Street segments (count or length)			+			
Off-street trail length	-			+		
Sidewalk length	+					
Four-way intersection (0/1 or %)			+			+
Intersection density						+
Major arterial road (0/1)			+			+
Minor arterial road (0/1)			+			
Transit stops (count)	+			+	+	+

Note: association with pedestrian volume: (+) positive; (-) negative.

$$\ln(Y'_i) = \beta_0 + \beta X_i + \gamma \ln X_i^* \quad (3.1)$$

$$Y'_i = e^{(\beta_0 + \beta X_i)} \times X_i^{*\gamma} \quad (3.2)$$

Where:

Y'_i = estimated AADPT or annual pedestrian volume at intersection i .

X_i = vector of non-transformed explanatory variables associated with intersection i .

X_i^* = vector of explanatory variables associated with intersection i that are ln-transformed.

β_0 = model constant.

β and γ = vector of model coefficients.

To the authors' knowledge, no efforts regarding the spatial transferability of existing DD models have been made yet. To contribute to this knowledge gap, an application of the six chosen models to three different jurisdictions is presented and discussed in the next sections.

3.3 Spatial transferability of Direct Demand models

With the knowledge of the explanatory variables and model coefficients from the selected models, their application to other jurisdictions depends only on the availability of the data for the explanatory variables. As indicated in Table 3-2, these explanatory variables generally fall into one of three categories, namely Census, Land Use, and Geometric and Operational. These data are typically available from national census data sources, jurisdiction open data sites, or regional transportation survey data sources. Regarding the data collection, there are two categories of variables: count/length and area based. The assemblage of the first is straightforward by using GIS functions to count points or sum lengths within the designated buffer areas around the site. The second category is often associated with census and transportation survey information, where the data is gathered into zones. Two strategies have been considered to deal with this kind of data: a) taking the arithmetic average of the

variable for every zone that is intersected by the buffer; and b) taking the weighted average of the variable based on the buffer intersection area. Some authors explicitly identified which strategy they followed: Schneider et al. (2021) used an arithmetic average, and Hankey et al. (2017) applied a weighted average. When no information is provided, the weighted average is assumed.

To assess the accuracy of the DD model application, it is also necessary to know the true (or observed) pedestrian exposure for each site. We have chosen to select sites and jurisdictions for which CCs are available so that AADPT can be directly computed (e.g., one year of CCs available) or estimated from time series of daily counts across a relatively large number of days. This avoids the introduction of errors from using one- or two-day STCs and expanding these to AADPT using expansion factors from CC stations. Three indicators were employed to measure the errors between observed and estimated AADPTs: MAE (Equation 2.8), MAPE (Equation 2.10), and the ratio of MAE and the average true AADPT (E, Equation 2.11).

3.3.1 Jurisdictions chosen for the spatial transferability assessment

Three jurisdictions, capturing a diverse range of geography, land use, urban form, population density, and climate were chosen for the DD model application: City of Milton, Canada; Pima County, USA; and City of Toronto, Canada. Each of these three jurisdictions has deployed a camera-based traffic monitoring system from the same vendor. The pedestrian volume counts (aggregated to intersection daily totals) were obtained from this system for each site. Counts from time periods that were substantially impacted by the COVID-19 pandemic were avoided. Due to jurisdiction specific COVID-19 impacts and particularities in the data, a different strategy was necessary to estimate AADPT for each jurisdiction:

- Milton ($n = 21$): Continuous count data were available from 6 sites from more than one year. Day-of-week and month-of-year expansion factors were developed from these 6

sites and applied to 15 sites, from which the AADPT was estimated from samples ranging between 31 and 258 days. The counts ranged from June-2018 to March-2020. No particular pattern was observed in the pedestrian volume of the sites employed for expansion factor development, hence no factor grouping was applied.

- Pima County (n = 81): Only three sites had sufficient data for the development of expansion factors in a pre-pandemic period (prior to March 2020), and a significant portion of the sites have counts only for the period after March 2020. It was decided to estimate the Seasonal Average Daily Pedestrian Traffic for the period between April-2021 and November-2021 and to expand it to AADPT using factors obtained from the three sites with complete pre-pandemic data from October-2018 to February-2020. By April-2021 most of the COVID-19 restrictions had been lifted in Pima County, including the school restrictions. The expansion factoring considered the presence of schools to appropriately deal with the reduced pedestrian volume in June and July due to school holidays.
- Toronto (n = 28): the data available allowed for the direct calculation of AADPT for all the sites using information from January-2018 to March-2020.

Figure 3-1 shows the location of the sites (signalized intersections) in each jurisdiction with an indication of the AADPT estimate. It is observed that the level of pedestrian activity differs significantly in each jurisdiction. Milton, for example, presents an average AADPT of 313, which is within the range of values used to calibrate existing DD models (Table 3-1). On the other hand, Pima County and Toronto represent two different extremes. The sites from Pima County have extremely low pedestrian activity (average AADPT = 28), with 30% of the sites comprise daily pedestrian volumes between two and seven. The sites from Toronto are situated in Downtown Toronto, a region that is characterized by its high employment and commercial densities, which are reflected in a significant pedestrian movement (average AADPT = 23,481). The choice of such distinct

jurisdictions was intention to provide the opportunity to assess the transferability of DD models in different contexts.



Figure 3-1 Site location: (a) Milton, (b) Pima County, and (c) Downtown Toronto

To verify to what extent the use of post-pandemic data affected the low pedestrian volume observed in the Pima County sites, Monthly Average Daily Pedestrian Traffic (MADPT) was calculated from October-2018 to February-2020 (pre-pandemic period) and compared to MADPT for the period considered in the study (from April-2021 to November-2021). Using data from nine sites, an average reduction of 5.65% of the MADPT was observed for the post-pandemic period. This is an indication that pedestrian activity had returned to pre-pandemic levels during the period of the study.

Figure 3-2 presents the distribution of several key explanatory variables for each jurisdiction. Figure 3-2 also shows the mean, minimum, and maximum statistics from the jurisdiction on which the model was originally developed. The distribution of the variables follows the level of pedestrian activity found for each jurisdiction (i.e., Pima County < Milton < Toronto). Of particular note are the extremely large population and employment densities of the Downtown Toronto sites. In addition, the percentages of people who walk to work and of households that do not own any vehicles in Downtown Toronto are also high – despite the high median income associated with the region – suggesting that the non-motorized transportation mode prevails in the region. It is also noted that some variables of the Toronto sites are outside of the range captured in the original DD model calibration datasets, which could cause issues when the DD models are applied. The variables for Milton and Pima County, in general, fall within the range of values of the DD model calibration datasets.

Two considerations are made regarding the explanatory variables used in the model's applications. The first is that one of the variables employed in (Hankey et al.2017) and Singleton et al. (2021) is household income. Because both models were developed in U.S. jurisdictions, US Dollars are the initial unit used. For the model's application to Milton and Toronto, data sources provide household income in Canadian Dollars. In the application of these models to the Canadian jurisdictions, household incomes were not converted to U.S. Dollars, as it is assumed that the variable in the original model has the objective of capturing the purchasing power of the population. The second point is that the models proposed by

Hankey et al. (2017) and Sanders et al. (2017) considered the count of residential addresses. The authors did not clearly state in their studies if multi-unit residential buildings are counted as one or as multiple addresses. Many open data sources report a single address per building regardless of the number of individual units in the building and consequently, the present work counted every residential address as one point.

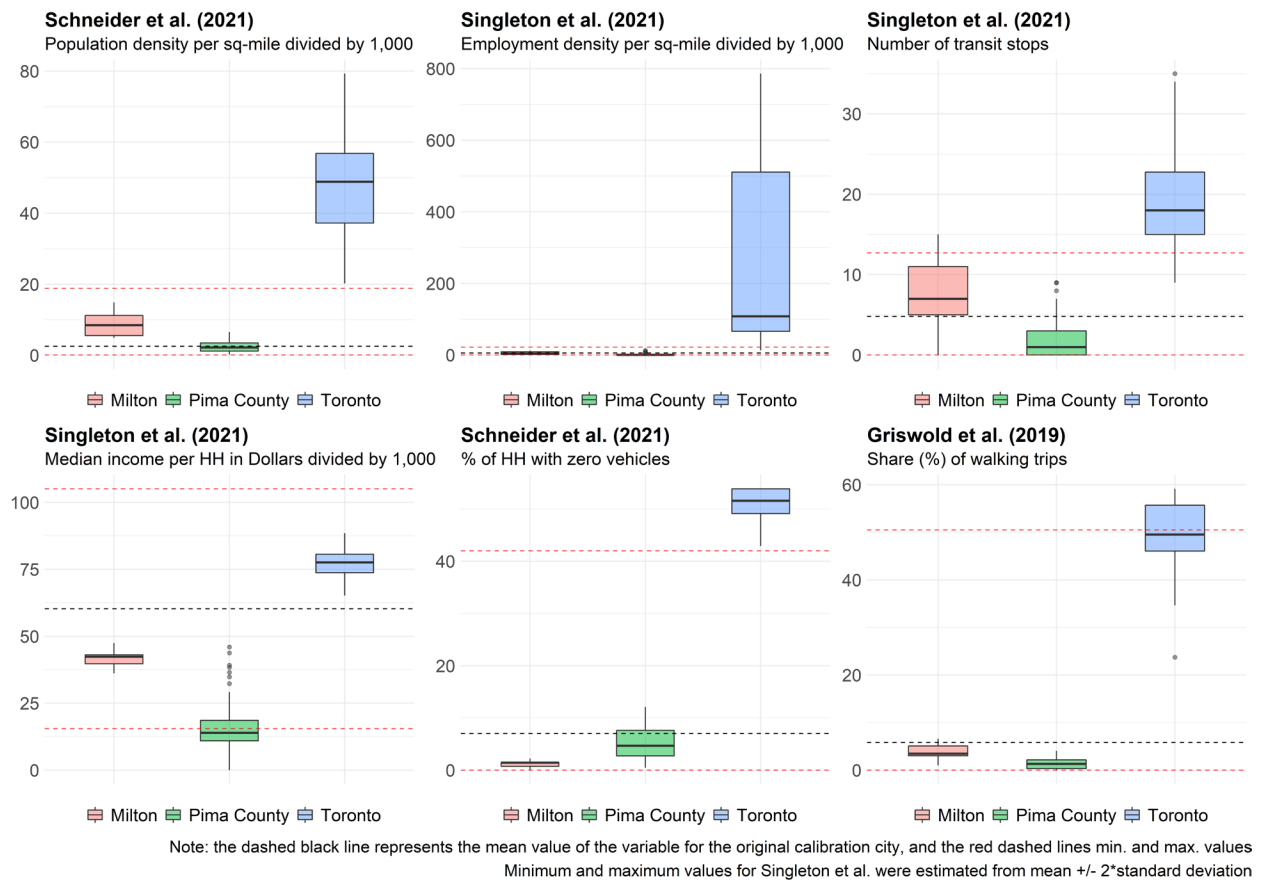


Figure 3-2 Boxplots of selected explanatory variables for selected DD models

3.3.2 Summary of the applications of the DD models

Table 3-3 presents a summary of the applications of the DD models to the sites in Milton, Pima County, and Toronto. Four measures of performance are shown: the average of

the AADPT estimate for all sites in the jurisdiction, MAE, the ratio of MAE to the average true AADPT, and MAPE. The models that estimate annual pedestrian volume were converted to AADPT. The table also provides the performance measures for the original study, whenever available, to compare the magnitude of errors between the application and the original model. We note that limited information on the goodness-of-fit of the original models was provided in the studies, particularly for conventional error indicators, such as MAE and MAPE.

In computing the summary performance indications presented in Table 3-3 we have made adaptations to two of the DD models. The first, indicated by footnote 3 in Table 3-3, was to constrain the value of the “Park Area” explanatory variable in the Singleton model to a maximum value of 75 acres. Examination of the model performance showed that for a small number of sites in Pima County, the Park Area was significantly larger than 75 acres, but these large values produced extremely large model estimates of pedestrian volumes that were clearly unrealistic.

The second adaptation, indicated as footnote 4 in Table 3-3, was to constrain the value of the “Number of people working from home” explanatory variable in the Munira model to a maximum of 100. For some sites in Toronto, the number of people working from home was much larger than 100 and for these sites the model predicted extremely large pedestrian volumes that were clearly unrealistic.

Both adaptations were made to avoid including model estimation errors that were clearly unrealistic and would not be used in practice. The threshold values were selected by examining the distribution of values of the explanatory variables and the resulting model estimates, but we recognize that establishing these constraints is subjective.

Table 3-3 Summary of DD model applications

Study	Original study				Milton (n = 21) Avg. True AADPT = 313				Pima County (n = 81) Avg. True AADPT = 28				Toronto (n = 28) Avg. True AADPT = 23,481			
	Avg. AADPT	MAE	MAE/ AADPT	MAPE	¹ Avg. AADPT	MAE	² MAE/ AADPT	MAPE	¹ Avg. AADPT	MAE	² MAE/ AADPT	MAPE	¹ Avg. AADPT	MAE	² MAE/ AADPT	MAPE
Griswold et al. (2019)	2,433	NA	-	NA	265	164	0.52	81%	91	69	2.46	574%	101,925	80,789	3.44	662%
Hankey et al. (2017)	192	NA	-	NA	329	244	0.78	117%	172	145	5.18	1440%	5,448	18,190	0.77	74%
Munira et al. (2020)	605	379	0.63	NA	138	243	0.78	68%	74	57	2.04	675%	38,473 ⁴	40,721	1.73	99%
Sanders et al. (2017)	NA	NA	-	39%	2,668	2,355	7.52	1611%	1,295	1,267	45.25	13013%	89,602	86,847	3.70	593%
Schneider et al. (2021)	186	120	0.65	NA	179	205	0.65	69%	47	29	1.04	224%	Inf.	Inf.	Inf.	Inf.
Singleton et al. (2021)	267	NA	-	NA	360	166	0.53	86%	112 ³	87	3.11	465%	13,185	15,461	0.66	86%

¹Average AADPT for the application.

²MAE divided by the Average True AADPT.

³A maximum value constraint of 75 acres to the park area was introduced.

⁴A maximum value constraint of 100 people to the population working at home was introduced.

3.4 Discussion and conclusions

The results of the spatial transferability analysis presented in Table 3-3 show that a) the performance of each DD model varied substantially across the different jurisdictions and b) the performance of the six DD models within a single jurisdiction also differed substantially. The model developed by Sanders et al. (2017) provided the poorest accuracy, especially for the Milton and Pima County sites. Besides being the simplest model (i.e., only three explanatory variables), the model's constant adds a significant value to the "baseline" estimate. For example, in sites with no residential addresses, commercial places, and universities, the model estimates an AADPT of 1,152, which is considerably greater than the average AADPT observed in Milton (313) and Pima County (28). None of the other models, which are discussed below, presented systematic problems.

The applications for Milton are the ones that provided the best performance. The ratios of MAE and average true AADPT ranged from 0.52 to 0.78, and are similar in magnitude to those reported for the original calibration/validation datasets of the DD models (Munira et al., 2020; Schneider et al., 2021). We believe that there are two reasons for this good level of spatial transferability. The first is that the average pedestrian volume observed in Milton is similar to those in the jurisdictions where the models were originally calibrated. The second is that the central tendencies (e.g., average and median) and distribution of the values of the explanatory variables for the Milton sites (Figure 3-2) fits the range of values used to calibrate the models. These points indicate that the sites in Milton are more consistent in terms of site characteristics and pedestrian volumes with the model calibration datasets than the sites in Pima County or in Toronto.

Figure 3-3 displays scatterplots of the observed and estimated AADPTs for analysis of the two models that demonstrated the highest accuracy. It is observed that both models underestimated the pedestrian volumes for the five sites with the greatest AADPTs (the same also happened for the other three DD models). An assessment of each site showed that all of them are associated with the presence of schools nearby. It should be noted that both Griswold et al. (2019) and Singleton et al. (2021) include the count of schools in their equations. This may suggest that the presence of a school in Milton generates more

pedestrian trips than in California and Utah (jurisdictions where the DD models were developed).

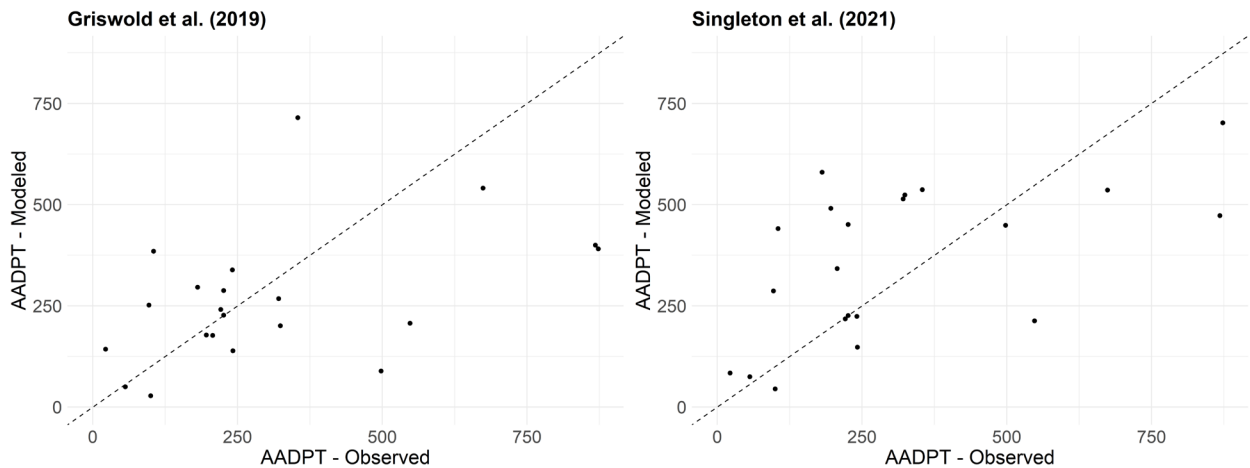


Figure 3-3 Estimated and observed AADPT for sites in Milton for the two most accurate DD models

Regarding the application for Pima County, it is observed that the average AADPT estimates are two to three times lower than for Milton. This was expected due to the reduced potential of pedestrian trip generation associated with the land use and socioeconomic attributes in the Pima County sites, as shown in Figure 3-2 (e.g., low population and employment densities and low percentages of people that walk to work). Nevertheless, systematic overestimation is still found for the estimations of all five DD models. The combination of these facts suggests that the pedestrian activity generated by a given attribute is lower in the Pima County sites than in the sites of the original models. For example, the presence of a commercial establishment or a school might produce fewer pedestrian trips in Pima County than in the sites from the original calibration. The Schneider et al. (2021) model performed the best in Pima County and is also the one with the lowest average AADPT. Different relationships between explanatory variables and the number of produced trips (i.e.,

model coefficients) across jurisdictions may be related to jurisdictional specific characteristics such as level of motorization, incentives to use active transportation, land use mix, weather, and others. In summary, none of the DD models was able to adequately account for the extremely low pedestrian activity in the Pima County sites (average AADPT = 28), probably because of the different contexts in which the models were calibrated.

Concerning Toronto, the performance of the DD models developed by Munira et al. (2020), Griswold et al. (2019), and Schneider et al. (2021) was severely impacted by the extremely large employment and population densities and the large number of people working at home observed in Downtown Toronto. The limits of these variables in the Downtown Toronto sites are significantly outside the range of the variables considered in the calibration of the models (Figure 3-2). For example, the maximum value of employment density observed in Schneider's model is 111,269 employees per square mile, whereas the average and maximum values in the sites of Toronto are 255,842 and 788,478 employees per square mile, respectively. The facts that the models were not calibrated for this range of variables and that independent variables are exponentiated resulted in some highly unrealistic AADPT estimates, especially for the Schneider et al. (2021) model.

Surprisingly, despite considering the population density as one of its explanatory variables, the Hankey et al. (2017) model systematically underestimated the AADPT predictions. In their study, no information regarding the range of the population density used for the model calibration is provided.

The Singleton et al. (2021) model provided the best accuracy for Toronto. The model includes population and employment densities in its formulation, and though the distribution of both variables for the Downtown Toronto sites differs considerably from that observed in the original dataset (Utah, US) (see Figure 3-2), the authors ln-transformed both variables before the model calibration, limiting the influence that very large values of these explanatory variables have on the AADPT estimate. Figure 3-4 shows the Singleton's model error as a function of the observed AADPT. It is noted that the model provided reasonable estimates up to values of AADPT close to 20,000. After that, consistent underestimation is

seen. This issue may be related to de-transformation bias (Wang & Currans, 2018), which can occur when converting predictions from the log scale back to the original scale, as the Singleton et al. model uses a log-linear formulation. Overall, it suggests a tendency of such models to underpredict pedestrian volumes at high-volume locations.

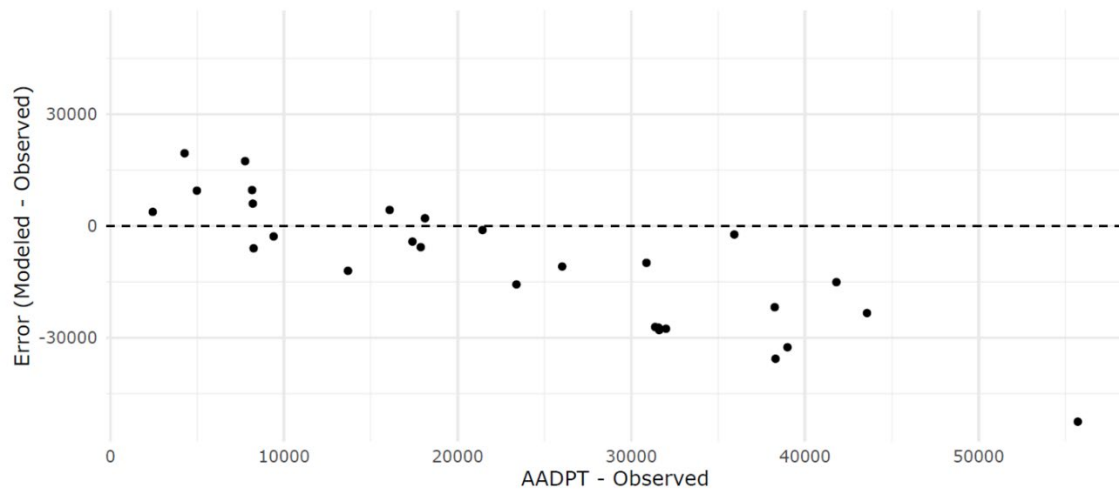


Figure 3-4 Toronto: Singleton et al. model error as a function of observed AADPT

To conclude the assessment of the DD models, Table 3-4 shows the rank of the models based on the ratio between MAE and true average AADPT. Overall, the models developed by Singleton et al. (2021) and Sanders et al. (2017) presented the best and worst transferability performance, respectively. The model proposed by Schneider et al. (2021) provided reasonable accuracy for Milton, the best performance for Pima County, but showed unrealistic estimates for Toronto. The model calibrated by Hankey et al. (2017) consistently overestimated for Pima County and underestimated for Toronto. The model proposed by Griswold et al. (2019) provided the best accuracy for Milton, while the model developed by Munira et al. (2020) showed consistent performance across all three jurisdictions.

The key takeaway from this Chapter is that the DD models evaluated in this research performed considerably differently across jurisdictions with distinct land use and socioeconomic features and levels of pedestrian activity. Five of the models performed reasonably well for Milton, a jurisdiction that has similar characteristics to the jurisdictions on which the original DD models were calibrated. However, these models performed poorly when applied to sites in Toronto and in Pima County. Furthermore, the relative performance of the models was not consistent across the different jurisdictions. It is clear that naively applying the models to jurisdictions that are substantially different from the jurisdictions where the DD models were calibrated can result in very large estimation errors and is not recommended. Establishing if the jurisdiction of interest is “substantially” different from the calibration jurisdiction can be done by examining the distribution of the explanatory variables from the jurisdiction of interest and comparing this to the range of values from the calibration data set. However, this can only be done if the model developers have reported the characteristics of the calibration dataset. Furthermore, establishing that the jurisdiction of interest is not substantially different from the original calibration data set does not guarantee that the application of that model will provide pedestrian volume estimates of acceptable accuracy.

Table 3-4 Summary of DD model performance

Study	E = MAE/AADPT			Average	Rank
	Milton	Pima County	Toronto		
Griswold et al.	0.52	2.46	3.44	2.1	3
Hankey et al.	0.78	5.18	0.77	2.2	4
Munira et al.	0.78	2.04	1.73	1.5	2
Sanders et al.	7.52	45.25	3.70	18.8	6
Schneider et al.	0.65	1.04	50.00 ¹	17.2	5
Singleton et al.	0.53	3.11	0.66	1.4	1
Best	0.52	1.04	0.66	1.4	-
Worst	7.52	45.25	50.00	18.8	-

¹An upper limit of 50 was introduced

After reviewing studies reported in the literature that developed DD models and making significant efforts to apply these models, some suggestions are made to researchers on how to make their methods and models more easily transferrable:

1. Clearly describe the explanatory variables used in the model. For example, defining if multi-unit residential buildings are counted as one or as multiple addresses and stating the criteria for classifying an establishment as a commercial one.
2. Provide a descriptive summary of pedestrian volume and explanatory variables, so other researchers can have a good understanding of the jurisdiction where the model was developed.
3. Include conventional indicators to assess the model's performance, such as MAE and MAPE. As stated in previous sections, it was hard to position the studies in terms of accuracy, since not always those kinds of indicators were available.
4. Place greater emphasis on using explanatory variables that are widely available across jurisdictions to enhance the opportunity for application of these models in different jurisdictions.

For future work, developing DD models using local data may help the assessment of spatially transferred models in terms of model coefficients and accuracy. There is also a need for further research to examine ways to improve the spatial transferability of existing DD models, including methods that enable locally calibrating the AADPT estimates using pedestrian volume data that is available for sites in the target jurisdiction. Another point that has not been explored yet is to what extent the compilation of different municipalities into a single model affects the quality of the modeling. For example, some studies used data from different municipalities within a county or state (Griswold et al., 2019; Singleton et al., 2021). If the relationship between explanatory variables and pedestrian volume (i.e., model coefficients) is different across municipalities, increasing the dataset sample size by combining distinct municipalities may not be beneficial to the model.

Chapter 4

Enhancing the Spatial Transferability of Direct Demand Models for Estimating Pedestrian Volumes at Intersections

This Chapter is based on the following journal paper:

Sobreira, LTP; Hellinga, B. Enhancing the Spatial Transferability of Direct Demand Models for Estimating Pedestrian Volumes at Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2678, issue 7, 2023.
<https://doi.org/10.1177/03611981231205881>

4.1 Overall context

Chapter 3 explored the naïve transferability of DD models, assuming that no sites with pedestrian volume data are available in the study jurisdiction—corresponding to Scenario 3-D in Figure 1-1. This represents an extreme case, as most jurisdictions likely have at least some pedestrian counts for their signalized intersections, even if they lack sufficient data to develop their own DD model (Scenarios 2-C and 3-C in Figure 1-1). Given this, jurisdictions with limited pedestrian count data may have the opportunity to enhance the naïve transferability of DD models by incorporating their available data – a process referred to as local calibration. This is the focus of Chapter 4.

The three DD models that presented the best performance when being naively transferred were selected for the local calibration task: Singleton et al. (2021), Schneider et al. (2021), and Griswold et al. (2019). This Chapter considered the same jurisdictions and databases used in the Chapter 3, as described in Section 3.3.1.

The remainder of this Chapter is organized as follows: the next three sections present the methods proposed for local calibration. This is followed by a discussion of the results, and the Chapter concludes with a summary section.

4.2 Proposed local calibration models

The idea behind the local calibration is that the pedestrian volume data that is readily available to target jurisdictions may be used to enhance the naïve AADPT estimates from DD models that were developed in other jurisdictions. Figure 4-1 presents the proposed method for estimating AADPT using the local calibration approach. Some key terms of the method are described below:

Y'_1 = predicted AADPT for the target jurisdiction (referred to as jurisdiction #2) using a DD model developed in another jurisdiction (referred to as jurisdiction #1). This represents the naïve transferability.

Y_2 = observed AADPT for the target jurisdiction.

Y'_2 = predicted AADPT for the target jurisdiction using the locally calibrated DD model.

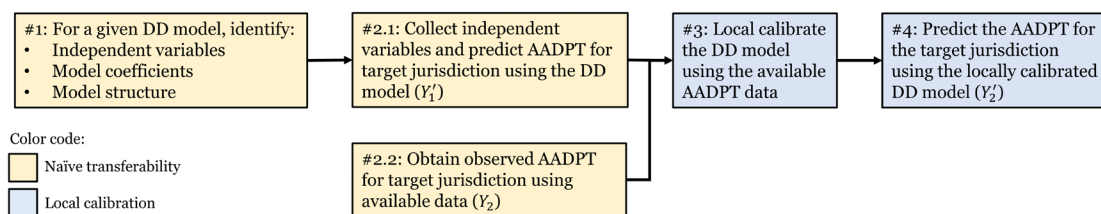


Figure 4-1 Method for estimating AADPT using local calibration

Steps #1, #2.1, and #2.2 represent the naïve transferability approach and are pre-requisites for the local calibration. More details on these steps, on the features of the DD

models selected from the literature, and on how the observed AADPTs were estimated are available in Chapter 3.

Five models (A – E) are proposed for the development of the local calibration (Equations 4.2 to 4.6). In addition, a base model representing the naïve transferability is also considered for comparison purposes (Equation 4.1). Models A and B correspond to scaling factor models. Models C and D consist of a calibration of linear and power regression, respectively, using Y_2 as the dependent variable and Y'_1 as the independent variable. In Model E, a coefficient is calibrated for each pair of independent variables and model coefficient in the DD model. It is expected that Models A-D can capture “macro” aspects of the differences in pedestrian activity between the two jurisdictions (e.g., different levels of motorization or transit supply), whereas Model E can incorporate variations in pedestrian activity at an “independent variable” level. For example, a school in jurisdiction #1 may produce N times more pedestrian trips than a school in jurisdiction #2.

The selected DD models for this task were developed using either log-linear or Negative Binomial distributions (Griswold et al., 2019; Schneider et al., 2021; Singleton et al., 2021). In both distributions, the dependent variable is ln-transformed before the model calibration. We tried the application of the local calibration models using both the natural logarithm of AADPT and AADPT itself. The former approach presented better results and was the one adopted. The exponential of the natural logarithm of AADPT was then taken to obtain the locally calibrated AADPT.

Base model: no change to the original model from jurisdiction #1 (naïve transferability)

$$\ln(Y'_{base,i,2}) = \ln(Y'_{i,1}) = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_j X_{i,j} \quad (4.1)$$

Model A: calibration factor (α)

$$\alpha = \frac{1}{k} \sum_{i=1}^k \frac{\ln(Y_{i,2})}{\ln(Y'_{i,1})} \therefore \ln(Y'_{A,i,2}) = \alpha \ln(Y'_{i,1}) \quad (4.2)$$

Model B: calibration factor version 2 (α)

$$\alpha = \frac{\sum_{i=1}^k \ln(Y_{i,2})}{\sum_{i=1}^k \ln(Y'_{i,1})} \therefore \ln(Y'_{B,i,2}) = \alpha \ln(Y'_{i,1}) \quad (4.3)$$

Model C: linear regression to obtain slope (α) and intercept (γ)

$$\ln(Y'_{C,i,2}) = \alpha \ln(Y'_{i,1}) + \gamma \quad (4.4)$$

Model D: power regression to obtain coefficients α and γ

$$\ln(Y'_{D,i,2}) = \alpha \ln(Y'_{i,1})^\gamma \quad (4.5)$$

Model E: linear regression to obtain adjustment (α) to each independent variable and coefficient

$$\ln(Y'_{E,i,2}) = \alpha_0\beta_0 + \alpha_1\beta_1X_1 + \dots + \alpha_j\beta_jX_j \quad (4.6)$$

Where:

i = a given signalized intersection for which AADPT is known.

j = number of explanatory variables in the DD model.

β = model coefficients from the literature DD model.

X = explanatory variables extracted from jurisdiction #2.

k = number of sites used where AADPT is known.

α and γ = local calibration coefficients.

$Y'_{M,i,2}$ = locally calibrated prediction of AADPT for site i in the target jurisdiction using local calibration model M , where $M = \{A, B, C, D, E\}$

4.3 Evaluation methodology

The performance of the different proposed local calibration models will depend on a) the number of sites (k) in the target jurisdiction for which AADPT are available, b) the characteristics of these sites, c) the DD model adopted, and d) the characteristics of the target jurisdiction. The first two factors can be considered random variables. As a result, a Monte-Carlo simulation method is proposed to assess the performance of each of the five local calibration models across a range of sites. The steps are described below. Steps 3-6 are simulated 300 times for each model and each value of k . The reason why we performed the simulation 300 times is discussed in the next section.

1. Identify sites where AADPT is known. Sample size is equal to n .
2. Set sample sizes of sites used for local calibration ($k = 5, 10, 15, \dots, n$).
3. Randomly select k sites from the entire sample to perform the local calibration.
4. Calibrate the coefficients α and γ .
5. Obtain Y'_2 for the entire sample.
6. Compute the error metric (E) as the ratio between MAE and average observed AADPT (Equation 2.11).

The median of the error indicators (E) obtained in Step 6 is calculated for each local calibration model and each value of k and then compared to the result achieved with the Base Model (i.e., naïve transferability). The magnitude of this indicator obtained in the development of the original models is typically between 0.60 and 0.65 (Table 3-3). This value is used as a benchmark for determining the quality of the local calibration. We note that the error metric carries errors from the original DD model development and from the spatial transferability itself. We are not able to separate these errors.

As the reader may have noticed, small samples of sites where AADPT is known are available in the jurisdictions of the study (Table 3-3: Milton = 21 sites, Pima County = 81, and Toronto = 28). This constrained our ability to carry out in-depth analyses, for example:

defining the optimal number or percentage of sites that maximize the cost-benefit of the local calibration. Thus, we complimented our evaluation using the empirical datasets with an evaluation based on synthesized AADPT data for real sites in the same three jurisdictions. The method of generating the synthetic AADPT data is described in the next section.

4.4 Generating synthetic AADPT data

The objective of synthesizing AADPT data is to create a larger dataset on which to evaluate the five local calibration models. The main challenge in this task is to generate synthetic AADPTs that are realistic given the characteristics of the site and the jurisdiction. Considering that, we proposed the use of DD models to generate synthetic estimates. The rationale behind this decision is that the DD models have been calibrated to capture the relationships between site characteristics and AADPT and were calibrated in different jurisdictions, so each DD model represents a particular environment in terms of the relationship between pedestrian activity and socioeconomic and land use features. These site-specific impacts and jurisdiction impacts are captured into the DD model via the explanatory variables and associated coefficients. As long as the explanatory variables of a given DD model are available for the sites, synthetic AADPTs can be generated. An example is provided to illustrate the method: Griswold et al. (2019) developed their DD model using sites from the State of California, USA. We apply this DD model to all sites in a target jurisdiction (for example Milton) to create synthetic AADPT to represent Y_2 . Then, it is possible to assess the spatial transferability and local calibration of a different DD model.

The generation of the synthetic AADPTs starts with the naïve estimate using a given DD model (Equation 4.7).

$$Y_{i,2}^* = e^{\beta_0 + \beta_1 X_{i,1} + \dots + \beta_n X_{i,n}} \quad (4.7)$$

Where:

$Y_{i,2}^*$ = initial synthetic AADPT estimate.

We decided to set an upper limit threshold ($Y_{max,j}$, where j refers to the jurisdiction) for $Y_{i,2}^*$ for two reasons: a) avoid unrealistic AADPT estimates and b) avoid datasets with a very wide range of AADPT values because sites with large AADPT estimates dominate the performance evaluation (i.e., the $E = MAE/\bar{Y}_2$ indicator would be biased towards the performance of these sites). Thus, we investigated the $Y_{i,2}^*$ estimates obtained for all sites in each jurisdiction and selected a value for $Y_{max,j}$ for each jurisdiction j by considering the maximum AADPT value observed in the “real” dataset and the number of sites that would be available for the analysis (i.e., if $Y_{i,2}^* > Y_{max,j}$, then site i was eliminated from the dataset). For example, in Pima County, $Y_{max,Pima}$ was set at 489. This number represents the maximum “real” AADPT value (163) multiplied by three and resulted in the removal of 160 out of 744 sites (signalized intersections).

After filtering the datasets based on the $Y_{max,j}$ threshold, an error term was included in the synthetic AADPT estimates to ensure realistic variability across sites is considered (Equation 4.8) and so that different populations of synthesized AADPTs can be generated for the sites in each jurisdiction from each DD model. The error term follows a normal distribution with a mean equal to zero and a given variance (Equation 4.9). A variance (represented by the standard deviation) was obtained for each site as a function of a given coefficient of variation (COV) and the $Y_{i,2}^*$ estimate (Equation 4.10). A single COV was calibrated for the application of each jurisdiction and DD model with the goal of generating synthetic datasets with an error metric ($E = MAE/\bar{Y}_2$) indicator of approximately 0.60 – 0.65 when comparing the AADPT estimates with and without the error term. In other words, we aimed to create synthetic datasets with similar AADPT variability (i.e., errors between observed and predicted AADPTs) than the one achieved in the original DD model calibration.

$$Y_{i,2} = Y_{i,2}^* + \varepsilon_i \quad (4.8)$$

$$\varepsilon_i \sim \text{Normal}(0, s_i^2) \quad (4.9)$$

$$s_i = COV \times Y_{i,2}^* \quad (4.10)$$

Where:

ε_i = random error assigned to site i . This term was constrained by $\pm 2s_i$.

s_i^2 = variance of ε_i .

s_i = standard deviation of ε_i .

COV = coefficient of variation calibrated for the application of each jurisdiction and DD model.

The inclusion of the random error resulted in the estimation of sites with zero or negative AADPTs, especially for sites with small values of $Y_{i,2}^*$ (i.e., the estimate obtained from the naïve transferability of the DD model). Although negative AADPTs could be treated as zero, we would get 10-15% of sites with zero AADPT. One would expect several sites with very low pedestrian activity, but this quantity of null activity is unrealistic. To prevent an over-representation of AADPT = 0, a secondary error term was added to $Y_{i,2}$ when its result is non-positive (Equation 4.11).

$$\varepsilon_i^* = \text{random between} \left(1, \max \left\{ 0.25 \times Y_{i,2}^*, 50 \right\} \right) \quad (4.11)$$

Where:

ε_i^* = secondary error term added to $Y_{i,2}$ when it is non-positive.

The functional form and parameters of Equation 4.11 are subjective but are based on examination of the observed AADPT data from the three jurisdictions used in this study and on the requirements to retain stochasticity in the synthesized data while limiting the number of sites with synthesized AADPT values equal to zero.

Due to the addition of the random error terms, there are an unlimited number of synthetic AADPTs for a given jurisdiction using a given DD model. As a result, ten sets of synthetic AADPTs were created for each jurisdiction and DD model combination, and 30 repetitions for each set were used in the simulation process outlined in the previous section. The 300 simulations used in the real AADPT approach were set to match the number of simulations in the synthetic approach. The performance of the five local calibration models considering both “real” and synthetic AADPTs are presented in the next section.

4.5 Results and discussion

Table 4-1 provides an overview of the number of sites and descriptive statistics of the site AADPT values as obtained from the real AADPT and synthetic AADPT approaches. For the synthetic approach, we decided to split the City of Toronto sites into two datasets – Downtown and Other – because we observed that the AADPT estimates in Downtown were much larger than the ones in the rest of the city. This division was made to avoid biasing the results of the local calibration towards the sites in Downtown. Synthetic datasets were generated for each jurisdiction using the following DD models: Singleton et al. (2021), Schneider et al. (2021), and Griswold et al. (2019), except for the application of Schneider et al. (2021) in Downtown Toronto which was not possible due to the generation of unrealistically large estimates (the Schneider et al. (2021) provides highly unrealistic AADPT estimates when employment density becomes large as is the case for the Downtown Toronto sites). The evaluation using the real AADPT data used the datasets from Milton, Pima County, and Downtown Toronto.

Table 4-1 Descriptive summary: sample size and AADPT

Jurisdiction	Y_2 obtained from	Number of sites	Statistics – AADPT ¹			
			Average	Median	Min	Max
Milton	Real AADPT	21	313	226	22	873
	Griswold et al.	94	180	153	0	922
	Schneider et al.	94	106	74	11	824
	Singleton et al.	94	120	73	2	500
Pima County	Real AADPT ²	79	28	16	2	163
	Griswold et al.	584	119	90	0	469
	Schneider et al.	584	89	51	12	478
	Singleton et al.	584	152	124	1	489
Toronto (Downtown)	Real AADPT	28	23,481	22,409	2,443	55,707
	Griswold et al.	197	32,174	18,200	368	107,873
	Singleton et al.	197	11,805	7,303	288	73,986
Toronto (Other)	Griswold et al.	1,275	317	217	0	3,979
	Schneider et al.	1,275	1,181	618	22	5,960
	Singleton et al.	1,275	1,109	664	9	5,871

¹The statistics for the synthetic approach were calculated before the inclusion of the error term ($Y_{i,2}^*$)

²Two sites were removed because they presented very small AADPT estimates using Griswold et al. model, which was biasing the local calibration process

From Table 4-1 it can be observed that the synthetic data show similar, but not exactly the same, tendencies across jurisdictions as do the real AADPT data. For example, the real data shows that the Downtown Toronto sites have much larger AAPDT than the other jurisdictions. This difference is also reflected in the synthesized data. However, the real data also exhibits a larger mean and median for Milton than Pima County, but for the synthesized data, the mean and median values are much more similar. It is also worth noting that some of the synthetic datasets presented AADPTs with similar central tendencies, for example the models of Singleton et al. (2021) and Schneider et al. (2021) provide very similar synthesized data when applied to Milton and when applied to Toronto-Other. This had an impact on the performance of the local calibration models, and it is discussed in the next section.

4.5.1 Local calibration using synthetic AADPTs

Figure 4-2 presents a sample of the results obtained with the local calibration using the synthetic datasets. Each graph represents the local calibration of a given DD model to a given target jurisdiction in which synthetic AADPTs were generated using a given DD model. In total, 31 graphs were created (For Milton, Pima County, and Toronto-Other, 9 graphs were generated for each jurisdiction. For Toronto-Downtown, Schneider et al. (2021) model could not be applied, so there were only four combinations for this jurisdiction). Note that we also considered the case where a DD model is transferred to a jurisdiction in which the synthetic AADPTs were generated using the same DD model (Figure 4-2F). Our intention was to evaluate the local calibration in scenarios where the target jurisdiction has an identical relationship between pedestrian activity and socioeconomic/land use features to the one observed in the jurisdiction where the model was originally developed.

Each point in the graphs in Figure 4-2 represents the median of the error indicator $E = MAE/\bar{Y}_2$ for the 300 observations considered in the simulation procedure. The x-axis presents the percentage of sites used for the local calibration, ranging from 2.5% to 100% (with the associated number of sites, k , provided in brackets). Each color represents one local calibration model (Equations 4.1 to 4.6). As a reminder, Models A and B correspond to scaling factor models; Models C and D consist of a calibration of linear and power regression, respectively. In Model E, a coefficient is calibrated for each pair of independent variables and model coefficient in the DD model. The calibration of Model E requires the number of observations to be greater than the number of explanatory variables in the DD model which is why Model E is not visible when the number of sites used for local calibration is small. The y-axis scale of each graph is selected to be appropriate for the associated jurisdiction. The performance of the Base Model (i.e., naïve transferability) does not vary with the number of sites available for local calibration and is therefore a horizontal line.

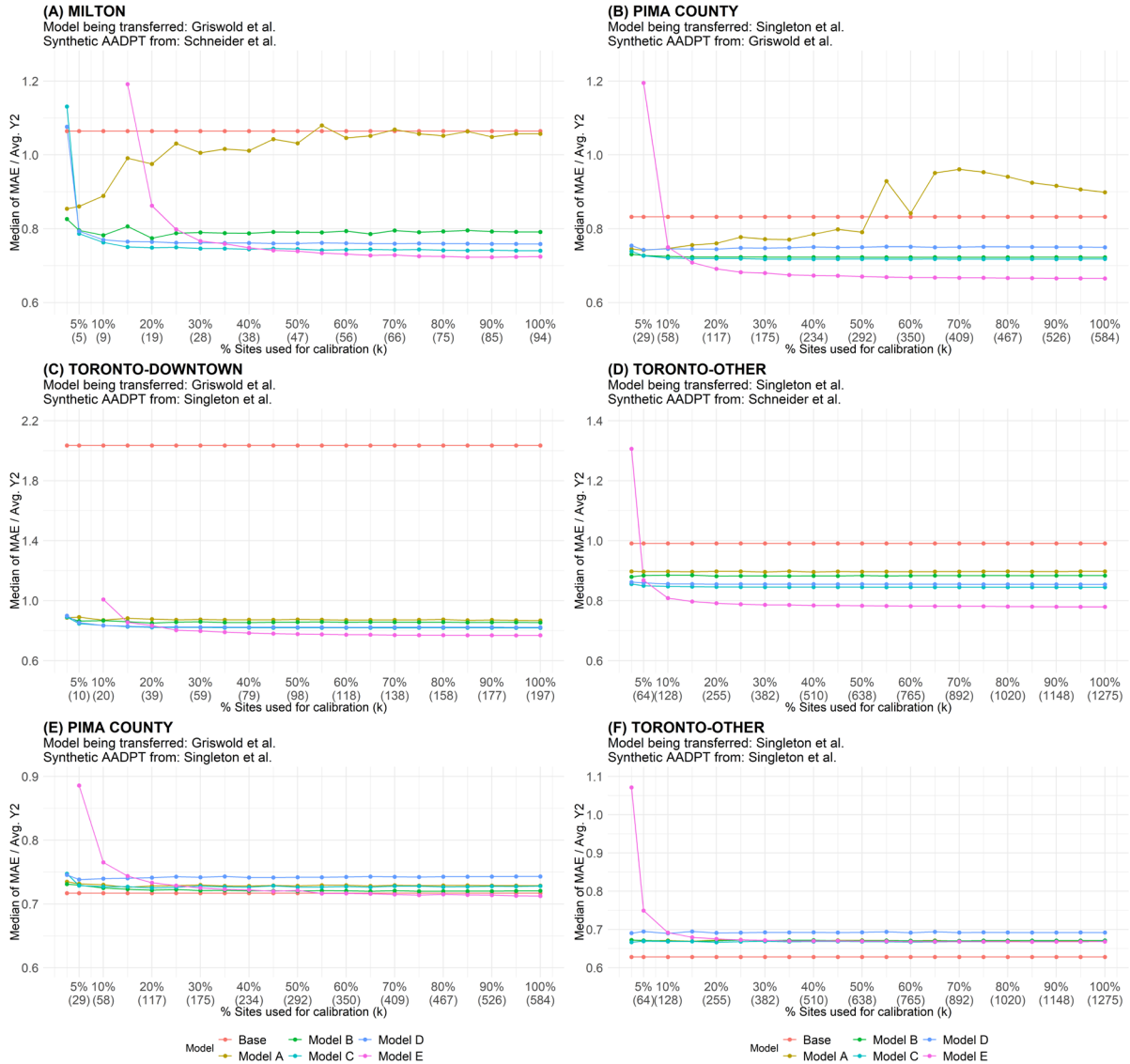


Figure 4-2 Results: local calibration using synthetic AADPTs

The performance of the local calibration models varies significantly across graphs in Figure 4-2. Figure 4-2A-D presents situations in which one or more local calibration models produce improvements when compared to the Base Model (i.e., naïve transferability). However, Figure 4-2E and F show cases where the effect of local calibration is null or negative. The main reason for the variability in the performance is associated with the level

of similarity of the relationship between pedestrian activity and socioeconomic/land use features in the target jurisdiction and in the one where the DD model was originally developed. One would expect the local calibration to provide improvements when the two jurisdictions exhibit different pedestrian activity relationships, especially if systematic underestimation or overestimation in the AADPT obtained from naïve transferability is observed. Local calibration Models A-D (Equations 4.2 to 4.5) are formulated to address scaling differences between pedestrian activity in the two jurisdictions. Conversely, it is expected that the local calibration is less beneficial when the two jurisdictions are very similar.

Model E (Equation 4.6) is expected to be beneficial when differences in the two jurisdictions are not predominantly a global scaling but reflect different levels of pedestrian activity for specific explanatory variables (e.g., a retail place generates X times more trips in the target jurisdiction than in the jurisdiction where the model was developed, but schools generate Y times more trips and X and Y are not similar). However, a review of all 31 graphs produced revealed that it is necessary for observed AADPT data to be available from at least 70% of the total sites before Model E provides results that are better than the Base Model. From a practical perspective, this suggests that local calibration Model E is only preferred under conditions when it would be both possible and desirable to develop a DD model directly for the target jurisdiction using the available data.

Figure 4-3 illustrates the effect of similarity of jurisdictions on local calibration model performance. Figure 4-3A shows a scatterplot between the naïve AADPT estimates using Griswold et al. (2019) and the synthetic AADPT estimates using Singleton et al. (2021) for Milton. This represents one of the ten datasets used to obtain Figure 4-2A, where the local calibration performed well. One may notice that overestimation prevails in Figure 4-3A, which is the reason why the $E = MAE/\bar{Y}_2$ indicator of 1.06 from the Base Model quickly improves to a magnitude of 0.80 when local calibration Models B, C or D are used. Figure 4-3B shows one of the 10 data sets used to produce Figure 4-2E, a scenario where the local calibration showed a null or negative effect. Although notable differences between the

estimates are observed, no systematic under/overestimation occurs, which is why the local calibration models are not able to provide any improvement over the Base Model.

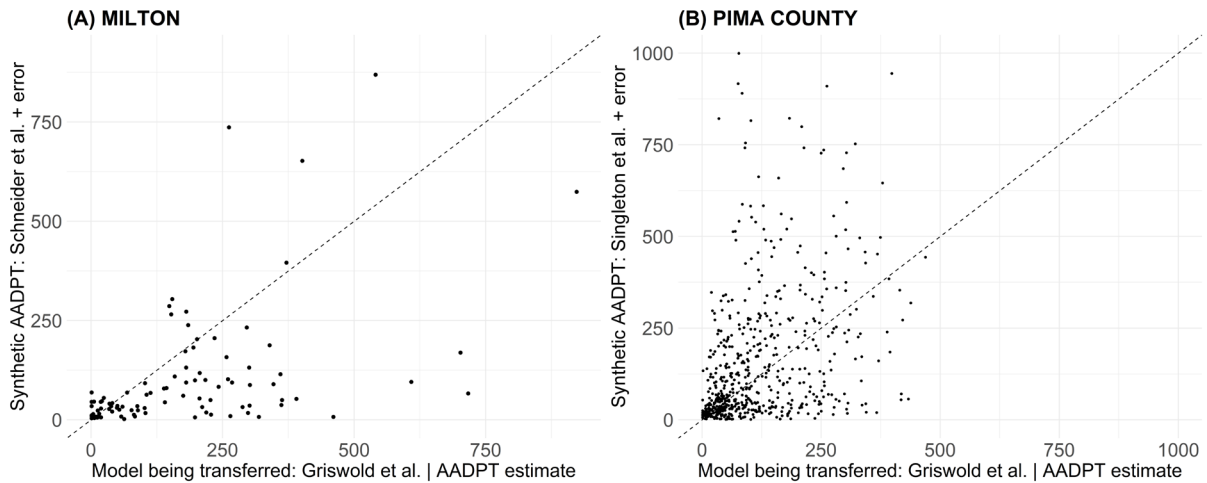


Figure 4-3 Scatterplots: estimated AADPT estimate vs synthetic AADPT

Local calibration Model A presents a different pattern (i.e., error increases as more sites are available for local calibration) than the other local calibration models in Figure 4-2A-B and this difference was also observed in some of the graphs that were not displayed in this document. This pattern occurred because of how Model A was formulated (Equation 4.2). Take the example where the ratio between observed AADPT (Y_2) and predicted AADPT (Y_1') for two sites is 2 and 0.5. Ideally, one would set the local calibration coefficient (α) to be equal to one. However, a value of 1.25 is obtained when using Equation 4.2. Therefore, the sites where Y_2 are greater than Y_1' dominate how Model A is calibrated, resulting in poor performances when no under/overestimation is systematically observed.

Consistent patterns for the performance of Models B-E across the graphs are observed in Figure 4-2. When small percentages of the total sites are available for local calibration (e.g., 2.5% and 5%), Models B-D outperform Model E. Model E turns out to be the best option when a relatively large sample of sites is available for the local calibration (between

50 and 100 sites, depending on the jurisdiction). However, if this amount of information is available, it would probably be preferable if the jurisdiction developed its own DD model instead of transferring and locally calibrating a DD model developed in another jurisdiction. Further research is required to confirm this statement.

Table 4-2 presents a summary of all 31 synthetic applications, considering the use of 10% of the sites for the local calibration. Each cell of Table 4-2 represents the median of $E = MAE/\bar{Y}_2$. The shading represents the comparison of the performance within each application. When the same DD model is used for both synthetic AADPT generation and local calibration, the error metric (E) of the Base Model is between 0.60 and 0.65, as imposed by the procedure of generating the synthetic estimates. Three main findings are made based on the results in Table 4-2:

1. Although similar performance is observed between Models B-D, Models B and C performed slightly better than the others.
2. The “cost” of applying the local calibration to a jurisdiction that is similar to the one where the DD model was originally developed is low. It is exemplified by the 11 cases where the same DD model is used for both synthetic generation and local calibration. For these 11 cases, the use of local calibration Models B and C results in errors that are, on average, 5.7% (95% CI: 4.4% to 7.0%) larger than using naïve transferability (Base Model). Note that in these cases, the Base Model represents the best estimate that one could get, since the only difference between observed and predicted AADPTs is the random error term that was added when generating the synthetic AADPTs. Therefore, the “cost” may be overestimated for practical applications because it is unlikely that two different jurisdictions present the exact same relationship between pedestrian activity and socioeconomic/land-use features.

Table 4-2 Summary of the synthetic AADPT approach using 10% of sites for local calibration

Jurisdiction	Y_2 obtained from	Sites used for local calibration		Model being transferred and locally calibrated (median of $E = MAE/\bar{Y}_2$)																	
				Griswold et al.						Schneider et al.						Singleton et al.					
		%	k	Base	A	B	C	D	E	Base	A	B	C	D	E	Base	A	B	C	D	E
Milton (n = 94)	Griswold et al.	10	9	0.63	0.70	0.67	0.70	0.72	-	0.76	0.81	0.81	0.79	0.82	-	0.75	1.28	0.88	0.75	0.77	-
	Schneider et al.	10	9	1.06	0.89	0.78	0.76	0.77	-	0.61	0.63	0.62	0.65	0.68	-	0.77	1.64	0.80	0.73	0.75	-
	Singleton et al.	10	9	0.95	0.79	0.77	0.78	0.79	-	0.73	0.73	0.73	0.77	0.79	-	0.62	0.70	0.65	0.67	0.71	-
Pima County (n = 584)	Griswold et al.	10	58	0.61	0.61	0.62	0.65	0.68	0.68	0.73	0.75	0.75	0.71	0.71	1.26	0.83	0.75	0.73	0.72	0.75	0.75
	Schneider et al.	10	58	0.85	0.73	0.72	0.73	0.74	0.76	0.60	0.60	0.60	0.62	0.63	0.70	1.04	0.75	0.73	0.73	0.75	0.76
	Singleton et al.	10	58	0.72	0.73	0.73	0.73	0.74	0.77	0.74	0.77	0.76	0.72	0.72	0.83	0.62	0.63	0.63	0.64	0.69	0.73
Toronto Other (n = 1,275)	Griswold et al.	10	128	0.62	0.66	0.65	0.67	0.68	0.68	2.60	0.77	0.77	0.74	0.75	0.76	2.56	0.85	0.84	0.78	0.80	0.74
	Schneider et al.	10	128	0.84	0.78	0.78	0.79	0.79	0.79	0.63	0.66	0.66	0.66	0.67	0.68	0.99	0.90	0.89	0.85	0.86	0.81
	Singleton et al.	10	128	0.86	0.86	0.85	0.82	0.83	0.83	1.03	0.92	0.90	0.82	0.82	0.83	0.63	0.67	0.67	0.67	0.69	0.69
Toronto Downtown (n = 197)	Griswold et al.	10	20	0.64	0.66	0.66	0.66	0.67	0.88	-	-	-	-	-	-	0.86	0.88	0.88	0.86	0.87	5.99
	Schneider et al.	10	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Singleton et al.	10	20	2.04	0.87	0.87	0.83	0.83	1.01	-	-	-	-	-	-	0.63	0.68	0.68	0.69	0.70	6.85
Average across jurisdictions				0.89	0.75	0.74	0.74	0.75	0.80	0.94	0.74	0.73	0.72	0.73	0.84	0.94	0.88	0.76	0.74	0.76	2.17
Best model				Model B - 0.74						Model C - 0.72						Model C - 0.74					

Note: the shading represents the comparison within each application. The shading scale is green = lowest value, gray = mid value, and red = highest value

3. The “benefit” of applying the local calibration to a jurisdiction that is not necessarily similar to the one where the DD model was originally developed is about 2.5 times greater than the estimated cost in the previous item. This scenario is represented by the 20 cases in which different DD models were used for generating the synthetic AADPTs and for transferability. For these cases, the use of local calibration models B and C results in errors that are, on average, 14.6% (95% CI: 6.7% to 22.4%) lower than using naïve transferability (Base Model).

Detailing the performance of Models B and C, Figure 4-4 presents the average and the 95% confidence interval around the average of the percent difference in error indicator E when comparing Models B and C to the Base Model. Each element of Figure 4-4 was calculated using information from the 31 different applications (i.e., including cases where jurisdictions are similar and when they are different). Positive values mean an increase in error indicator E with the local calibration, whereas negative values refer to improvements in the transferability (decrease in error). It is observed that Model B outperforms Model C when 2.5% of the sites are used for the local calibration, but we could not statistically confirm the difference in the average performance of the models (p-value for a two-tailed hypothesis test = 0.13). When 5% or more is available, the performance is similar, with Model C showing slightly better results. It is also seen that the performance is steady after reaching 10% of sites. Consequently, collecting AADPT for 10% of the sites may be the ideal number for jurisdictions that are interested in applying this approach. The application of Model C using 10% of sites resulted in an average improvement of 8.7% (95% CI: 0.6% to 16.9%). The smaller improvement values are associated with cases in which the jurisdictions are the same or similar, and large improvement values with cases in which jurisdictions exhibit greater differences.

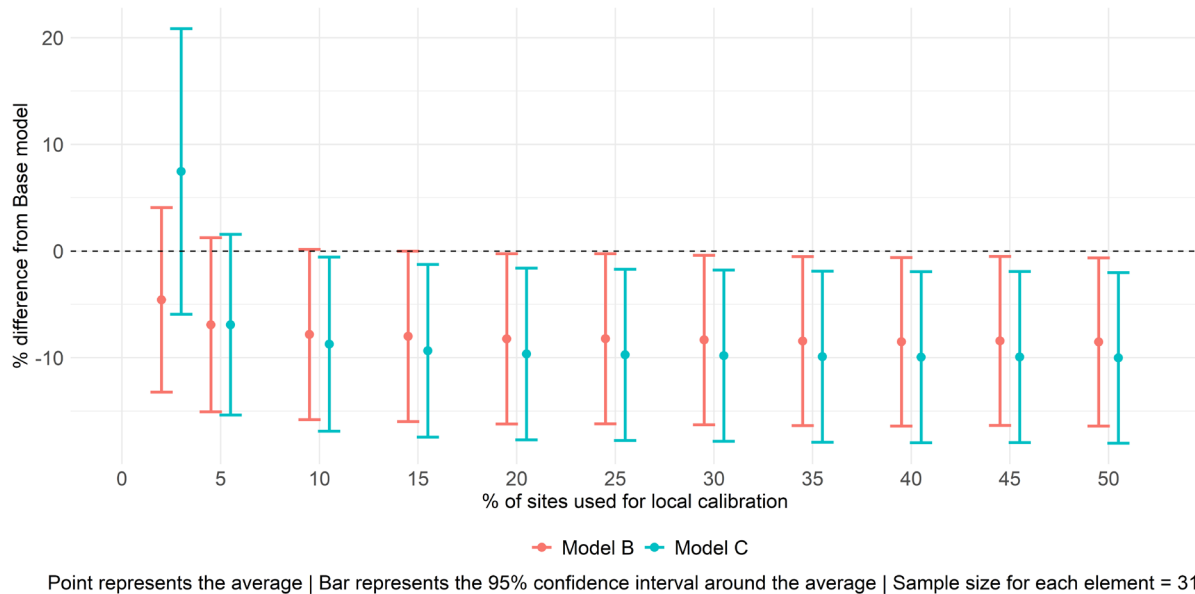


Figure 4-4 Performance of Models B and C when compared to the Base Model (negative values indicate improved accuracy vs Base Model)

In this section, we frequently associated the performance of the local calibration with the level of similarity between the target jurisdiction (where AADPT was synthetically generated using a DD model) and the jurisdiction where the DD model that is being transferred was developed. We tested three DD models that include different explanatory variables collected within different radii. In addition, we only have access to the brief descriptive analysis (e.g., average, standard deviation, minimum, and maximum) presented in the original papers to characterize each jurisdiction. Consequently, our ability to comprehensively understand the synthetic AADPTs that were generated and compare the land use and socioeconomic attributes of the jurisdictions was constrained. This is why we could not perform a thorough analysis in this section and limited ourselves to investigating the differences between synthetic and estimated AADPT (Figure 4-3). The local calibration using real AADPTs is assessed in the next section. This allows us to have a better control over the observed AADPTs, the characteristics of the target jurisdictions (Figure 3-2), and

the performance of the naïve transferability (Table 3-3); therefore, we can explore the effect of the level of similarity between the two jurisdictions in more detail.

4.5.2 Local calibration using real AADPTs

Table 4-3 summarizes the local calibration using real AADPTs. As the number of sites where real AADPT is known is very limited, the local calibration was conducted employing small samples for local calibration (5 or 6 sites). The results indicate that on average Model C performs best, an outcome that is consistent with the results from the synthetic approach.

Before detailing the results, we suggest the reader revisit Figure 3-2 and Table 3-3. It is observed that Milton is similar to the jurisdictions where the DD models were originally calibrated in terms of attributes (Figure 3-2) and pedestrian activity (similar average AADPT for the application in Table 3-3). Pima County presents reduced magnitudes in both attributes and pedestrian activity (average observed AADPT = 28); however, the transferability of the DD models resulted in systematically overestimation of the AADPTs, suggesting that the relationship between attributes and pedestrian activity (i.e., model coefficients) is different in Pima County when compared to the jurisdictions where the DD models were developed. Downtown Toronto is a particular case because overestimation and underestimation were observed when transferring the Griswold et al. (2019) and Singleton et al. (2021) models, respectively.

The naïve transferability (Base Model) of the DD models for Milton performed well, providing accuracy that was compatible with – and sometimes better than – the benchmark measure of $E = MAE/\bar{Y}_2$ between 0.60 and 0.65. As expected, the application of local calibration provided only modest improvements over the Base Model. The performance was slightly improved with the local calibration of Schneider et al. (2021) and Singleton et al. (2021) DD models and marginally worsened with the local calibration of the Griswold et al. (2019) model.

Table 4-3 Summary of the real AADPT approach

Jurisdiction	n	Sites used for local calibration		Model being transferred and locally calibrated (median of $E = MAE/\bar{Y}_2$)																	
		%	k	Griswold et al.						Schneider et al.						Singleton et al.					
				Base	A	B	C	D	E	Base	A	B	C	D	E	Base	A	B	C	D	E
Milton	21	24	5	0.52	0.57	0.55	0.57	0.58	-	0.65	0.98	0.83	0.61	0.62	-	0.53	0.51	0.51	0.52	0.52	-
Pima County	79	6	5	2.50	0.65	0.64	0.67	0.70	-	1.05	0.63	0.63	0.70	0.73	-	3.15	0.67	0.68	0.71	0.72	-
Toronto	28	21	6	3.44	0.75	0.75	0.60	0.61	-	-	-	-	-	-	-	0.66	0.85	0.80	0.57	0.57	-
Average across jurisdictions				2.15	0.66	0.64	0.62	0.63	-	0.85	0.80	0.73	0.66	0.67	-	1.45	0.68	0.66	0.60	0.60	-
Best model				Model C: 0.62						Model C: 0.66						Model C: 0.60					

Note: the shading represents the comparison within each application. The shading scale is green = lowest value, gray = mid value, and red = highest value

The results for the Pima County and Toronto jurisdictions are different. The Base Model (naïve transferability) performed poorly, and systematic overestimation and underestimation are observed for Pima County and Toronto, respectively. Significant improvements in the error metric (E) were observed even with the use of very small sample sizes for local calibration, leading to the reduction of the errors to the benchmark level of accuracy. This indicates that the locally calibrated DD model performs as well as we expect a locally developed DD model to do, demonstrating the potential of the technique for real world applications.

Considering the results of the eight applications in Table 4-3, Model C provided an average improvement of 35.0% (95% CI: 4.5% to 65.4%) when compared to the Base Model (naïve transferability). These results corroborate with the findings using the synthetic AADPT approach, specifically:

1. If the target jurisdiction and the one where the DD model was originally developed are similar in terms of the relationship between pedestrian activity and socioeconomic/land use features, local calibration provides modest improvements over naïve transferability. This situation can be observed in Table 4-3 for the application of Griswold et al. (2019) and Singleton et al. (2021) models to the Milton sites. The fact that local calibration does not provide large benefits for these cases is acceptable because when the jurisdictions are very similar, then naïve transferability can be expected to provide adequate results.
2. When the jurisdictions are different, then local calibration shows great potential for improving performance versus the naïve transferability, even when the number of sites where AADPT is known is very limited. This is evident from Table 4-3 for the application of the local calibration models to the Toronto and Pima County sites.

4.6 Conclusions

This Chapter examined the enhancement of the spatial transferability of DD models available in the literature. The models were locally calibrated in the target jurisdictions by making use of sites where AADPT is readily known. To perform the local calibration, five

models were proposed and tested using datasets with a) synthetic AADPT and b) real AADPT. The results showed great potential for the local calibration to improve the naïve transferability of DD models.

Regarding the synthetic application, 31 datasets were generated. In general, it was observed that the local calibration provides substantial benefits when the target jurisdiction and the one where the DD model was originally calibrated are different in terms of the relationship between pedestrian activity and socioeconomic/land use features. In this case, an average improvement of 14.6% (95% CI: 6.7% to 22.4%) was observed when compared to the naïve transferability. Conversely, when the two jurisdictions are identical, the use of the local calibration makes the prediction slightly worse (average: 5.7%; 95% CI: 4.4% to 7.0%). The benefits are about 2.5 times greater than the “costs”, and one would not expect two jurisdictions to show the exact same pattern between pedestrian activity and socioeconomic/land use features, indicating that the “costs” are probably overestimated.

In terms of the best model to perform the local calibration, Models B and C stood out. When 2.5% of the sites are used for local calibration, Model B outperforms Model C. The performance is comparable when 5% or more is available, with Model C producing slightly better results. The application of Model C using 10% of sites resulted in an average improvement of 8.7% (95% CI: 0.6% to 16.9%).

The results achieved from the real AADPT approach corroborated the findings from the synthetic approach. The naïve transferability for Milton already provided very reasonable results (E of 0.52–0.65), so local calibration provided only modest improvements. Concerning Pima County and Downtown Toronto, the naïve estimates showed systematic overestimation and underestimation, respectively. As a result, local calibration provided substantial improvements, even with the use of very small sample sizes. Considering the three jurisdictions, Model C provided an average improvement of 35.0% (95% CI: 4.5% to 65.4%) when compared to the naïve transferability. The magnitude of errors observed in the locally calibrated DD models is similar to that expected if a locally developed DD model was used, demonstrating the potential of the technique for real world applications.

From a practical point of view, some recommendations are made if one is willing to transfer and locally calibrate a given DD model:

1. Examine the similarities between the target jurisdiction and the one where the DD model was developed by comparing the magnitude of pedestrian activity (e.g., central tendencies of AADPT) and socioeconomic/land use features considered in the DD model. This may give a clue as to the level of similarity between the two jurisdictions.
2. If a very limited sample size of sites where AADPT is known is available (i.e., 2.5% or lower), the use of Model B is recommended. Otherwise, Model C should be used.
3. If the target jurisdiction contains zones with different relationships between pedestrian activity and socioeconomic/land use features, it is prudent to develop the local calibration separately for each zone. The City of Toronto illustrates this case. Downtown Toronto has a unique environment where, although trip generation attributes are also increased, the level of pedestrian activity is extremely high. One would expect the relationship between socioeconomic/land use features and AADPT (i.e., DD model coefficients) to be different in Downtown than in the rest of the city.

Finally, we make three suggestions for future work:

1. This study used a combination of synthesized data and empirical data because a sufficiently large empirical data set (in which pedestrian counts were not impacted by COVID-19) was not available. And though the empirical data analysis provided results that were consistent with those obtained from the analysis of the synthesized data, it would be prudent to confirm the results obtained from the synthetic approach. using a database with larger sample sizes of sites where the real AADPT is known.
2. This work used only sites where CC stations were available, so relatively reliable AADPT estimates were considered. Many jurisdictions have STCs for a large proportion of their intersection, but a relatively small number of continuous count stations. It is recommended to extend the work from this study to examine the potential for performing

local calibration of DD models with STCs expanded to AADPT or even unexpanded STCs.

3. Instead of the local calibration models proposed in this work, employing Bayesian methods to combine local information from the jurisdiction with estimates from DD models may be a viable alternative.
4. It is recommended to investigate the optimal threshold (i.e., number or percentage of sites with AADPT available) separating the use of the local calibration approach from the development of a DD model for the target jurisdiction.

Chapter 5

Estimating Pedestrian Volumes at Intersections using Artificial Intelligence: A ChatGPT Vision Approach

This Chapter is based on the following journal paper:

Sobreira, LTP; Hellinga, B. Estimating Pedestrian Volumes at Intersections using Artificial Intelligence: A ChatGPT Vision Approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2024. <https://doi.org/10.1177/03611981241308873>

5.1 Overall context

Chapters 3 and 4 explored, respectively, the naïve transferability and local calibration of DD models for estimating pedestrian exposure when STCs or CCs are not readily available at given sites and when not enough data is available for developing DD models, representing sites in Scenarios 2-C and 3-C in Figure 1-1. This Chapter proposes and applies a novel alternative method for sites in these scenarios.

A recent research direction for estimating pedestrian exposure involves the employment of Artificial Intelligence (AI). Many researchers have applied AI methods, such as machine learning and deep learning, to estimate pedestrian volumes based on bus dashcam videos (Lian et al., 2024) and street-view imagery (Chen et al., 2020; Hankey et al., 2021). However, a common challenge in these works and in the broader use of AI is the substantial effort required to calibrate complex models. In the context of transportation agencies, it would require personnel with AI expertise. Recently, Large Language Models (LLMs) – a type of AI algorithm – have become accessible to the general public, which may help practitioners with certain tasks by eliminating the need for developing their own AI models. The present work introduces a novel method for estimating pedestrian volume at

intersections that relies on the use of readily available LLMs (ChatGPT-4 Vision) to visually assess satellite images.

ChatGPT-4 Vision (or GPT-4V) (Open AI, 2024) is available to ChatGPT Plus users and offers image-to-text functionalities, allowing users to input images for visual assessment and interpretation. From a practical perspective, it is an easy-to-apply tool that can be utilized in a vast number of applications, including those in traffic engineering. For instance, Driessen et al. (2023) tested the potential of GPT-4V in assessing driver road safety risk perception from street view images. GPT-4V and humans scored 210 images based on risk perception, revealing a linear correlation of 0.69. In another application, Huang et al. (2024) utilized GPT-4V to predict pedestrian behavior from street view images for autonomous driving purposes.

Given the challenges of collecting pedestrian volume data across entire jurisdictions – whether through CCs, STCs, or DD models – there is a growing need for innovative, scalable methods. This Chapter explores the potential of GPT-4V to visually assess satellite images as a novel approach for estimating pedestrian volumes at intersections in jurisdictions which have limited pedestrian traffic data available. More specifically, the objectives of this Chapter are to:

- I. Explore GPT-4V's ability to visually assess satellite images, particularly its capacity to identify various types of land use and traffic elements such as crosswalks and transit stops.
- II. Propose and evaluate a method to rank intersections based on pedestrian activity using GPT-4V and satellite images.
- III. Propose and evaluate a method to integrate the site rankings generated in Objective II with observed pedestrian volume data from a small number of selected intersections, enabling the estimation of pedestrian volume for all sites. This objective is necessary as GPT-4V currently lacks the capability to correlate extracted features from satellite images with actual pedestrian volumes.

Figure 5-1 presents the framework for this Chapter. In line with the proposed objectives, the Chapter is divided into three main steps. The first step involves an exploratory investigation of GPT-4V's ability to visually assess images, aiming to determine its potential for estimating pedestrian exposure at intersections. The second step applies GPT-4V to rank the intersections based on satellite images. In addition, inconsistencies and potential biases in GPT-4V responses are also evaluated. The final step combines the rankings generated by GPT-4V with field-measured pedestrian volumes at specific sites, enabling the estimation of pedestrian volume at all locations.

Since each step is independent of the others, we have chosen to present the method and results for each step separately. Additionally, the next section describes the datasets used, while the final section presents conclusions and practical applications.

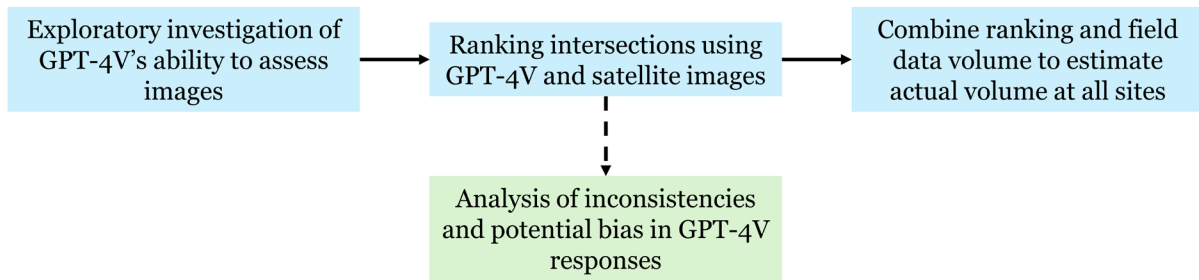


Figure 5-1 Framework

5.2 Datasets

Two types of datasets were utilized in this Chapter. The first dataset pertains to pedestrian volume information, which was used to calculate observed volumes at intersections and generate true rankings. The second dataset involves extracting satellite images for intersections where pedestrian data is available, which are then inputted into GPT-4V.

A total of 48 signalized intersections equipped with CC monitoring systems in the Region of Waterloo, Ontario, Canada, were considered for analysis. Due to the absence of a full year of data in a portion of the intersections, preventing the calculation of AADPTs, Monthly Average Daily Pedestrian Traffic (MADPT) for October 2023 was calculated to represent pedestrian volume at each site. The analyzed sites exhibited a wide range of MADPT, with the following statistics: average = 927; standard deviation = 1,395; minimum = 21; median = 382; and maximum = 7,093.

Satellite images from Google Maps (Google Maps, 2024) were extracted from each of the 48 sites using Google Maps Static API (Google, n.d.) and httr package from R (Wickham, 2022). Different types of images can be generated based on multiple factors, including: a) the type of map, including roadmap (standard Google Maps image), satellite view, hybrid view (satellite image with labels), street view, and others; and b) the zoom level, which ranges from 0 (farthest possible) to 20 (closest possible). Hybrid images were used due to their realistic representation, which aids in identifying land use types, and because they include additional information from labels. The decision on the zoom level was guided by existing DD models, which typically employ land use and socioeconomic variables within a radius of 100-800 meters of the intersection center. This led to potential zoom levels of 16 to 18. Squared images centered at the coordinates of each intersection were extracted, with different zoom levels representing real distances as follows: zoom level 16 = 1,108 x 1,108 meters; zoom level 17 = 554 x 554 meters; zoom level 18 = 277 x 277 meters (these are dimensions for the sites used in this study, but they may vary based on the site's latitude). After evaluating several images, zoom level 17 was chosen for its perceived optimal balance between the image dimensions and level of detail. This zoom level also allowed for a reasonable number of labeled locations and effectively depicted transportation features, such as transit stops.

The quality of the image is an input parameter in the Google Maps API. First, the size of the image is defined, followed by the quality (using the scale parameter). Two options are

available: scale = 1 (which means one pixel per unit of size) and scale = 2 (which means two pixels per unit of size). We used images with scale = 2 for better quality.

We acknowledge the limitations of using 2D satellite images for understanding pedestrian activity. These images do not capture socioeconomic features, such as vehicle ownership rates or household income, which are commonly used as explanatory variables in DD models. Additionally, 2D images cannot detect building heights, making it impossible to consider the "density" factor in the analysis; variables like population and employment density are frequently included in DD models. Therefore, future research could explore the use of different types of images, such as 3D images, and/or combine the images with quantitative measures of the site characteristics (e.g., population density within a given radius) to assist GPT-4V or other models in assessing the images more accurately.

To illustrate this, our initial plan was to create a collage with six images to represent each site: one satellite image showing the land use, one zoomed satellite image depicting the intersection's geometry (e.g., number of lanes, presence of crosswalks), and four street view images offering a 360-degree perspective. We believed that this approach would better characterize the intersection in terms of pedestrian activity compared to using a single satellite image. However, this plan was undermined by GPT-4V's limitation of assessing a maximum of ten images per inquiry (discussed in the next section).

5.3 Can GPT-4V visually assess satellite images?

The Google Maps satellite images (Google Maps, 2024) presented in Figure 5-2 were used for an initial assessment of GPT-4V's ability to identify different types of land use and traffic elements. Both sites represented in Figure 5-2 are from the Region of Waterloo, Ontario, Canada. Figure 5-2a shows a location with mixed land use, including commercial and residential uses, whereas Figure 5-2b is predominantly residential.

In the initial assessment using Figure 5-2a, GPT-4V was assigned several tasks. First, it was asked to count the number of commercial places in the image. GPT-4V responded that

it was unable to perform a physical count of objects within an image and recommended the use of GIS tools for such tasks. Subsequently, GPT-4V was asked to identify the different types of land use observed in the image. It listed eight types of land use, including commercial, recreational, and residential, providing detailed reasons for each identification. For example, it stated: *“Residential: although not as prominently marked, there may be residential use within the area, often integrated with commercial zones in urban settings”*. Additionally, GPT-4V was tasked with estimating the percentage of each identified land use type. It provided estimates such as 25% for commercial land use and 20% for transportation spaces like roads and parking lots.



Figure 5-2 Satellite images used for the preliminary investigation (29): (a) Erb at Caroline, depicting a scenario with mixed land use; (b) River at Manchester, depicting a scenario with mostly residential land use

A similar process was conducted using Figure 5-2b. GPT-4V was able to identify the dominance of residential land use: *“From this image, the land use appears to be primarily residential, with single-family homes dominating the landscape...”*, indicating that it detected residential places based on the roof’s characteristics. GPT-4V estimated the residential land use to be 80%, while commercial and recreational land uses accounted for 5% and 10%, respectively.

In a subsequent task, GPT-4V was challenged with ranking three images representing sites with varying pedestrian activity levels – high, medium, and low. These images also exhibited notable differences in land use. In response, GPT-4V succinctly described the land use of each location and presented a table ranking the sites from highest to lowest pedestrian activity. For instance, the site identified as having the highest activity was described as a *“diverse mix including cultural institutions, a public square, and commercial spaces”*. GPT-4V explained the rationale for ranking this site as the highest activity as follows: *“This area is likely to have the highest pedestrian activity throughout the year due to cultural events, daily commercial activity, and the presence of public spaces that encourage foot traffic”*. Notably, GPT-4V accurately ranked the sites according to their actual pedestrian activity levels.

These preliminary analyses confirmed the potential of using GPT-4V to rank sites according to pedestrian activity based on satellite images. The subsequent sections delve deeper into exploring how this capability can be applied to a larger context, such as assessing an entire jurisdiction. However, before moving on, it is necessary to highlight some limitations of GPT-4V's operation:

- When this work was developed (between December 2023 and May 2024 using model GPT-4 Turbo), GPT-4V had a couple of input limitations, including: a) a maximum of 30 inquiries within a 3-hour period, and b) a limit of 10 images per inquiry. These limitations add complexity to assessing entire jurisdictions, especially when ranking numerous intersections. They will be further discussed in this Chapter.

- GPT-4V exhibits inconsistency. While some level of stochasticity in its answers was anticipated, at times, different responses were observed for the same question. For example, GPT-4V occasionally indicated it could not rank sites due to a lack of visual assessment capability or provided varying estimates for land use percentages from the same images in different inquiries. These types of inconsistencies are another aspect explored in this Chapter.
- While GPT-4V can rank sites based on pedestrian activity, it cannot assign pedestrian volumes to the images. This limitation was somewhat expected, considering that the relationship between land use and pedestrian volume (i.e., DD model coefficients) varies across different jurisdictions (Sobreira & Hellenga, 2023a – Chapter 3).

5.4 Ranking intersections: GPT-4V and satellite images

This section is divided into three sub-sections. First, the method for ranking the sites using GPT-4V and satellite images is presented. Second, the results are presented and discussed. Finally, a comprehensive discussion of GPT-4V’s inconsistency and potential bias is provided.

5.4.1 Method

The first step was to create the prompt for the inquiries, an important aspect when working with AI applications (Schulhoff et al., 2024). The authors followed a trial-and-error process in the development of the prompt, with the final version of the prompt as follows:

“Each of the files uploaded here represents the satellite image of one intersection (highlighted by the red pin). The header of the images contains the site ID, address, and dimensions of the image. Your task is to rank the intersections in terms of their level of pedestrian activity based on the content of the images (i.e., land use, pedestrian facilities, or other relevant information). Create a descending rank where the first position represents the

intersection with the highest pedestrian activity. In your response, provide a table with the ID and Rank of each intersection. Do not include any additional text or information beyond the ranking.”

Two terminologies used to characterize the interactions with GPT-4V are presented. “Failed inquiry” represents an interaction in which images were uploaded to GPT-4V, but it did not provide a final response (i.e., ranking) due to a processing error. Conversely, “successful inquiry” refers to an interaction in which GPT-4V provided a ranking to the inquiry.

The second step involved determining the number of images to be input per inquiry. Ideally, the maximum allowed number of ten images would have been used. However, it was observed that using this number frequently resulted in failed inquiries. This led to significant delays in interactions with GPT-4V, making it impractical to obtain multiple rankings. Tests were conducted using four, six, and eight images, and it was found that six images proved to be a suitable number for proceeding with the following steps.

A methodological challenge emerged: there are 48 sites to be ranked, but only 6 sites can be input per inquiry. To address this, a simulation process was proposed. The rationale behind this approach is to rank each site (from 1 = highest activity to 6 = lowest activity) multiple times by considering different arrangements of sites. In the first round of the simulation ($k = 1$), the 48 sites are randomly divided into eight samples, each containing six sites. These samples are then individually ranked by GPT-4V ($R_{i,k}$). The process is repeated k times, and the average rank score for each site ($\bar{R}_i$) is calculated using Equation 5.1. Finally, the GPT-4V's final rank is determined by arranging the 48 sites in ascending order of $\bar{R}_i$. We did not implement any specific controls over the generation of the random samples. For future studies, we suggest employing strategies that ensure all (or as many as possible) site pairs are compared (i.e., each site appears in a sample with every other site at least once).

$$\bar{R}_i = \frac{1}{k} \sum_k R_{i,k} \quad (5.1)$$

Where:

k = Number of simulation rounds.

$R_{i,k}$ = GPT-4V rank associated with site i in round k .

$\bar{R}_i$ = Average rank score for site i .

The final step involves defining the parameter k . A higher value of k implies more interactions between different sites, potentially resulting in a higher quality final ranking. However, an increase in k also leads to increased time costs. For this work, k was set at 15, aiming to balance the number of simulation rounds and the time required to interact with GPT-4V. For context, we were able to obtain approximately 50 successful inquiries per day, considering GPT-4V's usage limit and the frequency of failed inquiries.

In an additional analysis not included in this work, we observed that low values of k (between 1 and 5) resulted in poor and highly variable linear correlations between GPT-4V and true rankings. Around $k = 8$ to 10, the correlation stabilized at levels similar to those achieved at $k = 15$. Our point is that the number of sites analyzed and the number of images input per inquiry affect the required k for achieving stable and accurate results. Future applications should investigate the stability of results when varying k .

5.4.2 Results

Figure 5-3 displays a scatterplot comparing the true rank obtained from ordering the MADPT of each site with the GPT-4V rank estimated using the process described above. A linear correlation of 0.728 and a MAE of 8.125 were achieved in this comparison suggesting that there is a moderately strong correlation between the GPT-4V ranks and the true ranks. It is important to note that there are several sources of error, including the inherent capacity of

GPT-4V to analyze satellite images and link them to pedestrian activity, as well as the method used to calculate $\bar{R}_i$. We were unable to find rank accuracy metrics from other studies in the literature; hence, we could not compare the performance of GPT-4V's estimates of sites ranks with rank estimates from other methods. However, in the next section of this paper we present a proposed method for estimating site MADPT values from the GPT-4V rankings and we compare the accuracy of those estimates to exposure estimation accuracy from the literature.

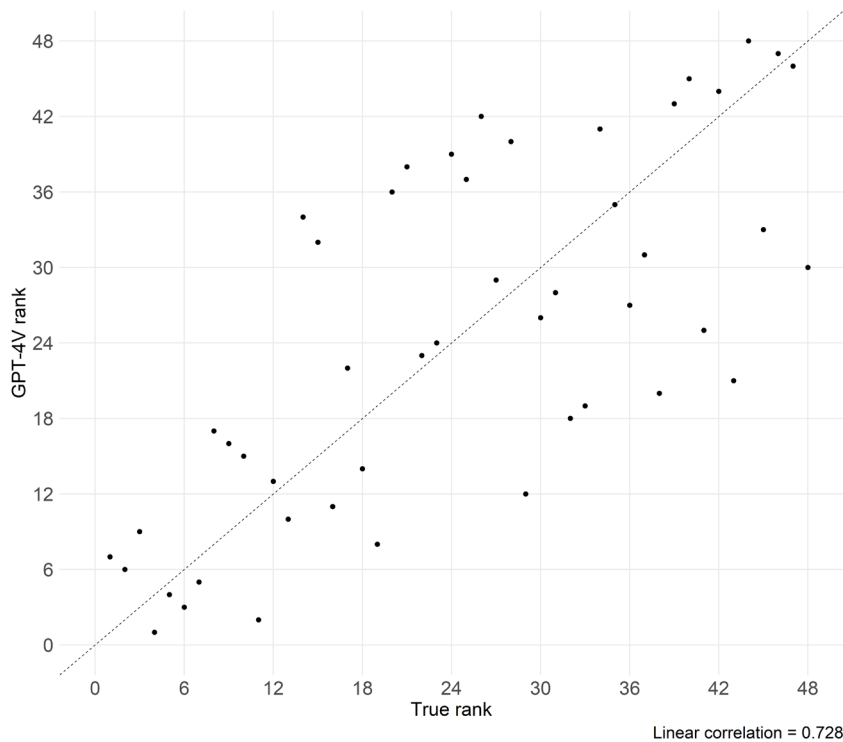


Figure 5-3 Scatterplot: true rank versus GPT-4V rank

Figure 5-4 illustrates the MADPT as a function of the true rank. For the analyzed sites in the Region of Waterloo, the relationship between rank and MADPT is highly non-linear. The differences between MADPT of the top-10 ranked sites are quite large, whereas the

differences between MADPT for sites ranked 15 or more are small (and probably of less practical significance). Consequently, for practical applications, practitioners should be more concerned about the ranking accuracy of the top-ranked sites than of the bottom-ranked sites.

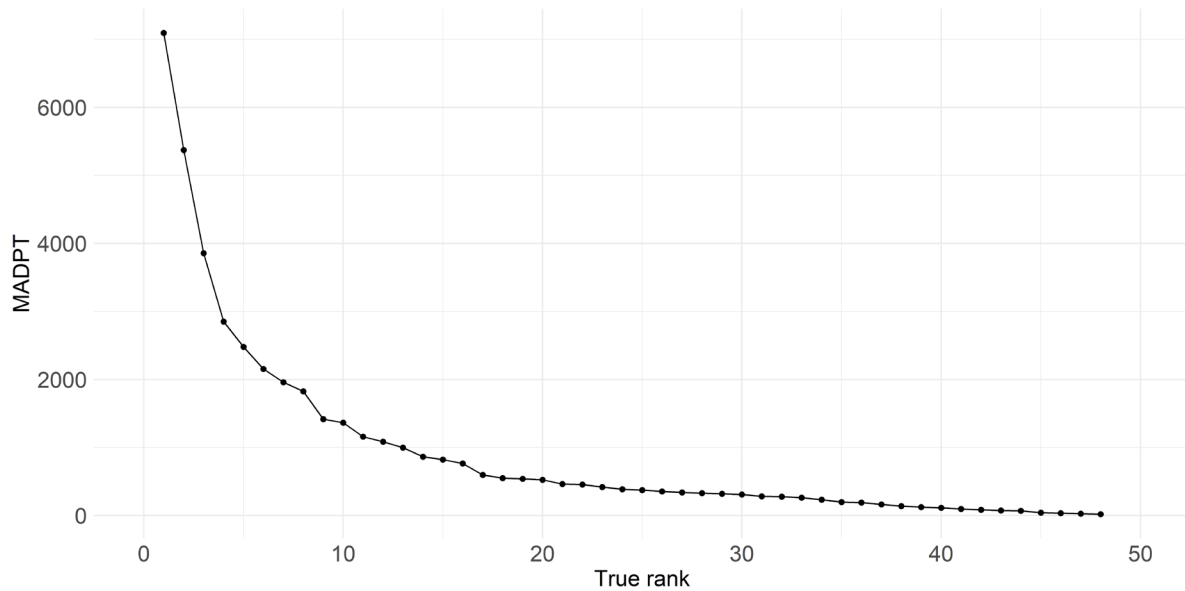


Figure 5-4 MADPT and true rank

Table 5-1 presents a confusion matrix between the true and GPT-4V ranks, categorized into bins. For example, the first cell of the first row represents sites correctly placed by GPT-4V into Bin 01-10, while the second cell of the first row shows sites belonging to Bin 11-20 but incorrectly placed into Bin 01-10. Additionally, the MAE of the rank error for each bin of the true rank is presented. As indicated, 41.7% of the sites were placed in the correct bins (highlighted in green cells). Considering adjacent bins as well (highlighted in yellow cells), the performance of GPT-4V improves to 79.2%.

Table 5-1 demonstrates that GPT-4V exhibited superior performance in ranking the top sites (Bin 01-10) compared to others, as evidenced by the lowest MAE and the highest classification accuracy (7 out of 10 sites) observed in this category. This enhanced

performance at the top of the ranking was somewhat expected, given the notably higher MADPT of these sites. It is expected that sites with very high MADPT will also exhibit distinguished levels of land use, socioeconomic, and transportation characteristics, facilitating their identification by GPT-4V. From a practical perspective, this is a good outcome considering the greater concern about sites at the top of the ranking.

Conversely, a higher magnitude of error is observed in the other bins. However, its impact on estimating pedestrian volume is less substantial: inaccuracies in ranking middle or lower-tier sites result in smaller discrepancies in MADPT compared to misallocations at the top. For instance, the MADPTs for sites at positions 10, 20, and 30 are 1365, 525, and 311, respectively.

Table 5-1 Confusion matrix: true rank versus GPT-4V rank

		True rank				
		01-10	11-20	21-30	31-40	41-48
GPT-4V rank	01-10	7	3	0	0	0
	11-20	3	3	1	3	0
	21-30	0	1	4	2	3
	31-40	0	3	4	2	1
	41-48	0	0	1	3	4
MAE		4.6	9.1	9.7	8.0	9.5

Note: Cells highlighted in green indicate correct bin classification, while those in yellow represent adjacent bins.

5.4.3 Exploring GPT-4V responses

After interacting with GPT-4V hundreds of times, several concerns about its operation were noted. The most frustrating aspect was the frequent processing errors, in which questions and images were input but GPT-4V failed to provide a final answer (failed inquiries). Though the exact frequency of failed inquiries was not computed, they occurred frequently enough to be a concern.

Another concern was that GPT-4V sometimes generated ranks identical to the order in which the images were input. Given that there are 720 different arrangements for ordering six input images, the likelihood of the images being presented to GPT-4V in the same order as their true ranks is very low, and therefore we would expect it to be equally rare for GPT-4V to produce a ranking identical to the image order. When such identical rankings were noted, we followed up with GPT-4V to inquire whether we should rely on that ranking or if an error had occurred. GPT-4V responded that “*The ranks provided were not based on actual analysis of the content of the images, as it was a placeholder to demonstrate the process...*”, indicating that an error had occurred during the processing of images and ranking generation. Therefore, when the resulting ranking matched the image order, that interaction was treated as a failed inquiry.

Additionally, inconsistencies in responses and potential bias towards the order in which images were input were observed, which will be discussed in the following sections.

5.4.3.1 Inconsistencies in GPT-4V responses

One noticeable aspect of GPT-4V was its inconsistency across different inquiries. More specifically for this application, GPT-4V derived different rankings when the same dataset of images was input multiple times. An experiment was conducted to demonstrate it.

The experiment involved identifying six sites with varying levels of pedestrian activity. The site selection process consisted of dividing the true ranking into six equal bins and selecting the top site from each bin (i.e., ranks 1, 19, 38, 57, 76, and 95). This site selection process maximized the differences between the levels of pedestrian activity at the sites and theoretically should have made it easier for GPT-4V to rank the sites due to expected visual differences in their land use. Subsequently, the site images were randomly ordered and GPT-4V was asked to rank the sites from 1 to 6. This process was repeated 30 times. Figure 5-5 illustrates the frequency with which GPT-4V ranked each of the six sites, with the teal-colored column highlighting the true rank.

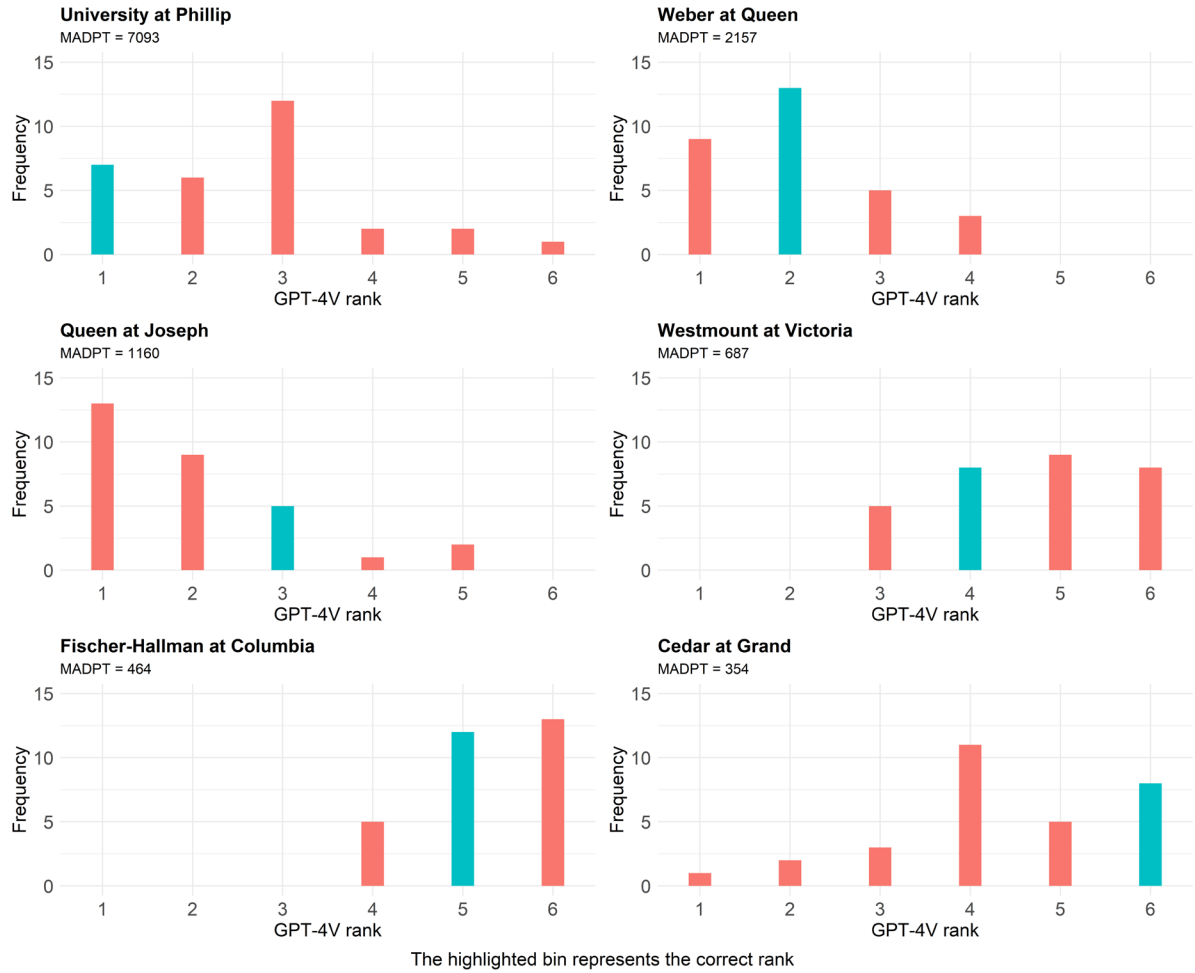


Figure 5-5 Exploring GPT-4V’s consistency in ranking the same six images

Figure 5-5 confirms GPT-4V’s inconsistency in generating different rankings for the same set of images. This outcome was somewhat expected, considering that stochasticity is inherent to GPT-4V's functioning. This inconsistency is another contributing factor to errors in ranking sites. However, despite the lack of consistency in the rankings, a discernible pattern is observed: sites with true ranks of 1 to 3 are often ranked at the top of the list, while sites with true ranks 4 to 6 are consistently ranked lower.

The combination of having ranks in the same order as the images were input and GPT-4V's general inconsistency raised suspicions about potential bias in how GPT-4V ranked the sites based on the image order. This led to the following question: does the order in which the images are input affect how GPT-4V ranks them? This question is explored in the next section.

5.4.3.2 Exploring potential bias in GPT-4V responses

The Friedman test was used to examine potential biases in GPT-4V outcome rankings based on the order in which the images were input. The Friedman test is a non-parametric statistical test used to analyze related samples. It shares similarities with repeated measures Analysis of Variance (ANOVA) but focuses on the ranking of measurements rather than the actual measurements themselves. An example is provided to clarify the Friedman test. Consider a scenario where five patients are experiencing back pain and are scheduled to undergo surgery to address the issue. Each patient is asked to numerically rate their pain from 0 to 100 at three different time points: before surgery, one week after surgery, and one month after surgery. The pain measurements from each patient are ranked from 1 to 3, and then the ranks for each time point are summed across all patients. The Friedman test is applied to determine if there are significant differences in pain ratings across the three time points. The null hypothesis for the Friedman test is that there are no differences in pain ratings between the time points.

The Friedman test was adapted to investigate potential bias arising from the order in which images were input. In this study, 15 rounds of simulation were conducted, with 6 images input per inquiry. With a total of 48 sites, each round required 8 different samples, leading to 120 interactions with GPT-4V in total (15 rounds x 8 samples). In the context of the example above, the interactions can be linked to patients, the order of image input (from 1 to 6) represents the “time points”, and the resulting rank from GPT-4V serves as the measurement. Since the samples were randomly generated (i.e., random image orders), it is

expected random distributions of ranks by GPT-4V if there was no bias. In other words, the null hypothesis was expected to hold and not be rejected.

The application of the test, however, resulted in the rejection of the null hypothesis ($p\text{-value} = 0.004$), indicating that GPT-4V is biased by the order in which the images are input. The sum of ranks for each image order increased progressively from position 1 (369) to position 6 (477), indicating that GPT-4V tended to rank images that were input first as having higher pedestrian activity (i.e. ranks closer to 1).

The Friedman test result was unexpected given the strong linear correlation between the true ranks and the GPT-4V ranks as presented in Figure 5-3. To better understand these results, it was decided to investigate each of the 120 interactions with GPT-4V using three indicators to measure the deviation between image order, true rank, and GPT-4V rank. $\bar{E}_{IG}$ and $\bar{E}_{IT}$ (Equations 5.2 and 5.3) represent the deviation between image order and GPT-4V rank, and between image order and true rank in each interaction, respectively. Evaluating both indicators simultaneously helps in identifying whether a GPT-4V ranking similar to the image order should be expected or not given the true ranking. Both indicators vary from 0 to 3, representing cases of identical order and completely different order, respectively. An additional indicator, $\bar{E}_{TG}$ (Equation 5.4), is also used to represent the deviation between GPT-4V and true ranks. Although it does not directly contribute to the bias issue, it indicates the ranking accuracy of each interaction.

$$\bar{E}_{IG} = \frac{1}{n} \sum_{j=1}^n |j - G_j| \quad (5.2)$$

$$\bar{E}_{IT} = \frac{1}{n} \sum_{j=1}^n |j - T_j| \quad (5.3)$$

$$\bar{E}_{TG} = \frac{1}{n} \sum_{j=1}^n |T_j - G_j| \quad (5.4)$$

Where:

n = Number of images per inquiry (six in this work).

j = Image order.

G_j = GPT-4V rank associated with image in order j .

T_j = True rank associated with image in order j .

Figure 5-6 shows a scatterplot of the indicators $\bar{E}_{IG}$ and $\bar{E}_{IT}$ for the 120 interactions with GPT-4V. The color of each dot is associated with its $\bar{E}_{TG}$. Four quadrants (Q1 – Q4) were created using the indicators' midpoint of 1.5. Each quadrant could be interpreted as follows:

- Q1 – High $\bar{E}_{IT}$ and high $\bar{E}_{IG}$ (68 interactions): true and GPT-4V ranks are different from the image order. This suggests interactions that are not biased and probably accurate.
- Q2 – Low $\bar{E}_{IT}$ and high $\bar{E}_{IG}$ (11 interactions): GPT-4V ranks are different from the image order, but true ranks are similar to the image order. This suggests interactions that are not biased and probably inaccurate.
- Q3 – Low $\bar{E}_{IT}$ and low $\bar{E}_{IG}$ (15 interactions): true and GPT-4V ranks are similar to the image order. This suggests interactions that could be biased but are probably accurate.
- Q4 – High $\bar{E}_{IT}$ and low $\bar{E}_{IG}$ (26 interactions): true ranks are different from the image order, but GPT-4V ranks are similar to the image order. This suggests interactions that are biased and probably inaccurate.

Interactions that fall in quadrants Q3 and Q4 raise concerns about potential bias. It is challenging to definitively state that there is bias in Q3 because the image order corresponds to the true order and therefore the fact that GPT-4V rankings correspond to the image order can indicate a bias or that the rankings are unbiased and accurate. Conversely, Q4 represents

a more severe case of potential bias. While it is known that the true ranks differ from the image order, GPT-4V ranked the sites similarly to how the images were provided.

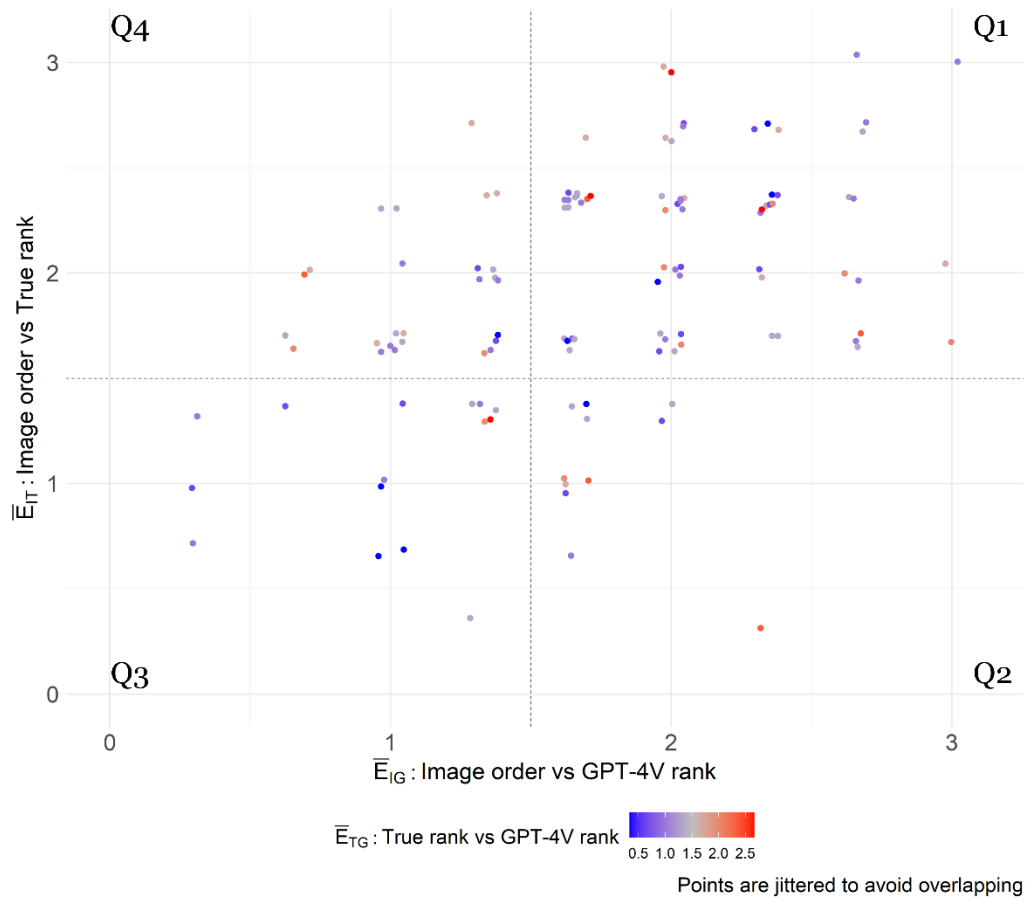


Figure 5-6 Exploring the bias in GPT-4V's responses

The Friedman test was conducted again considering two scenarios: a) sites in Q1 and Q2, and b) sites in Q1, Q2, and Q3. The null hypothesis was not rejected in either case, with p-values of 0.472 and 0.885, respectively. This suggests that the group of interactions in these quadrants is not biased. Consequently, this raised the suspicion that only a portion of the interactions (those in Q4) were biased by the image order.

A detailed investigation was conducted on the interactions in Q4 to evaluate whether the bias was triggered by a specific factor. Initially, the sites and satellite images frequently involved in these interactions were assessed, but no particular pattern was observed. Subsequently, it was decided to repeat some of the interactions in Q4 using the same images and image order. Each interaction was replicated 20 times, and the $\bar{E}_{IG}$ was computed. Stochasticity was observed across the different $\bar{E}_{IG}$, with more interactions occurring in Q1 and fewer in Q4. Despite our efforts, we were unable to identify the reasons for the bias towards image order that affected certain interactions. This bias could be attributed to factors such as GPT-4V's momentary processing capacity. Notably, these interactions appeared to occur randomly and did not target any specific group of sites, which meant that any bias-related errors were spread across multiple interactions.

In conclusion, it was decided to acknowledge the limitation that approximately 21.6% (26 out of 120 interactions) of the rankings generated by GPT-4V are located in Q4 and may be biased. However, there are reasons to believe that these biased interactions occur randomly; therefore, their error will be diluted among the other interactions, somewhat minimizing the problem, especially when a decent number of simulations are performed.

5.5 Estimating pedestrian volume from rankings

The method described in the previous section provides a ranking of sites with respect to the expected level of pedestrian activity at each site. These site rankings could be beneficial for jurisdictions, particularly in selecting locations for pedestrian centric interventions. However, just knowing the site rankings does not provide an estimate of the pedestrian volume, which could be used for road safety analyses. In this section, a method is proposed to estimate pedestrian volumes based on rankings. Subsequently, the results are presented and discussed.

5.5.1 Method

The proposed method integrates the ranking derived from GPT-4V with the collection of pedestrian volume data from key sites. Figure 5-7 presents the method, illustrating a case in which a dataset with 16 sites is used. The first step involves splitting the GPT-4V ranking into categories based on percentiles. In Figure 5-7, four categories were established, which could be interpreted as very high, high, low, and very low pedestrian activity. The second step is to collect field volume data (STC, daily or weekly volume, MADPT, or AADPT) for the sites that represent the category boundaries. For this work, it was assumed that October MADPT data is available for these sites. The third step is to check for inconsistencies in the ranks. Figure 5-7, for example, the site ranked as the fourth position has a lower volume than the site ranked as the eighth position, indicating an error in GPT-4V ranking. To rectify this, the positions for the two sites are swapped (i.e., the site originally ranked as fourth is now eighth, and vice versa). The fourth step involves estimating the volume of the sites with unknown volume using the category boundaries. Linear and non-linear methods were considered and are discussed below.

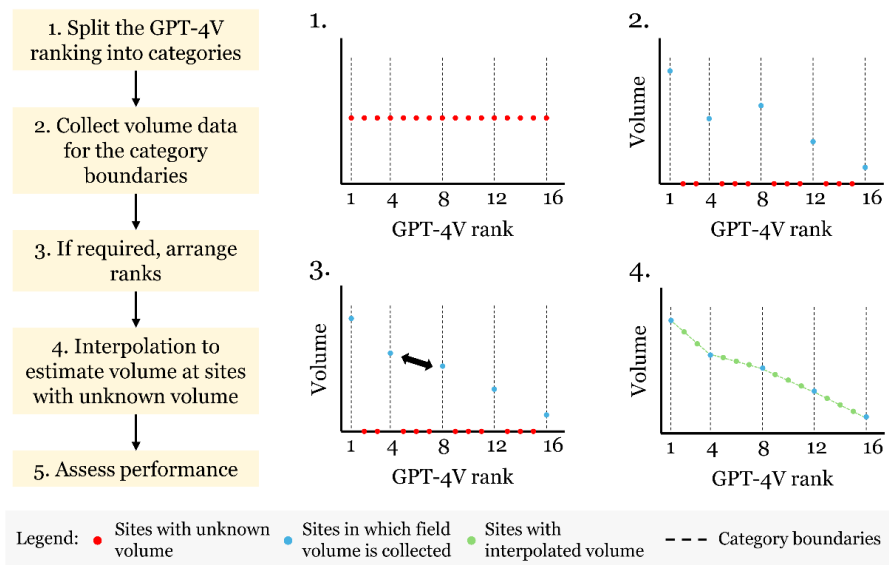


Figure 5-7 Method to estimate pedestrian volume from rankings

Linear interpolation was initially employed for all categories: the volume of the sites within each category was predicted considering the boundary volumes and the GPT-4V ranks. However, as can be seen from Figure 5-4, the pedestrian volumes of the sites at the top of the ranking (i.e., first categories) do not follow a linear distribution. Therefore, a large error was observed when using linear interpolation for these sites. This pattern may be specific to the Region of Waterloo; hence, it was decided to investigate the volume distribution across sites of three additional jurisdictions: Milton, ON, Canada; Pima County, AZ, US; and Toronto, ON, Canada. Table 5-2 presents a summary of the pedestrian volume, represented as the AADPT between 2022 and 2023, for each jurisdiction. Note the significant variation in pedestrian activity magnitude across jurisdictions.

Table 5-2 Additional jurisdictions: descriptive statistics

Jurisdiction	Number of sites	AADPT (2022-2023)		
		Average	Minimum	Maximum
Milton, ON, Canada	38	320	27	1,097
Pima County, AZ, US	87	29	2	2,528
Toronto, ON, Canada	49	12,193	651	30,675

Figure 5-8 displays the scaled volume plotted against its associated percentile, with each jurisdiction represented by a distinct color. Examining the sites in the first quartile, it is evident that Pima County and Waterloo exhibit a non-linear distribution, Toronto shows a distribution close to linear, and Milton presents a mix of linear and non-linear distributions. For sites in the other quartiles, distributions close to linear are observed across all jurisdictions.

To address the non-linearity observed in the first category in certain jurisdictions, non-linear models were developed to predict volume for these sites. Exponential, logarithmic, and polynomial models were calibrated using pedestrian volume observations from the first three boundaries (using all boundaries resulted in poorly-fitted curves in the

first category). The pedestrian volume (MADPT) served as the dependent variable, while the GPT-4V rank was the independent variable. Initial analysis revealed that the logarithmic model offered the best accuracy and was therefore adopted for this study. In summary, two methods are now used for estimating pedestrian volume at sites with unknown volume: a) linear interpolation for all categories, and b) logarithmic model for the first category and linear interpolation for subsequent categories.

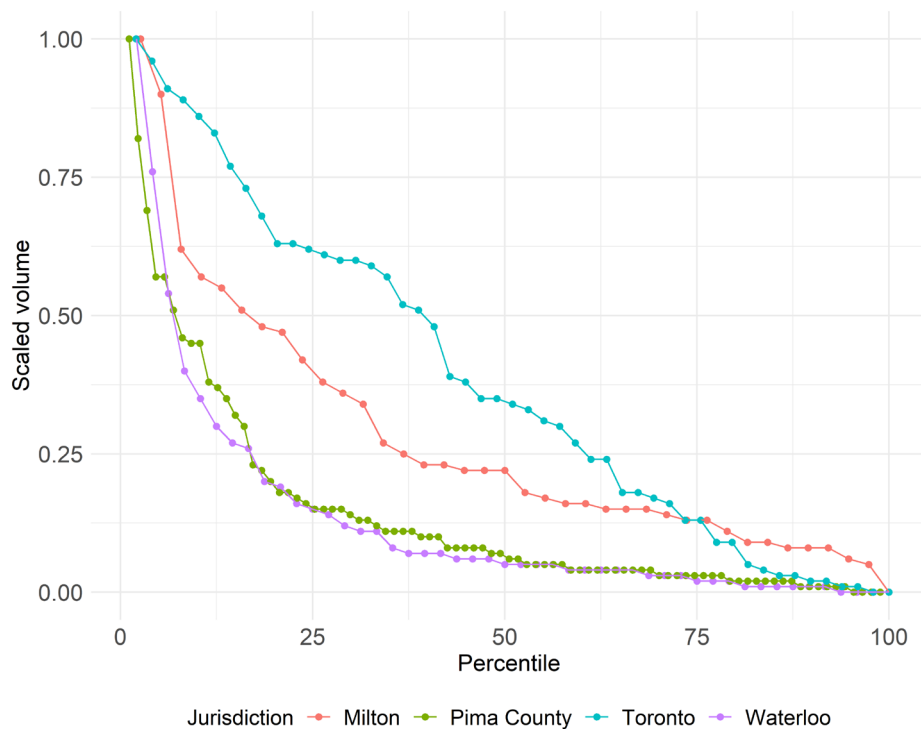


Figure 5-8 Pedestrian volume distribution across different jurisdictions

From a practical perspective, analyzing the shape of the distribution of the volumes of sites where field volumes were collected (Plot 3 from Figure 5-7) can provide insights into whether the logarithmic model method should be employed. Figure 5-9 illustrates two different scenarios. If there is a significant difference in the slope of a straight line fit to the

volumes of the first two boundaries compared to the slope of a line through the other boundaries, the logarithmic model method is recommended (Figure 5-9a). Otherwise, linear interpolation for all categories is advised (Figure 5-9b).

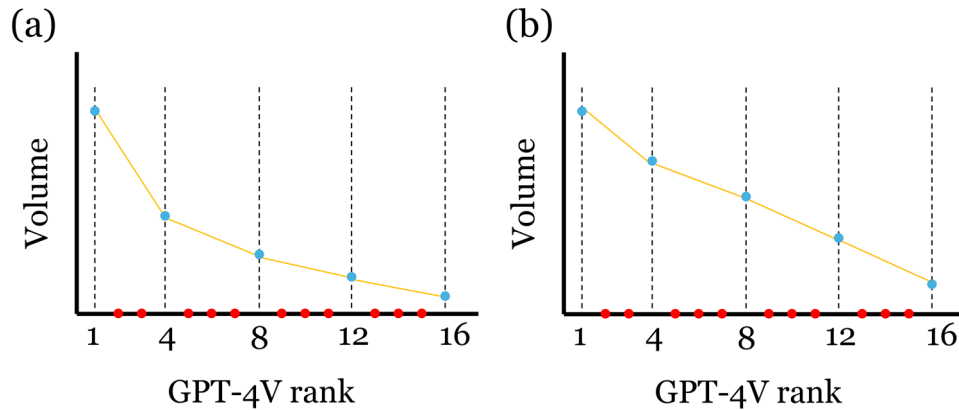


Figure 5-9 Determining if the Logarithmic model or linear interpolation approach should be adopted on the basis of field volumes

Returning to the method itself (Figure 5-7), the final step is to assess the performance of the predicted volumes. Error metrics E (Equation 2.11) and MAPE (Equation 2.10) are employed for performance assessment.

5.5.2 Results

Figure 5-10 presents the performance comparison of both the linear interpolation and the logarithmic model methods for different number of categories (ranging from 4 to 11). For this analysis, the maximum of 11 categories (resulting in 12 boundaries) was chosen to represent the collection of field volume data for 25% of the 48 sites in the dataset. In the figure, the horizontal dashed line at 0.60 serves as a benchmark performance level based on alternative approaches for estimating pedestrian volume when neither CCs nor STCs are

available for all sites within a jurisdiction. One such approach is the local development of DD models, which prior work suggests yield $E = MAE / \bar{Y}$ around 0.60 and 0.65 (Munira et al., 2020; Schneider et al., 2021). This approach typically necessitates access to 50 to 70 sites with available pedestrian volume data for model development. Another approach is the local calibration of an existing DD model developed in another jurisdiction. Prior work shows that this approach can provide similar performance to locally developed DD models but requires fewer sites for implementation (Sobreira & Hellinga, 2023b – Chapter 4). There is only a single study that investigated the local calibration of DD models, and it used limited sample sizes. Although it has shown potential for its simplicity and reasonable level of accuracy, further investigation is required to validate this method. Both locally developed and calibrated DD model approaches, besides requiring available sites with pedestrian volume data, also depend on the existence and usability of land use, socioeconomic, and transportation GIS databases, introducing additional complexity to their application.

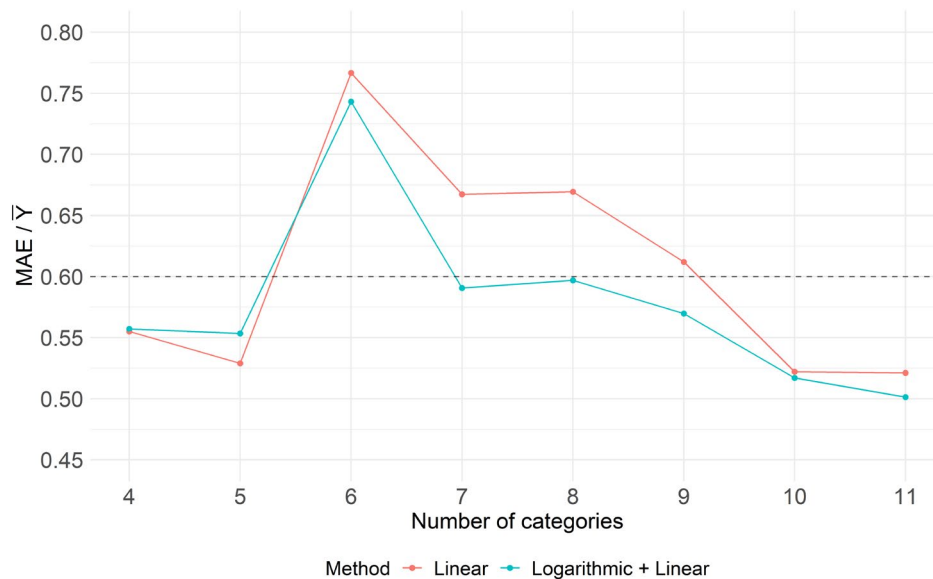


Figure 5-10 Performance assessment: estimating pedestrian volume from ranks

The performance shown in Figure 5-10 for estimating pedestrian volume from GPT-4V ranks is quite promising as it meets, and in some cases surpasses, benchmark approaches with $E = MAE/\bar{Y}$ ranging between 0.50 and 0.55 for certain categories. Moreover, this method is simpler to implement compared to DD model approaches: a) it requires pedestrian volume data for only a few sites, b) there is no need for land use, socioeconomic, or transportation GIS databases, and c) it does not involve developing complex statistical models. Instead, it relies on relatively simple tasks, such as: extracting satellite images, interacting with GPT-4V, and employing a straightforward method to estimate pedestrian volume at sites with unknown volume.

To provide additional context on GPT-4V's performance, Table 5-3 presents the accuracy in terms of MAPE when splitting the GPT-V4 rankings into five categories to identify sites for collecting field volumes and using the logarithmic model for interpolation. The results are compared to an alternative approach that jurisdictions might adopt in the absence of pedestrian data, which involves hiring third-party companies to provide volume estimates based on crowd-sourced data. For instance, StreetLight uses anonymized location data from various mobile apps to estimate pedestrian volume (StreetLight, 2023). The results are categorized by volume range to avoid bias towards low-volume sites, and both MAPE metrics are calculated based on MADPT estimates. Except for the first category, the GPT-4V method outperforms the crowd-sourced data approach.

Table 5-3 GPT-4V versus StreetLight

Volume range	GPT-4V		MAPE: StreetLight
	#Sites	MAPE	
Low (< 150)	23	273	151
Mid (150-500)	37	51	89
High (500-1,000)	14	51	82
Very high (>1,000)	27	49	71

Expanding on a datapoint from Figure 5-10, Figure 5-11 showcases the true MADPT versus predicted MADPT using four categories and the logarithmic model approach. Figure 5-11a provides an overview of the entire dataset, while Figure 5-11b focuses on sites with MADPT lower than 1,000. The proposed method achieved an accuracy performance of $E = MAE/\bar{Y} = 0.56$ with a strong linear correlation of 0.88. Notably, Figure 5-11a reveals a significant underprediction for sites at the top of the ranking (high volume). This discrepancy arose due to GPT-4V erroneously ranking the site at the top of the true ranking as the fourth position. Consequently, the first category boundary was incorrectly set at 2,852 instead of 7,093, resulting in the underprediction of sites occupying the top three positions in the true ranking. In general, Figure 5-11b demonstrates instances of accurate predictions but also highlights cases of under and overpredictions resulting from imperfections in the GPT-4V ranking.

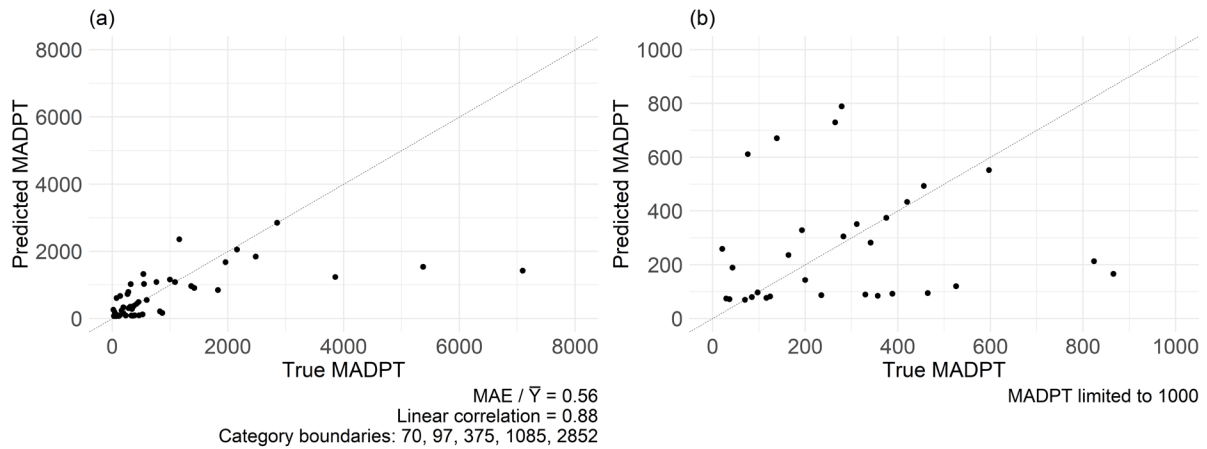


Figure 5-11 Predicted versus true MADPT using four categories and the logarithmic model approach: (a) all sites, (b) zoomed at sites with MADPT lower than 1,000

In terms of performance differences between the linear interpolation and logarithmic model approaches, Figure 5-10 reveals varying performances across different categories. Notably, the logarithmic model approach demonstrates superior performance in categories 7,

8, and 9, while similar performances are observed in the remaining categories. It was observed that the logarithmic model approach outperformed the linear interpolation approach when sites at the top of the true ranking (positions 1 or 2) – representing sites with extremely high pedestrian volume – were included in the category boundaries. This aligns with expectations, as linear interpolation for the first category did not perform well when there was a significant difference in volume between the first two boundaries.

The shapes of the curves in Figure 5-10 are somewhat unexpected. One might anticipate a systematic improvement in performance with an increase in the number of categories, but this trend is not observed (categories 4 and 5 are superior to categories 6 to 9). Additionally, the high magnitude of error in category 6 is atypical. To explore these further, the sites selected to define the category boundaries in these categories were closely examined. It was found that the good performance in categories 4 and 5 and the poor performance in category 6 are linked to specific characteristics of the sites chosen to form the category boundaries. In other words, the results depicted in Figure 5-10 are specific to the particular final ranking generated by GPT-4V and do not necessarily reflect overall (or populational) performance metrics. To ensure results are not limited by the specificity of the particular ranking generated by GPT-4V, two simulation approaches are proposed and implemented in the next section.

5.5.3 Simulation approaches

It was demonstrated previously in this work (Figure 5-5) that GPT-4V is inconsistent and provides different rankings for the same set of images. With this established, it is evident that the results presented in the previous section are subject to randomness. Put differently, if the interactions with GPT-4V were repeated another 120 times, a similar final ranking in terms of general trends would be expected, but individual sites would likely have different ranks. In an ideal scenario, we would replicate the 120 inquiries to GPT-4V multiple times to examine the statistical variation in the results in Figure 5-10. However, given the large

resource requirements associated with carrying out such replications, we decided to pursue two simulation approaches to estimate the statistical variation and obtain more generalizable results.

The first simulation approach relies on the original final ranking generated by GPT-4V, but with a different selection of sites that represent the category boundaries (Plot 1 from Figure 5-7). This would represent the shifted positions obtained from new rankings. To implement it, the ranks of the sites at the boundaries were randomly switched with a site that is close to each boundary. For example, when $n = 1$, an integer from the set $\{-1, 0, +1\}$ is randomly selected and added to the boundary rank. For instance, if the original boundary rank was 8 and the randomly selected number is -1, then site ranked 8 is switched with the site ranked 7. This process is repeated separately for each boundary to allow multiple random replications. After completing this procedure for all boundaries, steps 2, 3, and 4 from Figure 5-7 are conducted as usual. This procedure was carried out for $n = 1$ and $n = 2$.

The above simulation process is attractive in that it is intuitive. However, the error that is introduced (value of n) is arbitrary and uniform across all boundaries. The second simulation approach addresses these two issues by using the observed distribution of rank errors and varies this error across the boundaries. The procedure is outlined below:

1. Calculate the deviation between the true ranks and the original GPT-4V ranks for all 48 sites. This is the observed distribution of rank errors.
2. As shown in Figure 5-3, rank errors are larger when the true rank exceeds 12. To account for this, two distributions of rank errors were created: one for sites ranked between 1 and 12 and another for sites ranked between 13 and 48.
3. Define an auxiliary variable $V_i = T_i + \varepsilon_a$ where T_i is the true rank of site i , ε_a is the random error from the error distribution for sites ranked between 1 and 12 when $T_i \leq 12$, otherwise from the error distribution for sites ranked between 13 and 48. This was done for all sites ($i = 1, 48$).
4. Sort auxiliary variable V_i in descending order to create a new simulated GPT-4V ranking.

5. Check the linear correlation between simulated and true ranking. If it is greater than 0.75, return to step 3. Otherwise, steps 1 to 4 from Figure 5-7 are conducted as usual. This step ensures that the simulated ranking maintains a linear correlation similar to the original ranking (0.728)

The simulation approaches were applied to create 100 simulated rankings, and their performance, along with that of the original method, is summarized in Figure 5-12. In Figure 5-12b-d, the line represents the median of the accuracy across the 100 replications, while the background depicts the 25th and 75th percentiles.

Figure 5-12 presents intriguing findings. Firstly, the curves' trends now align with expectations for the simulation approaches, showing improved performance as the number of categories increases. This supports the hypothesis that the unexpected trend observed in the original application arose from random outliers in the original GPT-4V ranking. Furthermore, the results once again demonstrate performances surpassing those of benchmark approaches, highlighting the methodology's potential.

Secondly, notable differences in performance between the linear interpolation and logarithmic model methods were observed. While the first simulation approach yielded similar results between the methods, the second approach showed better outcomes for the logarithmic model method, especially with fewer categories. This outcome was expected due to GPT-4V's initial misranking of the top-2 sites in the true ranking, placing them at positions 6 and 7. The inclusion of these sites as part of category boundaries significantly impacted the performance gap between the two methods. The first simulation approach, which often excludes these sites as boundaries, resulted in comparable outcomes for linear interpolation and logarithmic model methods.

Lastly, the linear interpolation and logarithmic model methods tend to converge as the number of categories increases (Figure 5-12d). This convergence was expected for two reasons: a) assuming linear interpolation adversely affects overall performance when the first

category is wider and top-ranked sites are included in the first boundary; and b) with more categories, the application of the two methods becomes more similar since the logarithmic model is solely applied to the first category.

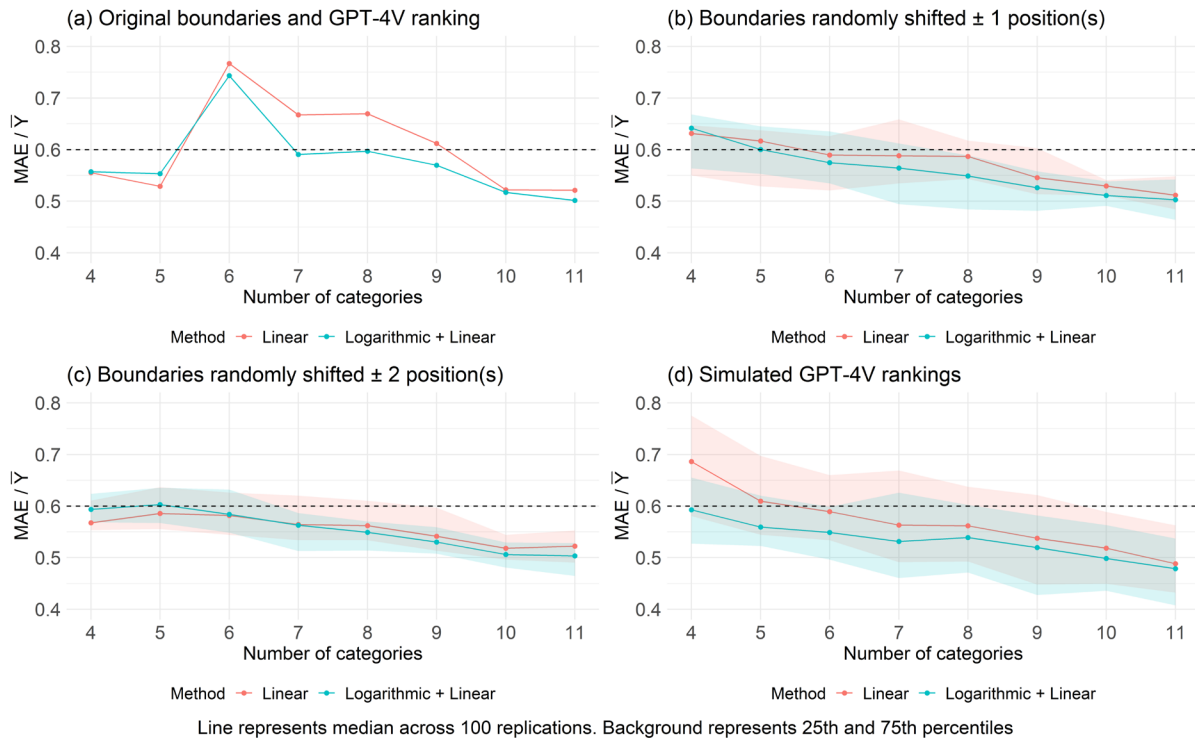


Figure 5-12 Performance assessment of original and simulated approaches

5.6 Conclusions and practical application

This Chapter explored the use of GPT-4V as a tool to aid jurisdictions in assessing pedestrian activity within their intersections. It involved several key steps, including an initial assessment of GPT-4V's capabilities in interpreting satellite images, the development of a method to rank sites based on pedestrian activity using GPT-4V and satellite images, and the creation of a method to estimate pedestrian volume from the GPT-4V ranking.

The initial evaluation of GPT-4V revealed its potential in interpreting satellite images, particularly in identifying different types of land use and estimating their shares within the image. Additionally, GPT-4V demonstrated its ability to rank sites based on pedestrian activity using satellite imagery.

A method utilizing GPT-4V and satellite imagery to rank intersections was proposed and applied to 48 sites within the Region of Waterloo, ON, Canada. The method yielded promising results, including a linear correlation of 0.728 between GPT-4V ranking and true ranking. Additionally, it demonstrated strong accuracy in identifying intersections with high pedestrian activity, with seven out of the top-10 sites in the true ranking appearing within the top 10 positions estimated by the proposed method. From a practical perspective, these findings underscore the potential of GPT-4V in identifying intersections with high pedestrian activity, which could help with site selection for implementing pedestrian-centric interventions.

The use of GPT-4V for ranking purposes is not without limitations. Apart from the restrictions on inputting a maximum of ten images per query (which was further reduced to six in this work) and conducting 30 inquiries within a 3-hour period, inconsistent and biased responses were observed. GPT-4V provided different responses (i.e., rankings) for the same image datasets. More significantly, we observed that 26 of the 120 rankings generated by GPT-4V may be associated with some level of bias toward the order in which the images were presented. Despite investigation, the reasons for this bias could not be identified. However, it was observed that these biased responses were not triggered by specific sites and appeared to occur randomly. This randomness somewhat mitigated the impact on ranking generation, as it introduced a degree of random error that is diluted across the rounds of simulation.

In possession of the GPT-4V ranking, a method combining the ranking and the collection of field pedestrian volume in key intersections was proposed to estimate the actual pedestrian volume for the sites in the ranking. Again, very promising results were attained, with performance similar to (and sometimes better than) the existing conventional

approaches that are typically used to estimate pedestrian volumes in sites that lack pedestrian data information, such as the development of DD models.

The greatest benefit of the GPT-4V method is its simplicity. Unlike the development of DD models, it does not require dozens of sites with available volume data. Comparable results were obtained using five categories, which would necessitate data collection at only six sites (12.5% of the entire dataset). Furthermore, it does not require land use, socioeconomic, or transportation GIS databases, nor the development of complex statistical models. Instead, it relies on extracting satellite images, interacting with GPT-4V, and applying the relatively simple methods proposed in this work.

For future work, it is suggested: a) applying the method to larger datasets to investigate whether its performance is dictated by the number or the fraction of sites used for category boundaries, b) applying the method to other jurisdictions to validate its potential, c) exploring and comparing different zoom levels and types of satellite images (including 3D images) to develop the GPT-4V ranking, d) employing other types of pedestrian volume data to represent the category boundaries, specifically counts with shorter durations than the monthly average used in the present work, e) exploring alternative LLMs besides GPT-4V for satellite image interpretation and ranking generation, f) examining GPT-4V's capability to identify traffic volume variations both between days (e.g., weekdays versus weekends) and within a day (e.g., peak hours) based on satellite imagery, g) assessing whether including additional information directly to the satellite image, such as the number of transit stops, commercial establishments, etc., within a given radius of the intersection, would enhance GPT-4V's performance in ranking the sites, and h) evaluating GPT-4V's performance in visually identifying transportation-related elements from satellite images, such as transit stops or schools, that are important for explaining pedestrian activity, using metrics like Visual Question Answering.

Chapter 6

Expanding Short-Term Pedestrian Volume Counts to Average Annual Daily Counts at Intersections: Temporal Pattern Indices and a Comparison of Statistical Land Use and Simple Threshold Methods

This Chapter is based on the following journal paper:

Sobreira, LTP; Hellinga, B. Expanding Short-Term Pedestrian Volume Counts to Average Annual Daily Counts at Intersections: Temporal Pattern Indices and a Comparison of Statistical Land Use and Simple Threshold Methods. *Transportation Research Record: Journal of the Transportation Research Board*, 2024.
<https://doi.org/10.1177/03611981241281736>

6.1 Overall context

The focus changes to the expansion factor method in Chapter 6. Given the existing limitations and literature gap on the identification of factor groups using temporal indicators and on the magnitude of error associated with expanding 8-hour STCs to AADPT identified in Chapter 2, the objectives of this Chapter are to fill these knowledge gaps by: (I) proposing and applying pedestrian volume temporal indicators that capture seasonal trends in order to identify factor groups; (II) determining the thresholds of these indicators for the identification of factor groups; (III) developing multinomial logit regression models to link STC sites with a particular factor group; (IV) quantifying the error associated with the expansion of 8-hour STCs to AADPT; and (V) evaluating the implementation of simpler models and their trade-off regarding complexity versus performance. Objective V is a practical investigation that may help jurisdictions in their decision-making based on the level of information they have available.

The structure of this Chapter follows the steps a practitioner would take to apply the expansion factor method to STCs using factor groups (Figure 6-1). The first step involves identifying factor groups from sites with CCs and calculating expansion factors for each set of factor groups. In the second step, factor groups are modeled as a function of land use, socioeconomic, and transportation attributes, enabling the association of STC sites (i.e., sites with unknown traffic patterns) with factor groups. The final step encompasses the expansion of STCs to AADPT, applying the expansion factors and models obtained from steps 1 and 2, respectively.

Additionally, the two subsequent sections outline the data sources and present an assessment of types of expansion factors. These sections provide complementary information required for the implementation of the core method. The Chapter is concluded with a section presenting conclusions and practical recommendations.

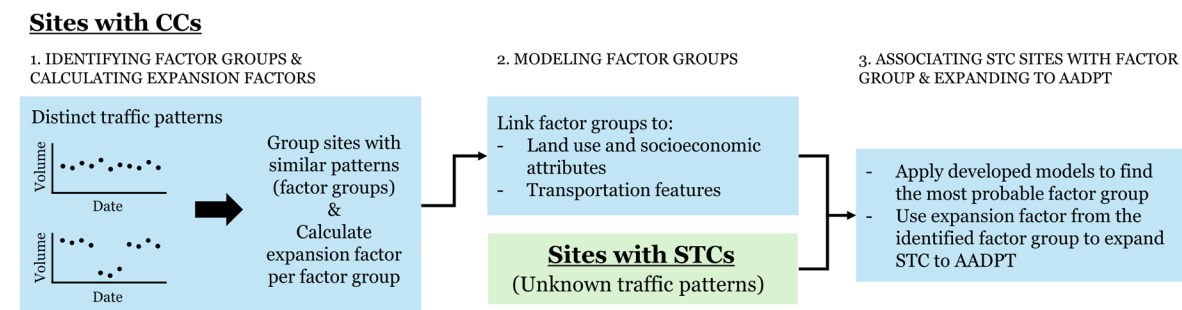


Figure 6-1 Proposed framework

6.2 Data sources

This Chapter uses data from the Region of Waterloo, Ontario, Canada. The Region of Waterloo is formed by three cities (Cambridge, Kitchener, and Waterloo) and four townships. It has an area of 1,369 km² and a population of about 650,000. The region shares a similar latitude with the City of Toronto, resulting in four distinct seasons, including cold

winters (average daytime high temperatures $\leq 0^{\circ}\text{C}$ (32°F) for December through February) and hot summers (average daytime high temperatures $\geq 24^{\circ}\text{C}$ (75°F) for June through August) (Weather Spark, 2023). The seasonality of the weather may have a significant impact on pedestrian traffic patterns.

Pedestrian volume data were collected from camera-based CC stations located at 78 signalized intersections, deployed in the three cities that compose the region. These stations were primarily installed for signal timing and coordination purposes, placing them in urban areas. Consequently, this dataset may underrepresent sites with predominantly recreational pedestrian trip activity. The CC stations record motorized vehicles, cyclists, and pedestrian counts for each turning movement and crosswalk at one-minute intervals on a continuous (24/7) basis. Only sites that met the following two data quality/quantity criteria were included in the study:

1. The CC station was considered to be collecting data properly on a given day if: a) at least 72 (out of 96) 15-minute count intervals had at least one motorized vehicle counted on any approach; and b) the total daily pedestrian volume > 0 .
2. The CC station was operational for the 12-month period from May 1, 2022, to April 31, 2023, and had properly collected data from at least 95% (345) of these days.

In addition, a filter was applied to the pedestrian counts each month of each site to remove days that represent outliers: daily (24 hour) pedestrian volumes that were outside the limits of either a) the 3rd quartile plus 1.5 times the interquartile range or b) the 1st quartile minus 1.5 times the interquartile range. This filtering criteria resulted in the removal of 1.1% of days across the 78 sites.

In the subsequent sections of this Chapter, results are presented based on 8-hour STCs obtained from all available days throughout the year (i.e., days when the CC station was operating properly and were not removed by the filtering criteria), as well as from a subset of days referred to as “candidate” days. It is a common practice for jurisdictions to collect

STCs, particularly for signal timing purposes, on days characterized as typical in terms of traffic volume. In this Chapter, the “candidate” days specifically refer to non-holiday Tuesdays, Wednesdays, and Thursdays in the months of April, May, June, September, October, and November. These months are chosen to exclude periods of inclement weather and school holidays.

6.3 Assessing types of expansion factors

Three types of expansion factors are commonly employed in practice. The first and simplest involves using 7 day-of-week (DOW) and 12 month-of-year (MOY) factors for a total of 19 factors. The second requires estimating a factor for every day-of-week-of-month (DOWOM) for a total of 84. The third and most complex approach utilizes factors for each individual day-of-year (DOY), often referred to as the disaggregated method (for a total of 365 factors). As discussed in Section 2.1, implementing DOY factors have practical challenges. Thus, this Chapter exclusively considered the DOW & MOY and DOWOM factors.

The following procedure was employed to determine the superior performance between the DOW & MOY and DOWOM factors. The two types of expansion factors were calculated for each of the 78 sites with CC stations. Given the availability of CCs for all sites, multiple STCs can be retrieved from the CCs (and for each AADPT is estimated). The absolute percent error ($APE_{i,j}$) between observed and estimated AADPT was computed for every individual STC observation j of each site i and then averaged across all sites and STC observations. It is worth noting that this procedure does not account for the existence of factor groups. As a result, STCs and expansion factors were obtained from the same CC station. Subsequent sections of this Chapter make use of factor groups, and a similar procedure is applied to assess the performance of the expansion. The main difference in the procedure is that expansion factors for each factor group are calculated based on the average of the factors of the sites that belong to a given factor group.

Using data for the 78 sites available and STCs extracted from candidate days, the DOWOM factor method provides better accuracy (MAPE = 13.6%) than the DOW & MOY method (MAPE = 15.9%) and therefore the subsequent steps of this Chapter are conducted using DOWOM expansion factors.

6.4 Identifying factor groups

This section outlines the steps taken to identify factor groups based on the 78 sites with CCs. The analysis focused on the time series of DOWOM expansion factors, chosen for their representation of scaled pedestrian volume. This scaling allowed for comparisons across sites with different magnitudes of pedestrian volumes.

A three-step method was applied. First, a visual inspection was performed to subjectively identify the factor groups. Second, temporal indicators were proposed to characterize the identified groups. Third, a clustering technique, specifically the k-means algorithm, was applied to group similar sites based on their temporal indicators. The optimal number of clusters (i.e., factor groups) was determined through three tests:

1. Total within-cluster sum of squared errors (TWSS): measuring the distance between elements within a cluster and the centroid of the cluster.
2. Silhouette test: comparing the distance between the centroid of each element and the centroid of its cluster with the distance between the centroid of each element and the centroid of the closest cluster. The test verifies whether the characteristics of each element are strongly associated with its grouping or whether the attributes of that element could be easily related to another group.
3. Gap statistic: comparing the sum of errors between groupings of the sample used with the sum of errors if a sample with random distribution (i.e., without grouping structure) was applied.

In consideration of practical applications and replicability of the results, an additional step was taken to investigate whether the factor groups identified using the clustering technique could have been determined using a simpler approach relying solely on the thresholds of the temporal indicators.

6.4.1 Visual inspection

Figure 6-2 illustrates the three primary temporal pedestrian activity patterns identified from the time series of DOWOM for the 78 sites with CCs in the Region of Waterloo. Each plot in Figure 6-2 corresponds to a specific site and displays its 84 DOWOM expansion factors. The dots are arranged chronologically from Monday to Sunday, and the vertical gridlines align with the Monday for each month. The color of the dots distinguishes between weekdays and weekends. Note that expansion factors above one are obtained due to the conversion from 8-hour volumes to daily volumes.

Factor Group #1 (Figure 6-2a) represents sites with consistent volumes on both weekdays and weekends throughout the year (i.e., low seasonality effects), implying probable associations with recreational and commercial activities due to the uniform exposure throughout the week. In contrast, Factor Group #2 (Figure 6-2b) illustrates sites characterized by reduced volume on weekends compared to weekdays (indicated by a high expansion factor value) and low seasonality effects, suggesting a correlation with working activities. Factor Group #3 (Figure 6-2c) also exhibits reduced activity on weekends but demonstrates low pedestrian exposure during school holidays (July and August in Ontario, Canada), indicating a likely association with the presence of schools.

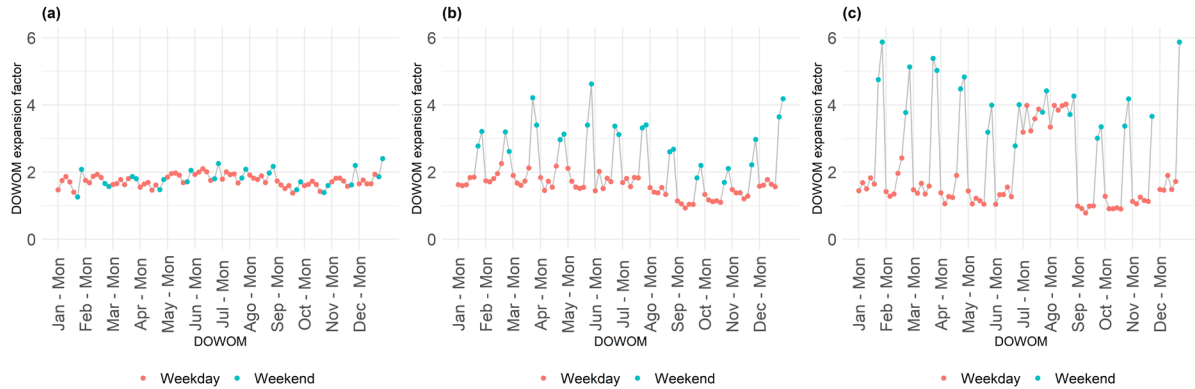


Figure 6-2 Different temporal pedestrian activity patterns for three different sites. (a) Fairway Road at Wilson Avenue; (b) King Street at Mt. Hope Street; (c) Charles Street at Cameron Street

In addition to the three patterns identified in Figure 6-2, it was also observed that the pedestrian volume at some sites presents a seasonal trend. Figure 6-3 elaborates on this observation, illustrating two sites that exhibit characteristics of Factor Group #1 (i.e., consistent volumes on both weekdays and weekends). One site maintains a consistent volume throughout the year (Figure 6-3a), while the other experiences reduced volume during the winter season (Figure 6-3b). This seasonality is believed to be associated with the “level of obligation” associated with different types of trip-making. For example, mandatory trips, such as work-related activities or home grocery shopping, may contribute to consistent volumes, whereas discretionary trips may exhibit seasonality.

The same seasonality pattern was also noted in Factor Group #2 sites, but not in Factor Group #3. This is expected from a phenomenological point of view: sites with pedestrian activity predominantly influenced by the presence of a nearby school are anticipated to have consistent volume throughout the year, except for school holidays. The next subsection introduces the temporal indicators proposed to capture the identified patterns in both Figure 6-2 and Figure 6-3.

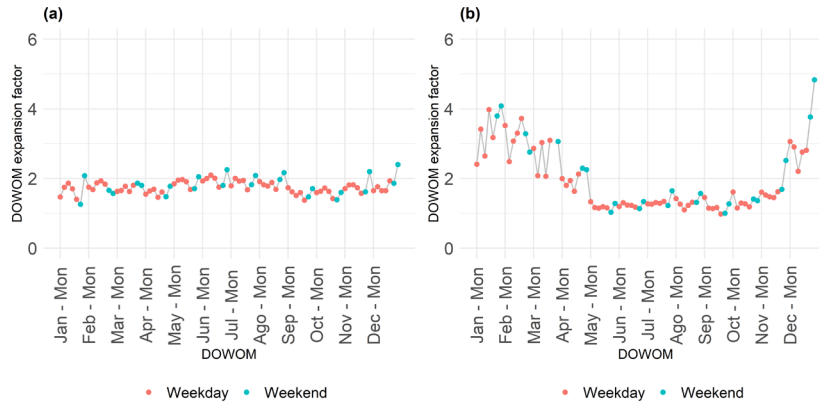


Figure 6-3 DOWOM expansion factors for two sites showing the effect of seasonality on pedestrian activity patterns. (a) Fairway Road at Wilson Avenue; (b) King Street at Highway 85 Ramp

6.4.2 Proposed temporal indicators for identifying factor groups

Temporal indicators calculated using the DOWOM expansion factors of each site are proposed for the identification of factor groups. The median was consistently chosen over the mean in the development of these indicators. This choice aims to prevent the generation of potentially skewed indicators that might arise if the mean was employed.

The first indicator captures the level of similarity between weekdays and weekends. An adapted version of the Weekend-Weekday Index (WWI), as developed by Miranda-Moreno et al. (2013), is proposed. The adapted WWI (WWI_i^*) is computed in two steps. First, the ratio of expansion factors between weekends and weekdays for each month is computed using Equation 6.1. Second, the median of these monthly ratios is calculated (Equation 6.2). It is important to note that months affected by school holidays (i.e., July and August) were excluded from the analysis. This decision is based on the understanding that the weekend-weekday ratio during this period may not be representative at some sites.

The second indicator assesses the level of similarity between school holidays and typical periods through the School Holiday-Typical Index (SHTI). SHTI is calculated as the

ratio between the median of the expansion factors on weekdays during school summer holidays (months of July and August) and the median of the expansion factors on weekdays during typical periods (Equation 6.3). In the context of this study, the typical period is defined as months not impacted by the winter season, spanning from April to November.

The third and final indicator, named the Deviation Across Months (DAM), is designed to capture seasonality by the magnitude of the deviation of expansion factors across months. DAM is computed in two steps. First, the expansion factor for each month is computed as the median of weekday factors for weekdays in that month (Equation 6.4). Second, the standard deviation of these monthly medians is calculated, considering all months except those affected by school holidays (Equation 6.5).

WWI*:

$$WWI_{i,m}^* = \frac{\frac{1}{2} \sum_{d=6,7} EF_{i,m,d}}{\frac{1}{5} \sum_{d=1-5} EF_{i,m,d}} \quad (6.1)$$

$$WWI_i^* = \text{median}_{m=1-6,9-12}(WWI_{i,m}) \quad (6.2)$$

SHTI:

$$SHTI_i = \frac{\text{median}_{m=7-8;d=1-5}(EF_{i,m,d})}{\text{median}_{m=4-6,9-11;d=1-5}(EF_{i,m,d})} \quad (6.3)$$

DAM:

$$EF_{Typical_{i,m}} = \text{median}_{d=1-5}(EF_{i,m,d}) \quad (6.4)$$

$$DAM_i = \text{standard deviation}_{m=1-6,9-12}(EF_{Typical_{i,m}}) \quad (6.5)$$

Where:

i = a given site

m = month of year

d = day of week (1 = Monday; 7 = Sunday)

$EF_{i,m,d}$ = DOWOM expansion factor at site i

WWI^* = Adapted Weekend-Weekday Index

$SHTI$ = School Holiday-Typical Index

$EF_{Typical_{i,m}}$ = typical expansion factor at site i and month m

DAM = Deviation Across Months

For all three indicators, the application of the above equations results in a value ≥ 0 . To avoid biasing due to very large values of the indicator, all indicators were truncated to a maximum value of 2.

These indicators considered school holidays and typical periods for the context of the Region of Waterloo. However, the conceptual framework is designed to be transferable to other jurisdictions. Practitioners can apply the temporal indicators by adjusting the months to align with the seasonal characteristics of a different jurisdiction.

6.4.3 Clustering sites

The k-means clustering algorithm was utilized to group sites based on their similarity in terms of the three temporal indicators defined in the previous section. The exploration of the optimal number of clusters (or factor groups) through the TWSS, silhouette test, and gap statistic indicated an optimal configuration with four or five factor groups. The choice of five groups was made as it aligned precisely with the factor groups previously visually identified.

Table 6-1 summarizes the characteristics and number of sites for each factor group. The configuration with four groups erroneously combined sites from Factor Groups #4 and #5 into a single group.

Table 6-1 Characteristics of each factor group based on their pedestrian volume

Factor group	# sites	Similar on weekdays and weekends	Reduced on weekends	Constant throughout the year	Reduced in school holidays	Affected by seasonality ¹
#1: Multipurpose	29	✓		✓		
#2: Commuter	24		✓	✓		
#3: Primarily school trips	15		✓		✓	
#4: Discretionary multipurpose	7	✓				✓
#5: Discretionary commuter	3		✓			✓

¹ Affected by seasonality other than school holidays

Figure 6-4 summarizes the DOWOM expansion factors for each of the factor groups identified in Table 6-1. The lines and dots in the figure represent the median of the factors within each group, while the grey background illustrates the 10th and 90th percentiles. The y-axis was limited to a value of 8, which is why the background is not displayed in some cases of Figure 6-4e. Consistent patterns are observed within each factor group, aligning with the traffic patterns that were visually identified. This consistency serves as an indication that the proposed temporal indicators effectively captured the characteristics of each site. The next subsection provides a characterization of the temporal indicators associated with each factor group.

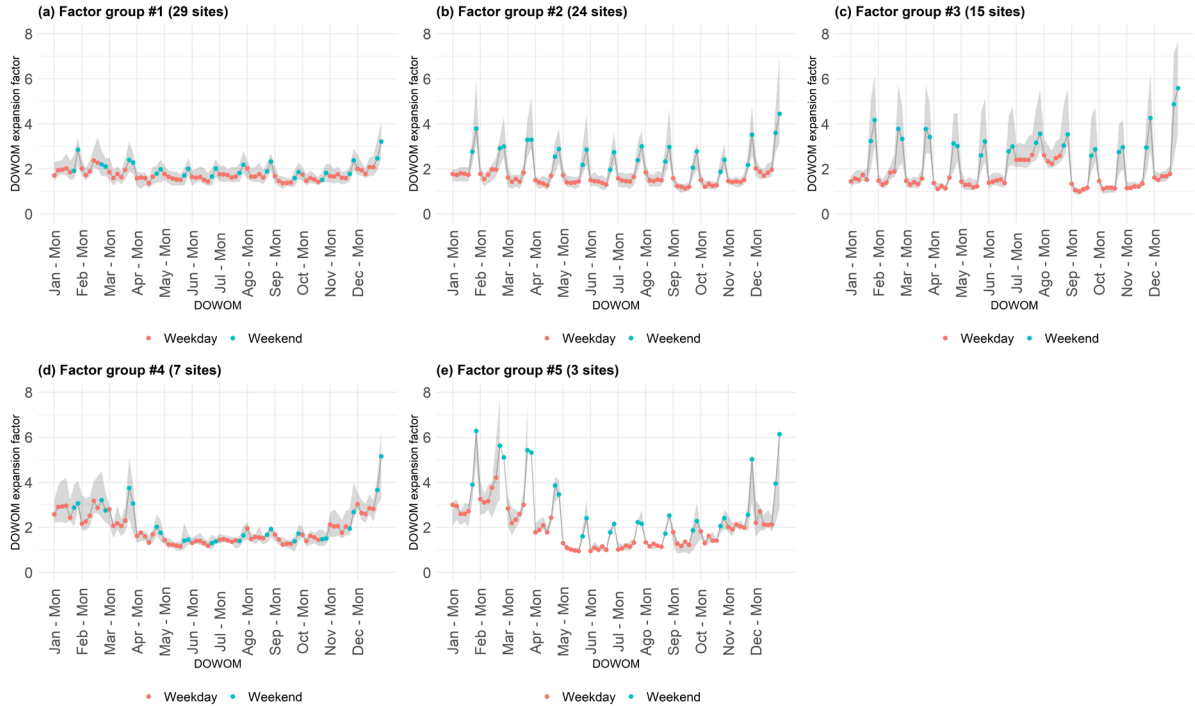


Figure 6-4 DOWOM expansion factor plots for factor groups identified through K-means clustering

For comparison, Medury et al. (2019) visually identified four factor groups based on hour-to-week expansion factors, using data from nine US states. The authors categorized these groups as urban commuter trails, central business districts, isolated recreational trails, or campgrounds/summer destinations. In contrast to our work, Medury et al. (2019) observed sites where the volume on weekends was significantly greater than on weekdays, possibly due to the location of CC stations on recreational trails. In the Region of Waterloo, the CC stations are all located at signalized intersections and were primarily installed for signal timing and coordination purposes, positioning them in urban – and likely commercial – areas. These locations are less likely to experience pedestrian travel patterns that are predominately recreational. Three sites present a WWI* lower than one, with a minimum value of 0.83, indicating weekend volumes greater than weekday volumes on average. However, the values

for the indicator were not low enough for the k-means algorithm to create a new factor group to represent “recreational” or “discretionary recreational” traffic patterns, and the sites were classified as multipurpose. It is possible, though, that these additional groups exist in the Region of Waterloo, but none of the sites monitored encompassed such characteristics.

In a separate study, Humagain & Singleton (2021) identified five factor groups based on hour-to-week factors and three groups based on monthly patterns. The primary distinction among the hour-of-week groups lies in the number of observed intraday peaks, either one or two. Their analysis, conducted in Utah, US, did not reveal any sites with greater volume on weekends compared to weekdays. Concerning the monthly patterns, one group showed the least variation throughout the year, another exhibited greater variation linked to weather sensitiveness, and a third displayed reduced activity during the summer months (June through August).

6.4.4 Temporal indicator thresholds for identifying factor groups

This section introduces an additional step to the core method for identifying factor groups presented in the previous three sections. It assesses the possibility of identifying factor groups by applying thresholds to the temporal indicators. This alternative approach, which replaces the clustering technique with a simpler method, may prove useful for practitioners looking to replicate the results of this work. We acknowledge that further research is needed to assess the spatial transferability of these indicators and thresholds.

Using the factor group classification obtained from the k-means procedure as the “true” factor group, different combinations of thresholds were tested in search of the optimal configuration. Thresholds were tested in increments of 0.05, ranging from 1.2 to 1.8 for WWI* and SHTI, and from 0.1 to 0.8 for DAM. All combinations were assessed based on the accurate labeling of sites.

The best combination of thresholds presented an almost perfect fit to the k-means classification. Out of 78 sites, only one site classified by the k-means clustering as Factor

Group #1 was misclassified by the indicator thresholds method as #3. This indicates a strong separation of the factor groups in terms of their temporal indicators. Figure 6-5 illustrates a decision tree with the thresholds and the logic for classifying a given site into a factor group.

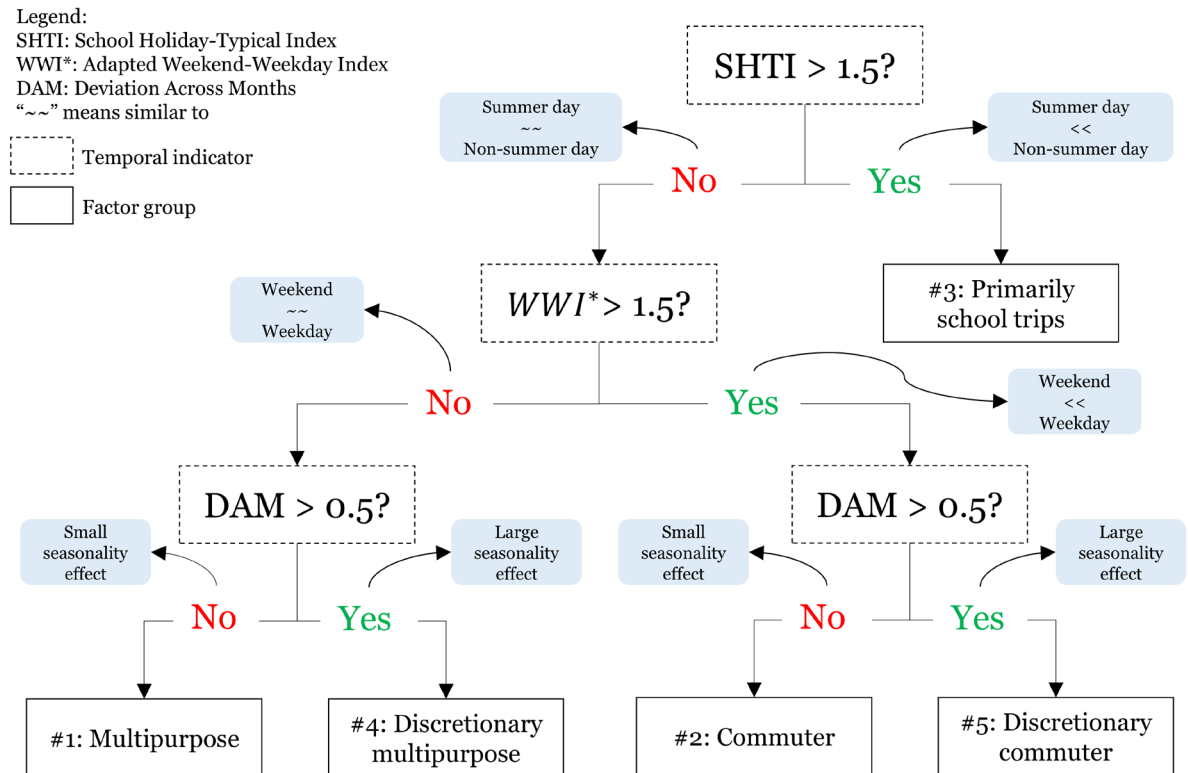


Figure 6-5 Decision tree: temporal indicator thresholds for identifying factor groups

The first step in the classification process involves assessing whether pedestrian volume at the site is primarily influenced by the presence of a school. If SHTI is greater than 1.5, then the site is associated with Factor Group #3. As discussed in the “Visual inspection” section, sites where pedestrian activity is strongly linked to schools also exhibit reduced volumes on weekends and a steady pattern throughout the year. Therefore, there is no need to assess WWI* and DAM for these sites. If SHTI is less than 1.5, the subsequent steps involve

assessing the values for WWI* and DAM to determine the site's factor group. The factor group classification derived from the thresholds was applied in the following steps of this work.

To provide context for the thresholds used in other studies, Nordback et al. (2019), while investigating non-motorized traffic (i.e., cyclists and pedestrians), employed the WWI to categorize sites across six US cities into three groups: Weekday Commute (average WWI ≤ 0.8), Weekly Multipurpose ($0.8 < \text{average WWI} \leq 1.2$), and Weekend Multipurpose (average WWI > 1.2). According to the authors, no substantial month-to-month variation was observed, leading to the omission of monthly adjustments. Note that WWI is calculated based on actual volumes, whereas WWI* is derived from expansion factors. As a result, they possess inverse interpretations: a higher WWI indicates higher weekend traffic compared to weekdays, while a higher WWI* suggests higher weekdays traffic relative to weekend. Additional sources for thresholds can be found in the literature (Hankey et al., 2014; Johnstone et al., 2018; L. Miranda-Moreno et al., 2013).

6.5 Modeling factor groups

After identifying the factor groups among the sites with CCs, the challenge lies in associating sites where only one or a few STCs are available with a specific factor group with the objective of employing an appropriate factor for expanding the STC to AADPT. To address this challenge, a Multinomial Logistic Regression (MLR) is proposed to establish a link between each factor group (i.e., the dependent variable) and land use, socioeconomic, and operational explanatory variables. The explanatory variables were collected in radii of 100, 200, 400, and 800 meters. The list of explanatory variables is provided in Table 6-2 as well as descriptive statistics for each variable for a radius of 400m.

Table 6-2 Site characteristic variables and descriptive statistics for radius of 400m

Variable	Unit	Source	Statistics (400m radius)			
			Avg.	SD	Min.	Max.
Pedestrian volume						
AADPT	Volume per intersection	Jurisdiction	710	852	41	4,546
Land use						
Commercial area (%) ¹	%	Region of Waterloo Open Data Portal	38.9	24.2	0.0	85.6
Industrial area (%) ¹	%		9.9	17.6	0.0	88.0
Residential area (%) ¹	%		40.9	27.8	0.0	96.6
Park area	Sq-m / 1,000		22.8	35.0	0.0	174.6
School - All types	Count		0.8	1.2	0	5
School - High and post secondary	Count		0.6	1.1	0	5
Places of worship	Count		2.1	2.8	0	12
Census-based						
Employment density ²	Employment per sq-km / 1,000	Environics: 2020	3.5	2.9	0.0	13.1
Household income ³	Median income per sq-km / 1,000	Census: 2016	66.9	33.8	0.0	181.4
Household size ³	Avg. number of people per household	Census: 2016	2.0	0.5	1.0	3.3
Population density ³	Population per sq-km / 1,000	Census: 2016	2.2	1.3	0.0	5.9
Walk to work (%) ³	%	Census: 2016	11.0	9.2	0.0	32.6
Operational						
Trails	Length (km)	Region of Waterloo Open Data Portal	0.7	0.7	0.0	2.7
Transit stops	Count		12.6	6.9	2	32

Note: Avg. = Average; SD = Standard deviation; Min. = Minimum; Max. = Maximum.

¹Percentage based on building footprints: area for a given land use divided by total footprint area.

²Environics is a point database that provides employment information. Employment density was calculated by dividing the sum of employment within a given buffer by the buffer area.

³Census-based information is available for each census zone. The value of the variable for each site was calculated as the weighted average, considering the overlapping area between the buffer area and each census zone.

A rule-of-thumb in the development of MLRs is to ensure that at least ten observations are available for each class of dependent variables. However, Factor Groups #4

(Discretionary multipurpose) and #5 (Discretionary commuter) have seven and three observations, respectively. In light of this, it was decided to exclude group #5 from the analysis while retaining group #4. The model results were interpreted with caution, given the small sample size used for this group.

Fourteen explanatory variables (Table 6-2) were collected within four different radii. This resulted in the possibility of generating millions of variable sets, but testing each one is impractical. To address this, a simulation process was implemented with three steps. First, a radius was randomly assigned to each variable. Second, the linear correlation of every pair of variables was examined, and one variable of the pair was randomly removed if the correlation exceeded 0.50. Third, using the remaining variables, the MLR was calibrated using a stepwise process. The simulation was repeated 100,000 times. To identify the best model (or something close to it), the evaluation included accuracy of classification (confusion matrix), McFadden pseudo- R^2 , Akaike Information Criterion (AIC), and Variance Inflation Factor (VIF). Additionally, a 5-fold cross-validation procedure was employed to inspect the best model for overfitting and coefficient stability.

In anticipation of the detailed results discussed in the following section, it was observed that the magnitude of error in expanding STCs of sites from Factor Group #3 (the group that is likely related to the presence of schools) is significantly higher than in sites from other factor groups. Consequently, it was decided to propose simpler approaches based on only two factor groups: one encompassing sites from Factor Groups #1, #2, and #4, and the other including sites from Factor Group #3. Three approaches were tested:

1. Binary logistic regression, following a procedure similar to MLR.
2. Presence of schools: a site is categorized in Factor Group #3 if there is a school within a specified radius. We considered two models: one considered the presence of any school, the other only high/post-secondary schools. Radii of 100, 200, 400, and 800 meters were tested for optimal classification.
3. Presence of schools and commercial land use: a site is categorized in factor group #3 if there is a school (or high/post-secondary) within a specified radius, and the percentage of

commercial land use within a specified radius does not exceed a given percentage. The land use constraint aims to minimize the misclassification of sites where the pedestrian activity is influenced not only by the nearby school but also by other land uses. Radii of 100, 200, 400, and 800 meters, and percentages from 0% to 100% in steps of 5%, were tested to obtain optimal classification.

6.5.1 Multinomial logistic regression

Table 6-3 presents the MLR coefficients and odds ratios (ORs) of the best model obtained from the simulation approach described in the previous section. The OR is derived from the natural logarithm of the coefficient and should be interpreted as a comparison with the reference factor group. For instance, the OR of 4.181 for “Schools” in Factor Group #3 indicates that, while holding the other variables constant, a one-unit increase in the “Schools” variable makes the odds of the Factor Group being #3, on average, 4.181 times higher compared to Factor Group #1.

Table 6-3 MLR coefficients and odds ratio

Variable	Radius (m)	Factor group (reference: #1: Multipurpose)						VIF
		#2: Commuter		#3: Primarily school trips		#4: Discretionary multipurpose		
		Coeff.	OR	Coeff.	OR	Coeff.	OR	
Intercept	-	-0.597	-	1.048	-	2.389	-	-
Commercial area	400	-0.009	0.991	-0.091	0.913	-0.059	0.943	1.05
Industrial area	400	0.036	1.037	-0.065	0.937	0.087	1.091	1.05
Schools	400	1.034	2.812	1.431	4.181	-1.384	0.251	1.04
Transit stops	100	-0.068	0.934	0.091	1.095	-1.311	0.269	1.04
Accuracy of classification		69.3%						
McFadden pseudo-R2		0.288						
Sample size		75						

Statistical significance: bold for p-value < 0.05

The overall accuracy of classification for the MLR using the entire dataset was 69.3%. In a 5-fold cross-validation procedure run 100 times, the average training accuracy was 67.6%, while the average cross-validation accuracy was 57.2%. This suggests that the model is not overfitting. In a similar study, Medury et al. (2019) achieved an overall accuracy of 67% when classifying four factor groups using MLR. Furthermore, our model presented low VIF values for all variables, suggesting the absence of multicollinearity.

The selected model is simple, involving only four explanatory variables. We were able to generate more complex models with an accuracy of classification around 85%; however, these models exhibited overfitting and stability of coefficient issues, likely due to the relatively small sample size used for calibration.

The interpretation of the model coefficients shows interesting findings. The variable “Commercial area” exhibits a negative coefficient for Factor Groups #2 and #3, indicating that Factor Group #1 is linked to commercial activity. This matches expectations, given the characteristic of consistent pedestrian volume throughout the week in Factor Group #1 and reduced volumes on weekends in Factor Groups #2 and #3. However, the presence of a statistically significant and negative coefficient for Factor Group #4 lacks clarity from a phenomenological perspective. Factor Groups #1 and #4 share the feature of having relatively consistent volume throughout the week but are different in the seasonality aspect. As a result, it would be expected that Factor Group #4 is also associated with a certain level of commercial activity. A potential approach for addressing this could involve segregating commercial land use to distinguish establishments based on their “attendance patterns”. For instance, differentiating between grocery stores, which likely maintain a more constant activity throughout the year (i.e., Factor Group #1), and pubs or restaurants, which might experience more seasonal variations (i.e., Factor Group #4). Increasing the precision in categorizing land use could potentially lead to a more comprehensible interpretation of the outcomes, but it requires more disaggregated and sophisticated land use databases. Humagain & Singleton (2021) also reported a significant negative coefficient for “Commercial area”

associated with their monthly factor group that is affected by seasonality, though they did not provide a detailed interpretation.

The positive coefficient associated with the variable “Industrial area” for Factor Group #2 is expected, although it is not statistically significant at a 95% degree of confidence. This expectation arises from the decreased pedestrian activity on weekends. Meanwhile, the coefficient for Factor Group #4 may be interpreted as indicative of commuting seasonality: individuals may choose to take public transportation and/or walk to work in months unaffected by severe cold, opting for driving or carpooling during winter months. This observation aligns with the findings of Humagain & Singleton (2021), who found a similar outcome in their monthly factor group affected by seasonality, showing a correlation with increased employment density.

The presence of schools indicates association with Factor Groups #2 and #3, with a stronger relation to Factor Group #3, as anticipated. The positive and significant coefficient for Factor Group #2 can be understood as intersections with nearby schools in which the pedestrian activity is not mainly dictated by the school to the extent of observing substantial changes during school holidays. This may be linked to either or both the type and size of school (e.g., large high schools are expected to generate more pedestrian activity than smaller primary schools) and the presence of other land uses that generate pedestrian trips.

The negative and significant coefficient of the variable “Transit stops” for Factor Group #4 indicates that sites with nearby transit stops are associated with more consistent traffic patterns throughout the year. This aligns with expectations, considering that transit stops are part of the means of transportation. Thus, it is expected that the “attendance pattern” generated by a transit stop would be steadier than that of a commercial land use, for example.

Table 6-4 presents the confusion matrix for the MLR. “Observed factor group” in this table is the factor group assigned by the temporal indicator threshold method. The information in parenthesis represents the value of the cell divided by the total number of sites

observed in each factor group. Overall, the classification errors are evenly distributed across the matrix, except for one cell: among the 24 sites classified by the temporal indicator threshold method in Factor Group #2, 8 were incorrectly predicted to belong to Factor Group #1. Further examination of these sites revealed that half of them have WWI* values falling between 1.50 and 1.60. The WWI* threshold distinguishing groups #1 and #2 is set at 1.50 (Figure 6-5). This indicates that these particular sites are in a transitional area, sharing characteristics of both groups. Consequently, it is not expected that using expansion factors from Factor Group #1 instead of Factor Group #2 would significantly degrade the accuracy of the estimated AADPT for these sites.

Table 6-4 Confusion matrix for MLR approach

		Predicted factor group				Total
		1	2	3	4	
Observed factor group	1	21 (75%)	2 (7%)	4 (14%)	1 (4%)	28
	2	8 (33%)	13 (54%)	3 (13%)	0 (0%)	24
	3	1 (6%)	1 (6%)	14 (88%)	0 (0%)	16
	4	2 (29%)	1 (14%)	0 (0%)	4 (57%)	7
	Total	32	17	21	5	75

6.5.2 Logistic regression and school-based approaches

The calibration of Logistic Regression (LR) was conducted based on two categories: a) a combination of Factor Groups #1, #2, and #4 (called Non-School Trips), and a) Factor Group #3 (Primarily School Trips). In Table 6-5, LR coefficients and ORs are presented. The overall accuracy of classification using the entire dataset reached 92.0%. In the 5-fold cross-validation procedure executed 100 times, the average training accuracy was 91.4%, while the average cross-validation accuracy was 84.7%. These results suggest that the model is not

overfitting. Additionally, the greatest VIF value of 3.86 is relatively far from the value of 5 normally considered for flagging multicollinearity.

Variables such as “Commercial area” and “Schools”, which were statistically significant for factor group #3 in the MLR (Table 6-3), are also observed to be significant in the LR. The simplicity of the model (i.e., two dependent variables instead of four) allowed the inclusion of additional explanatory variables without causing overfitting issues. The variables are “Employment density”, “Household size”, and “Places of worship”.

Table 6-5 Logistic regression coefficients and odds ratios for factor group #3 (Primarily School Trips)

Variable	Radius (m)	Coeff.	OR	VIF
Intercept	-	-9.820	-	-
Commercial area	200	-0.154	0.857	3.86
Employment density	100	0.173	1.189	2.26
Household size	800	4.430	83.923	3.07
Places of worship	200	0.801	2.227	2.33
Schools	800	0.806	2.239	2.99
Accuracy of classification		92.0%		
McFadden pseudo-R ²		0.486		
Sample size		75		

Statistical significance: bold for $p < 0.05$

Reference group: Non-School Trips (Factor Groups #1 + #2 + #4)

In the context of the school-based approaches, the simplest approach that relies solely on the presence of schools within a given radius demonstrated that the variable "Schools – all types" within a 400-meter radius yielded the highest classification accuracy at 72.0%. Concerning the school-based approach that also considers the percentage of commercial land use, the combination of variables "Schools – all types" and "Commercial area" within the same 400-meter radius and with a threshold set at 50% provided the highest accuracy of 84.0%. This means that for a site to be classified into Factor Group #3, it must have at least

one school within a 400-meter radius, and the percentage of commercial land use within the same radius must be below 50%.

Table 6-6 presents the confusion matrices for the LR and school-based approaches. Overall, all three approaches exhibit similar classification for the sites of factor group #3, with the school-based approaches slightly outperforming the LR (13 versus 12). Conversely, the LR demonstrates superior performance in correctly classifying sites in the Combined Factor Group (57 vs. 50 vs. 41). Comparing the two school-based approaches, the 13 sites correctly identified as Factor Group #3 are identical in both approaches. Additionally, it is observed that the inclusion of the commercial land use constraint was beneficial for accurately classifying sites in the Combined Factor Group (50 vs. 41). This suggests that there are indeed sites with nearby schools whose pedestrian activity is not primarily influenced by the schools.

Table 6-6 Confusion matrices for Logistic regression and school-based approaches

		Approach / Predicted factor groups						Total
		LR		Schools and Commercial		Schools		
		1+2+4	3	1+2+4	3	1+2+4	3	
Observed factor groups	1+2+4	57 (97%)	2 (3%)	50 (85%)	9 (15%)	41 (69%)	18 (31%)	59
	3	4 (25%)	12 (75%)	3 (19%)	13 (81%)	3 (19%)	13 (81%)	16
	Total	61	14	53	22	44	31	75

Factor Group 3 = “Primarily School Trips”

Factor Groups 1+2+4 = “Non-School Trips”

6.6 Assessing the accuracy of expanding STCs to AADPT

Inspired by the methodology introduced by Medury et al. (2019), five different methods for assigning sites to factor groups were considered in the assessment of the expansion of the STCs to AADPT:

1. From STC Site Method: in this method, both the STC and the expansion factor originate from the same site. This represents the “best expansion possible” considering the use of DOWOM factors. In this case, every site represents one factor group. This method cannot be used in practice because expansion factors cannot be computed for STC sites.
2. Single Factor Method: this method employs a single expansion factor for each DOWOM, determined by the average of expansion factors across all sites. It represents the “worst expansion possible”, grouping all sites into a single factor group.
3. Empirical Factor Group Method: sites are categorized into factor groups based on their temporal indicator threshold classification. Although this method cannot be used in practice (because clustering on the basis of temporal indicators can only be done for CC sites not STC sites), this method eliminates modeling error, allowing for the identification of sources of error such as combining sites into factor groups and assigning STC sites to factor groups via models.
4. Predicted factor group – maximum probability: MLR and LR outcomes provide probabilities of a site belonging to a specific factor group. This approach assumes that the DOWOM expansion factor is selected from the factor group with the highest probability.
5. Predicted factor group – weighted probability: unlike the previous approach, this method does not directly assign an STC site to a particular factor group. Instead, it calculates the DOWOM expansion factor based on the weighted average of the DOWOM expansion factors from each factor group where the weights are the probabilities derived from MLR and LR outcomes.

Table 6-7 summarizes the performance of expanding STCs to AADPT by considering the various approaches (i.e., MLR, LR, and school-based) and site assignment methods listed above. The MAPE for each approach was derived using a process similar to the one employed for the initial determination of the appropriate type of expansion factor (Section 6.3). The results are categorized based on the strategy type (i.e., four or two factor groups) and on the potential days for the collection of STCs (i.e., all days or candidate days).

Additionally, the MAPE values for certain approaches are shown as the difference from the Single Factor Method. For instance, an overall value of -2.6 for the Empirical Factor Group Method using candidate days indicates that the MAPE for this method is 17.6% (calculated as 20.2% minus 2.6%). Figure 6-6 displays the DOWOM expansion factors for each factor group and strategy type, with days considered as candidates highlighted in pink. The expansion factors for the Single Factor Method are also included. The highlights from Table 6-7 and Figure 6-6 are presented and discussed below:

1. Candidate Days versus All Days: consistently improved performance is noted when restricting the analysis to Candidate Days. This is a well-established finding, as it aligns with the understanding that pedestrian activity on typical days exhibits less variability compared to considering All Days. For example, pedestrian volume on Tuesdays in April tends to be more stable than on Saturdays in December.
2. Expansion factor from STC site method: the MAPEs observed of 13.2% (Candidate Days) and 16.8% (All Days) represent the best possible performance when using DOWOM factors. This error corresponds to the variability of the STC within each DOWOM.

Table 6-7 Expanding STC to AADPT: summary of results

Type of days	Strategy	Obs. factor group	# sites	MAPE		Performance relative to Single Factor Method (MAPE i - MAPE Single Factor)						
				EF from STC site	Single Factor Method	Emp. factor group	MLR		LR		Schools & Comm.	Schools
							Max prob	Weighted prob	Max prob	Weighted prob		
Candidate days	4 factor groups	Overall	75	13.2	20.2	-2.6	-1.2	-1.2	-	-	-	-
		1	28	12.9	17.7	0.0	0.2	-0.5	-	-	-	-
		2	24	11.6	18.0	-1.9	1.6	0.9	-	-	-	-
		3	16	13.7	26.7	-8.9	-7.9	-5.3	-	-	-	-
		4	7	18.6	23.4	-1.0	-0.9	-1.2	-	-	-	-
	2 factor groups	Overall	75	13.2	20.2	-1.4	-	-	-0.8	-1.0	-1.0	-1.0
		1+2+4	59	13.0	18.5	0.6	-	-	0.6	0.1	0.7	0.7
		3	16	13.7	26.7	-8.9	-	-	-5.8	-5.4	-7.3	-7.3
All days	4 factor groups	Overall	75	16.8	26.9	-4.3	-1.7	-2.0	-	-	-	-
		1	28	15.0	24.6	-4.5	-1.6	-1.9	-	-	-	-
		2	24	15.9	23.4	-1.8	1.8	0.3	-	-	-	-
		3	16	18.9	32.9	-5.9	-5.4	-4.4	-	-	-	-
		4	7	22.0	34.8	-7.9	-6.0	-4.9	-	-	-	-
	2 factor groups	Overall	75	16.8	26.9	-2.3	-	-	-1.7	-1.6	-1.0	0.0
		1+2+4	59	16.2	25.3	-1.4	-	-	-1.0	-0.9	0.3	1.7
		3	16	18.9	32.9	-5.9	-	-	-4.1	-4.0	-6.1	-6.1

Note: Obs. = observed; EF = expansion factor; Emp. = empirical. Green and red shadings indicate that a particular approach is better or worse, respectively, than the Single Factor Method

3. Single Factor Method versus Empirical Factor Group Method:
- a) Overall, employing the Empirical Factor Group Method yields somewhat modest improvements compared to the Single Factor Method. For instance, when considering Candidate Days and employing four factor groups, MAPE for the Single Factor Method (worst performance) is 20.2% and MAPE when the expansion factor is computed from the STC site (best performance) is 13.2%. Thus, at best there is an opportunity to improve on the Single Factor Method by 7% MAPE. Employing the Empirical Factor Group Method to group the sites and utilizing the expansion factor from the correct site only reduces the Single Factor Method MAPE by 2.6% (a 12.9% reduction in error). This indicates that the remaining 4.4% of the MAPE is attributed to the variation within sites of the same factor group.
 - b) Candidate Days and four factor groups: Factor Group #3 exhibits the highest sensitivity to correct classification, as evidenced by the substantial error associated with the Single Factor Method and the notable improvement when applying the Empirical Factor Group Method. This sensitivity motivated the alternative proposal of having only two factor groups.
 - c) Clustering Factor Groups #1, #2, and #4 into a single factor group worsened the Empirical Factor Group Method results (increased the magnitude of MAPE by 1.2% for Candidate Days and 2.0% for All Days).
 - d) Candidate Days and two factor groups: it was unexpected that the Empirical Factor Group Method produced a poorer result than the Single Factor Method for the Combined Factor Group. A detailed exploration revealed two causes contributing to this result: a) a limitation in how the expansion factors are calculated. Expansion factors are computed using Equation 2.5, where for each site, AADPT is divided by the average of the STCs for a given day of the week and month. This formulation does not guarantee minimal averaged errors when expanding STCs to AADPT but instead results in a more even distribution of errors across observations; and b) the expansion factors derived from the Single Factor Method consistently exhibited lower values (averaging 8% lower) compared to the factors obtained from the Combined Factor Group (Figure 6-6b). These lower expansion factors improved the performance

of some STCs – particularly those associated with larger volumes, which are likely to exhibit a greater magnitude of error – resulting in lower averaged errors. In other words, the misclassification of sites from the Combined Factor Group into #3 accidentally improved the overall expansion performance.

- To illustrate this, consider a hypothetical scenario where a site with an AADPT of 300 has two STCs: one with a volume of 100 and another with a volume of 200. The average for the STCs is $(100 + 200) / 2 = 150$. Thus, an expansion factor of $300 / 150 = 2.00$ is obtained. Applying these factors to each STC results in absolute errors of 100 for each STC (e.g., $100 \times 2.00 = 200$, then $300 - 200 = 100$), which translates to percent errors and MAPE of 33.3%. If the expansion factor is arbitrarily reduced to 1.80, we would expect a deterioration in the overall performance, but this does not occur. The STC with a volume of 100 gives an estimated AADPT of 180, resulting in absolute and percent errors of 120 and 40%, respectively. For the STC with a volume of 200, the estimated AADPT is 360 (absolute error = 60, percent error = 20%). The MAPE in this case is 30%, which is lower than the 33.3% obtained using the empirically calculated expansion factor. This illustrates how the formulation used to calculate the expansion factors does not guarantee minimal averaged errors but rather leads to a more even distribution of errors. Note that the same unexpected results would occur if different error metrics – such as mean absolute error or root mean squared error – were employed.
- To further support this finding, when considering all days, the Empirical Factor Group Method outperforms the Single Factor Method.

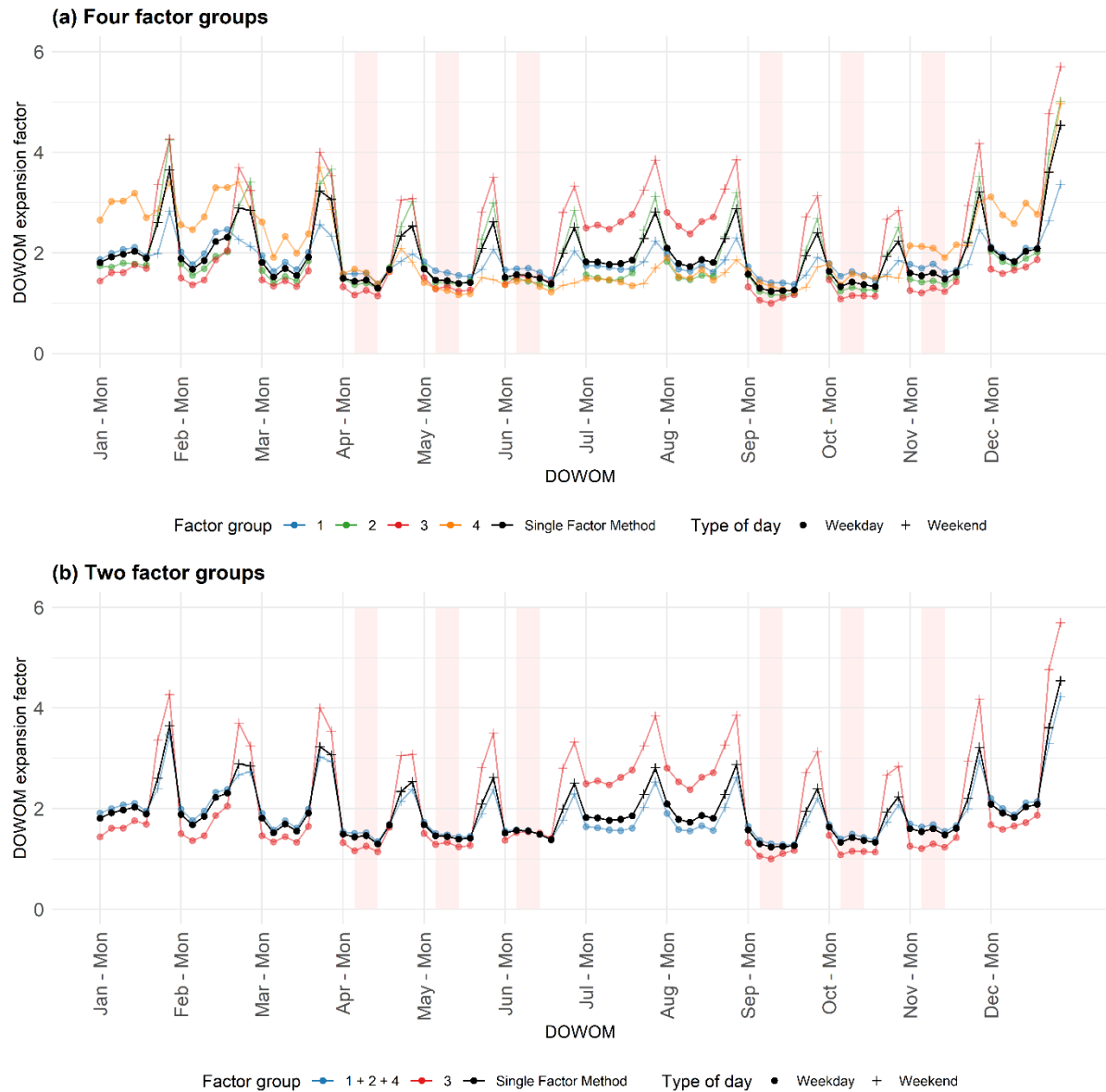


Figure 6-6 DOWOM expansion factors per factor groups

Note: the pink background highlights the candidate STC days.

4. MLR Method:

- a) Using MLR Method to identify the factor group provides a 1.2% reduction in the MAPE compared to the Single Factor Method (Candidate Days). This improvement is

relatively modest when contrasted with the Single Factor Method. This aligns with the findings of Medury et al., 2019), who reported comparable magnitudes of improvement. However, the significance of these improvements becomes more apparent when contextualized within the "room for improvement" between the best and worst methods, which is calculated as 7% (20.2% minus 13.2%) for Candidate Days. In this context, a 1.2% improvement translates to 17% of the total potential improvement that could have been attained.

- b) Candidate Days: a comparison with the Empirical Factor Group Method, which essentially represents the outcomes of a perfect model, reveals that 1.4% of the MAPE (2.6% minus 1.2%) can be attributed to model misclassification (i.e., the MLR predicts the incorrect factor group).
- c) The poor performance of Factor Group #2, particularly for the Maximum Probability Method, can be linked to the relatively poor classification of sites within this group (Table 6-4).

5. LR Method and School-Based Approaches:

- a) The School-Based Approaches exhibit superior performance compared to LR Method for Factor Group #3. This is attributed to the fact that the School-Based Approaches demonstrate slightly more accurate predictions for sites within Factor Group #3 than those made by the LR Method (Table 6-6).
- b) In the context of using Candidate Days, for which expansion factor values are relatively similar across factors groups (Figure 6-6b), the misclassification of sites in the Combined Factor Group (Table 6-6) does not substantially degrade the performance of the School-Based Approaches, resulting in comparable performance to that of LR Method. Conversely, when considering All Days, the significance of the misclassification becomes more apparent. This is due to the more substantial difference in expansion factors between the Combined Factor Group and Factor Group #3 in this scenario (Figure 6-6b).

6. Overall evaluation:

- a) The Weighted Probability MLR Method typically exhibits slightly superior performance compared to the Maximum Probability MLR Method, a trend consistent with the findings of Medury et al. (2019).
- b) Results obtained from MLR and LR Methods demonstrate comparability in both Candidate Days and All Days. This suggests that developing a model focused on identifying sites associated with schools may be sufficiently effective when contrasted with a more complex approach considering all existing traffic patterns.
- c) In the context of using Candidate Days, the performance of the School-Based Approaches is comparable to the MLR and LR Methods. As discussed, this happens because the misclassification of sites from the Combined Factor Group does not degrade model accuracy when only Candidate Days are considered. Conversely, the performance of the school-based approaches is degraded when using All Days.

6.7 Conclusions and practical recommendations

This Chapter examined the expansion of pedestrian STCs to AADPT at intersections. It followed a procedure that involved the identification of factor groups, the calibration of models to associate sites where long-term volumes are not available with factor groups, and the measurement of the error related to the expansion considering approaches with different levels of complexity.

Five factor groups were visually identified, and then three temporal indicators were proposed to characterize them: WWI*, SHTI, and DAM. The use of thresholds for these indicators allowed for an adequate representation of the factor groups. Further research is required to assess the spatial transferability of these indicators and thresholds. Additionally, none of the sites considered in this work presented a primarily recreational pattern (i.e., pedestrian volume on weekends is greater than on weekdays). So, it is possible – and likely –

that extra factor groups representing pedestrian volumes predominantly associated with recreational trips, may exist in the Region of Waterloo and in other jurisdictions.

A MLR was developed to link factor groups to land use, socioeconomic, and operational attributes. Explanatory variables that represent commercial area, industrial area, number of schools, and number of transit stops were found to be significant. The MLR provided a reasonable accuracy of classification of 69.3%. Although this performance is consistent with magnitudes reported in the literature, we believe that an improved model could have been developed if a larger sample size was available, especially for Factor Groups #4 and #5.

In addition to the MLR, three simpler approaches were also proposed with the focus of identifying sites where pedestrian activity is highly influenced by the presence of schools (i.e., Factor Group #3). A LR and two school-based approaches were developed and demonstrated good accuracy of classification: a) overall: 72-92%; and b) Factor Group #3: 75-81%.

The assessment of the expansion of STCs to AADPT considered other methods (for reference) in addition to the modeling approaches aforementioned: a) Single Factor Method: averaged expansion factor across all sites, representing the “worst scenario”; b) Empirical Factor Group Method: averaged expansion factor across sites that belong to a given factor group; and c) From STC site: both STC and expansion factor originate from the same site, representing the “best scenario”. Overall, modest improvements to the MAPE in the order of 0.8-1.2% were obtained when using the modeling approaches in comparison to the Single Factor Method when considering Candidate Days. The significance of these improvements is more noticeable when contextualized within the "room for improvement" between the best and worst methods, which is calculated as a MAPE of 7%. In this regard, the improvements are in the magnitude of 11-17%.

More importantly, different magnitudes of error were observed across factor groups, with Factor Group #3 (sites for which pedestrian volumes are primarily associated with trips

to/from school) being the one with the highest error when considering Candidate Days. This means that the simple Single Factor Method provided reasonable results for most of the factor groups, suggesting that – if data and/or personnel resources are constrained – practitioners may concentrate their efforts on identifying the sites impacted by the presence of schools.

We make the following recommendations for practitioners who need to expand pedestrian STCs to AADPT:

1. As established in the literature, this work reinforces the use of STCs from days with typical pedestrian activity rather than STCs from any day throughout the year.
2. If a jurisdiction has sufficient CC stations (i.e., ten or more stations per factor group) and qualified personnel to develop statistical models, it is recommended to develop a MLR, as this approach has been demonstrated to provide the best results.
3. If a jurisdiction lacks sufficient CC stations for developing a MLR (i.e., less than ten stations per factor group), it is suggested to apply one of the proposed simpler approaches. This is possible if CCs are deployed at sites that are impacted by schools.
 - a. If STCs from All Days are used, the LR Method is recommended.
 - b. If STCs from Candidate Days are used, the School-Based Approaches are recommended as they provide comparable accuracy and are much simpler to apply.
4. If a jurisdiction lacks any CC stations but is willing to install the minimum number possible, it is recommended to strategically deploy CC stations in at least five sites (Nordback et al., 2019) of two distinct groups based on the degree of pedestrian activity impacted by schools (i.e., highly impacted and not impacted). This should be sufficient for employing the School-Based Approaches if STCs from Candidate Days are used.

Chapter 7

Spatial Transferability of Expansion Factors for Estimating Pedestrian Volume at Intersections

This Chapter is based on the following journal paper:

Sobreira, LTP; Hellinga, B. Spatial Transferability of Expansion Factors for Estimating Pedestrian Volume at Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, under review.

7.1 Overall context

This Chapter explores the spatial transferability of expansion factors for estimating pedestrian exposure at intersections. More specifically, this work aims to answer the following practical questions a practitioner would face when attempting to estimate pedestrian exposure from short term counts and without the ability to compute expansion factors for the study jurisdiction:

1. Does the application of expansion factors spatially transferred from another jurisdiction provide sufficiently accurate AADPT estimates to be of practical value?
2. Does transferring models for associating STC sites with factor groups provide improved AADPT estimation accuracy that is of practical benefit as compared to using a single average factor group when transferring expansion factors?
3. How does the accuracy of spatially transferring expansion factors compare to that of alternative methods (e.g., crowd-sourced data, Large Language Models, and direct-demand models) when continuous monitoring is not available?

The next section introduces the jurisdictions considered in this study. This is followed by a description of the Chapter's framework and organization.

7.2 Study sites

This Chapter was developed considering four jurisdictions. Three of them are located in Ontario, Canada: the Region of Waterloo (including the cities of Cambridge, Kitchener, and Waterloo), the City of Toronto, and the City of Milton. These jurisdictions are within a similar range of latitude and share comparable weather conditions: four distinct seasons, including cold winters (average daytime temperatures $\leq 0^{\circ}\text{C}$ (32°F) from December through February) and hot summers (average daytime temperatures $\geq 24^{\circ}\text{C}$ (75°F) from June through August) (Weather Spark, 2023). Additionally, these jurisdictions have similar school holiday periods in July and August. The fourth jurisdiction is Pima County, Arizona, USA, which experiences different weather conditions compared to the Ontario jurisdictions (i.e., milder winters and hotter summers) and has distinct school holiday periods in June and July.

Given the differences among the jurisdictions, it is expected that expansion factors spatially transferred across the jurisdictions in Ontario will perform better than those from Pima County. We acknowledge that, from a practical perspective, one would not typically consider transferring expansion factors from Pima County to Toronto, for example. However, we included Pima County to represent a scenario where poor transferability is likely to be observed.

Table 7-1 presents the number of sites and descriptive statistics of the AADPT in each jurisdiction. The average AADPT highlights the differences in pedestrian activity between jurisdictions. Pima County and Toronto exhibit very low and very high pedestrian activity, respectively. This variation is explained by the site locations used in this work: none of the sites from Pima County are in the City of Tucson (the most populated city in the county), while 19 out of the 34 sites in Toronto are in the Downtown area, which has very high pedestrian activity. Milton and Waterloo present mid-levels of AADPT.

Table 7-1 Descriptive statistics of AADPT

Jurisdiction	#Sites	AADPT			
		Average	Standard Deviation	Minimum	Maximum
Waterloo	101	852	1,152	16	6,146
Toronto	34	14,282	10,557	1,283	33,629
Milton	33	350	265	26	1,124
Pima County	20	52	50	16	202

Note: AADPT = Annual Average Daily Pedestrian Traffic

Expansion factors are calculated as the ratio between AADPT and short-term volumes (Equation 2.5). Therefore, differences in overall pedestrian activity levels across jurisdictions do not necessarily imply that expansion factors are unsuitable for transferability. If jurisdictions exhibit similar temporal patterns, transferred expansion factors may still yield reasonable accuracy, even if pedestrian activity levels differ significantly. This is why Ontario jurisdictions may have transferable expansion factors: as they share similar weather and school holidays, they may also share similar temporal patterns.

This work considered sites equipped with camera-based CC stations, all located at signalized intersections, with the analysis period spanning from April 01, 2023, to March 31, 2024. Filters were applied to ensure that CC stations were collecting data properly and to remove outliers:

1. Days of proper functioning: at least 72 out of 96 15-minute count intervals had to register at least one motorized vehicle, and the total daily pedestrian volume had to be greater than 0.
2. Outlier removal: for each month and each site, a filter was applied to the pedestrian counts to remove outliers. This involved excluding daily (24-hour) pedestrian volumes that were either a) above the 3rd quartile plus 1.5 times the interquartile range, or b) below the 1st quartile minus 1.5 times the interquartile range.

After applying the filters, the average number of days available for each site was 359, with a minimum of 242 days. Of the 188 sites, only 14 had fewer than 350 days of data available.

7.3 Framework

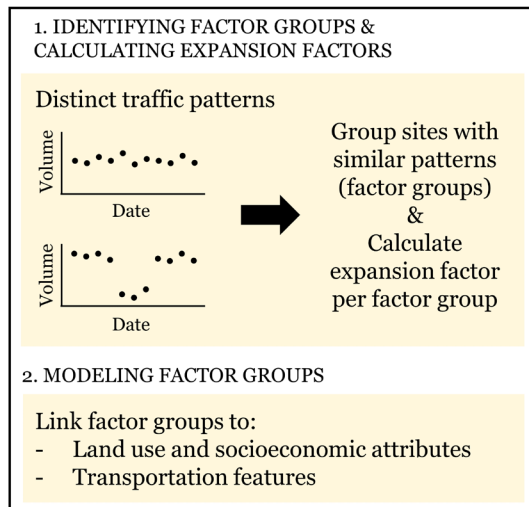
Before presenting the framework, two key terms are defined: base jurisdiction and target jurisdiction. The base jurisdiction refers to the jurisdiction where CCs are available. Expansion factors are calculated in this jurisdiction and then transferred to the target jurisdiction. In the target jurisdiction, it is assumed that only STC sites are available.

Figure 7-1 displays the proposed framework for spatially transferring expansion factors. The framework involves four major steps: two in the base jurisdiction and two in the target jurisdiction. The first step is the identification of factor groups and the calculation of expansion factors in the base jurisdiction. The second step consists of modeling factor groups with the goal of allowing the association of STC sites with a given factor group. Moving to the target jurisdiction, the third step involves assigning STC sites with a given factor group using the models developed in the second step. In the final step, the STCs in the target jurisdiction are expanded to AADPT using the expansion factors retrieved from the base jurisdiction and the associated factor group.

Note that Steps 1, 2, and 3 are strictly related to Practical Question 2, which examines whether transferring models to associate STCs with factor groups provides practical benefits. The application of single expansion factors averaged across all sites (i.e., one unique factor group) does not require these steps. To emphasize the discussion of the three practical questions while reducing the complexity and density of this Chapter, we have presented the methods and results for Steps 1, 2, and 3 in the appendix while including an overview in the main body. Step 4 represents the most critical portion of this study and is linked to all three practical questions; therefore, its methods and results are presented in the body of the document.

BASE JURISDICTION

Sites with continuous counts



TARGET JURISDICTION

Sites with short-term counts

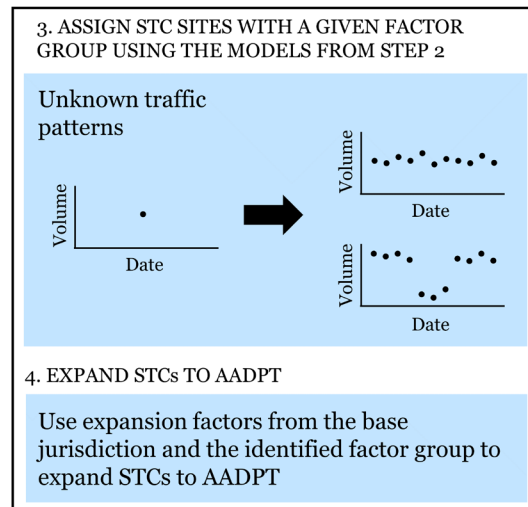


Figure 7-1 Framework

The remainder of this document is organized as follows: The next section introduces the types of expansion factors employed, followed by an overview of Steps 1, 2, and 3. Subsequently, the methods and results for Step 4 are presented. Finally, the document concludes with a discussion of the three practical questions, along with conclusions and practical recommendations. Similarly to the previous Chapter, DOWOM expansion factors were employed.

7.4 Steps 1, 2, and 3: overview

Step 1 involves identifying factor groups at CC sites within the base jurisdiction. Three traffic indices were used to characterize the temporal patterns at each site, followed by a clustering technique to group sites with similar characteristics. Figure A1 illustrates the factor groups for each jurisdiction. This procedure is identical to the one used in the previous Chapter.

Similar factor groups were identified in Waterloo and Toronto: a) Factor Group #1: multipurpose sites with relatively constant pedestrian activity throughout the year, b) Factor Group #2: commuter sites with reduced volumes on weekends, and c) Factor Group #3: sites primarily affected by schools, showing reduced activity on weekends and during school-holiday months. Milton exhibited only Factor Groups #1 and #3. Pima County resulted in three factor groups, but with characteristics different from those in the Ontario jurisdictions.

Step 2 involves associating the factor groups identified in Step 1 with land use, socioeconomic, and transportation attributes. Due to small sample sizes and differences in pedestrian activity patterns compared to Ontario, this step was not applied to Pima County.

For the Ontario sites, four different models were considered. The first used Multinomial Logistic Regressions (MLRs) for cases with more than two factor groups. The remaining models focused on distinguishing Factor Group #3 (associated with schools) from the others. These included binary Logistic Regression (LR) and simple threshold-based models, which identified sites in Factor Group #3 based on the presence of schools and the combined presence of schools and non-dominant commercial land use. Due to a limited sample size, MLR and LR models were not developed for Toronto. Table A1 and Table A2 summarize the models developed for Step 2.

Step 3 switches to the target jurisdiction, where only STC sites are available. The models developed in Step 2 for the base jurisdiction are transferred and applied to the target jurisdiction to assign each site to a specific factor group. Transferability performance was inconsistent, with some models transferring well to certain jurisdictions and others performing poorly. Table A3 summarizes the transferability results.

Again, detailed explanations of the methods and results from Steps 1, 2, and 3 are provided in the appendix. The next section focuses on Step 4, which represents the actual application of the expansion factor method in the target jurisdiction.

7.5 Step 4: expanding STCs to AADPT in the target jurisdiction

7.5.1 Method

Five different methods were employed to expand STCs. The first method expands the 8-hour counts to daily volumes, while the others expand to AADPT. The first three methods are only applicable when the base and target jurisdictions are the same; these methods serve as a reference for comparing the spatial transferability considered in the remaining methods. The methods are described below:

1. **Ratio 24/8 Method:** this method applies when no CCs are available in the target jurisdiction, and no expansion factors can be transferred from another jurisdiction. To expand 8-hour STCs to daily (24-hour) volumes in this scenario, a simple solution is to collect one-day 24-hour volume data at a selected number of sites (k). This allows for the calculation of an average "Ratio 24/8" factor, which is then applied to STCs. We tested using $k = 5$ and $k = 10$ sites and found that both yielded similar results; therefore, we adopted $k = 5$ for the remainder of this paper. Due to the inherent variability in selecting the five sites and the days for calculating the ratio, the procedure was repeated 100 times, with the results presented as an average. Theoretically, this method represents the "worst-case scenario" for estimating daily pedestrian exposure.
2. **From STC Site Method:** in this method, both the STC and the expansion factor come from the same site, meaning that each site represents one factor group. This method cannot be used in practice because expansion factors can only be computed from CC sites and not from STC sites. It represents the "best-case scenario".
3. **Empirical Factor Group Method:** in this method, the observed (or true) factor group for each site, as obtained in Step 1, is used. Although this method cannot be applied in practice (since clustering based on temporal indicators is only possible for CC sites, not STC sites), it eliminates modeling error. This allows for the identification of sources of error, such as issues with combining sites into factor groups and assigning STC sites to factor groups via models.

4. Single Factor Method: this method uses a single expansion factor for each DOWOM, determined by the average of expansion factors across all sites (i.e., one unique factor group).
5. Modeling Methods: for the school-based models, a specific factor group is predicted for each site. In the case of MLR and LR models, probabilities of a site belonging to a specific group are available. Therefore, two strategies can be used: a) maximum probability: select the expansion factor from the factor group with the highest probability, and b) weighted probability: use an expansion factor calculated as the weighted average of the DOWOM expansion factors from each factor group, with weights derived from the probabilities provided by the MLR and LR model outcomes. (Medury et al., 2019) and (Sobreira & Hellenga, 2024) (Chapter 6) found that the weighted probability strategy performed slightly better; thus, it was adopted in this work.

This work uses sites with CCs due to the necessity of knowing the observed AADPT at each site to evaluate the quality of the STC expansion. Multiple STCs can be retrieved from each site, resulting in multiple predicted AADPT estimates. The performance of each individual prediction is assessed using the absolute percent error ($APE_{i,j}$) and then averaged across all observations j and sites i to obtain the mean absolute percent error (MAPE).

Results are presented based on 8-hour STCs obtained from all available days throughout the year (i.e., days when the CC station was functioning properly and were not excluded by the filtering criteria) and from candidate days. In this work, candidate days refer specifically to non-holiday Tuesdays, Wednesdays, and Thursdays during April, May, June, September, October, and November for the Ontario sites, and from January to May and August to December for Pima County. These months were selected to avoid periods of inclement weather and school holidays.

7.5.2 Results

It is important to note that the AADPT estimation accuracy when transferring expansion factors using the models to assign factor groups depends on two elements: a) the quality of factor group classification, shown in Table A3, and b) the similarity of expansion factors within the same factor group across different jurisdictions. In other words, while classification accuracy can influence the success of transferring expansion factors, the primary factor is the consistency of expansion factors across jurisdictions. This aspect is explored and discussed in more detail in the following sections.

Table 7-2 and Table 7-3 present the performance of the expansion methods described in the previous section for STCs collected on candidate days and all days, respectively. The results are categorized by overall performance (representing the average of all sites) and by factor group. Minor observations regarding the tables are provided below, with a discussion on the spatial transferability of expansion factors in the next section.

The first observation is the noticeable difference in the magnitude of error across different jurisdictions. When considering the most accurate method (From STC Site), the overall MAPE follows the order: Toronto << Waterloo < Milton << Pima County. This pattern correlates with the level of pedestrian activity in each jurisdiction (Table 7-1): AADPT Toronto >> Waterloo > Milton >> Pima County. The observed order of error magnitude is expected, given that we used a relative measure of error, which tends to yield higher errors at sites with low AADPT (the denominator in Equation 2.9). If we had used an absolute measure of error, such as the mean absolute error, we would likely observe an opposite pattern. This highlights the importance of the type of error measure employed, as it suggests that direct comparisons across different jurisdictions using these metrics alone may not be entirely fair.

The second observation concerns the methods used as references. When comparing the worst- and best-case scenarios (i.e., Ratio 24/8 and From STC Site, respectively), the Ratio 24/8 Method, on average, produces an error 2.78 times greater than the From STC Site Method for candidate days. Additionally, when expansion factors are grouped into factor

groups (Empirical Factor Group Method), the overall MAPE for candidate days increases by 8.42%, on average, in absolute terms compared to the From STC Site Method.

The third observation involves some unexpected results. For example, the School model outperforms the LR model (MAPE = 24.5% vs. 25.4%) for candidate days when Waterloo is both the base and target jurisdiction. This is surprising since the LR model provides a better factor group prediction (#1+2 = 93%; #3 = 62%) compared to the School model (#1+2 = 76%; #3 = 67%) (Table A3). This is a similar issue to the one discussed in Chapter 6 in Section 6.6 originated from the way expansion factors are calculated.

7.6 Discussion and conclusions

This section addresses the three practical questions regarding the spatial transferability of expansion factors posed at the end of the introduction. Additionally, it discusses further findings related to cases where the base and target jurisdictions are the same and concludes the work with a summary of key points and practical recommendations. To support the discussion, smaller tables derived from Table 7-2 and Table 7-3 are used, with a primary focus on STCs collected on candidate days.

Table 7-2 Accuracy of estimated AADPT (MAPE): STCs from candidate days only

Target jurisdiction	Observed factor group	#Sites	Reference methods			Factor group identification methods for each base jurisdiction												
			Ratio 24/8	From STC Site	Emp. Factor Group	Waterloo					Toronto			Milton				PC
						SF	MLR	LR	S + C	S	SF	S + C	S	SF	LR	S + C	S	SF
Waterloo	Overall	101	36.8	15.5	23.6	26.0	24.7	25.4	24.9	24.5	31.9	30.7	30.5	27.5	29.3	31.7	28.8	92.3
	#1: Multipurpose	48	27.8	15.4	23.8	21.9	21.9	23.8	24.4	23.8	24.3	28.6	27.8	23.0	27.8	31.5	27.9	77.1
	#2: Commuter	32	39.4	13.9	22.2	25.3	24.7				33.5			27.3				101
	#3: Schools	21	53.5	18.4	25.5	36.5	31.0	31.4	26.6	28.7	46.8	38.8	40.8	38.0	34.8	32.3	32.3	113
Toronto	Overall	34	26.0	7.9	14.9	20.0	19.7	20.0	22.5	23.5	19.1	18.2	18.8	19.3	22.4	26.5	27.6	71.2
	#1: Multipurpose	21	15.2	7.3	13.3	18.3	17.0	18.8	21.6	23.4	13.2	17.3	17.8	17.8	21.1	26.8	28.2	57.2
	#2: Commuter	7	38.5	8.1	14.0	18.0	23.2				24.4			17.8				92.0
	#3: Schools	6	47.0	9.6	21.2	27.7	24.6	25.3	26.6	23.7	32.4	22.1	23.2	25.8	28.0	24.8	24.8	93.0
Milton	Overall	33	48.8	16.9	24.3	31.8	29.2	29.7	28.1	26.9	36.3	33.0	32.6	31.2	27.8	27.7	26.1	93.3
	#1: Multipurpose	18	29.9	18.2	24.6	23.7	25.1	24.8	24.1	23.3	23.2	24.0	23.4	23.1	23.1	25.1	24.1	70.1
	#3: Schools	15	71.3	15.3	23.9	41.5	34.2	35.6	32.8	31.3	51.8	43.9	43.6	40.8	33.3	30.7	28.5	121
Pima County	Overall	20	66.8	26.1	37.2	59.4	-	-	-	-	63.9	-	-	84.8	-	-	-	53.9
	Multipurpose	8	39.0	21.4	27.3	35.5	-	-	-	-	37.0	-	-	55.7	-	-	-	35.4
	S & S	5	119	28.8	48.2	97.3	-	-	-	-	108	-	-	133	-	-	-	106
	Seasonality	7	61.7	29.7	40.8	59.8	-	-	-	-	63.1	-	-	84.1	-	-	-	37.7

Note: STC = Short-term count; Emp. = Empirical; SF = Single Factor; MLR = Multinomial logistic regression; LR = Logistic regression; S + C = School & Commercial model; S = School model; PC = Pima County; S & S = School and seasonality

Table 7-3 Accuracy of estimated AADPT (MAPE): STCs from all days

Target jurisdiction	Observed factor group	#Sites	Reference methods			Factor group identification methods for each base jurisdiction												
			Ratio 24/8	From STC Site	Emp. Factor Group	Waterloo					Toronto			Milton				PC
						SF	MLR	LR	S + C	S	SF	S + C	S	SF	LR	S + C	S	SF
Waterloo	Overall	101	35.5	19.1	27.9	30.9	29.0	29.7	30.4	30.7	34.2	33.3	33.6	45.5	44.8	47.0	50.3	97.6
	#1: Multipurpose	48	29.5	18.2	26.8	29.9	28.0	28.8	30.0	30.3	30.6	31.9	32.0	45.9	43.8	46.5	50.8	101
	#2: Commuter	32	37.1	18.4	27.0	28.7	27.8				33.5			44.1				102
	#3: Schools	21	46.7	22.1	31.9	36.2	33.3	33.3	31.8	32.3	43.5	38.9	39.6	46.8	48.5	48.8	48.8	84.2
Toronto	Overall	34	23.7	10.6	16.8	23.3	21.8	22.4	26.3	29.8	20.6	19.2	22.3	35.1	32.4	40.4	54.1	91.2
	#1: Multipurpose	21	16.4	9.1	14.2	22.3	19.7	21.5	25.7	30.4	16.9	17.9	21.4	35.9	30.7	40.9	57.7	91.1
	#2: Commuter	7	30.7	11.9	16.9	22.5	23.7				21.5			32.5				100
	#3: Schools	6	39.4	14.4	25.1	27.7	26.6	26.4	28.9	26.9	32.0	25.1	26.2	35.2	39.7	38.2	38.2	81.3
Milton	Overall	33	43.7	22.6	32.5	37.5	36.0	36.4	35.4	33.8	38.6	35.9	35.9	41.1	36.3	36.0	41.5	97.9
	#1: Multipurpose	18	33.0	21.8	27.2	33.5	33.5	34.2	33.2	31.6	30.5	29.3	29.6	38.6	32.2	32.2	42.7	108
	#3: Schools	15	56.4	23.4	38.8	42.3	39.1	39.1	38.0	36.4	48.1	43.8	43.3	44.0	41.1	40.6	40.1	85.4
Pima County	Overall	20	64.5	28.9	45.9	61.3	-	-	-	-	64.9	-	-	81.7	-	-	-	68.0
	Multipurpose	8	41.4	22.3	29.9	40.1	-	-	-	-	41.3	-	-	58.4	-	-	-	68.5
	S & S	5	100	35.3	72.6	84.2	-	-	-	-	92.3	-	-	104	-	-	-	83.8
	Seasonality	7	65.5	32.0	45.3	69.3	-	-	-	-	72.3	-	-	92.6	-	-	-	56.1

Note: STC = Short-term count; Emp. = Empirical; SF = Single Factor; MLR = Multinomial logistic regression; LR = Logistic regression; S + C = School & Commercial model; S = School model; PC = Pima County; S & S = School and seasonality

7.6.1 Question 1: Does the application of expansion factors spatially transferred from another jurisdiction provide sufficiently accurate AADPT estimates to be of practical value?

Short answer: Yes, for jurisdictions with similar weather and school-holiday periods, transferring expansion factors across jurisdictions provided AADPT estimates of practical value. However, poor results were observed when transferring across jurisdictions with differing characteristics.

Table 7-2 and Table 7-3 demonstrate that applying expansion factors developed in Pima County to the Ontario jurisdictions resulted in poor performance (MAPE varied from 71.2% to 97.9%). This outcome is attributed to two factors: a) differences in school holiday months and weather seasonality between the regions, leading to distinct factor group configurations (Figure A1), and b) the significantly lower pedestrian activity observed in Pima County, which produced expansion factors of different magnitudes compared to those in Ontario. Given the available data, it is challenging to determine the relative influence of each factor. Therefore, it cannot be definitively stated that expansion factors from regions with differing school holiday schedules and weather seasonality are non-transferable.

Regarding the Ontario jurisdictions, Table 7-4 provides a summary of the spatial transferability using the Single Factor Method. The results are presented for two conditions: average across all sites (“Overall”) and just for sites in Factor Group #3: Schools. Column 6 (C6) displays the average MAPE when the base and target jurisdictions differ. For instance, the 29.7% recorded for Waterloo – Overall represents the average of 31.9% and 27.5%. C7a compares the values in C6 with the performance when the base and target jurisdictions are the same (i.e., the blue-highlighted cell in C5).

Overall, the spatial transferability of expansion factors across the Ontario jurisdictions provides good results. When comparing the spatially transferred factors to those obtained when the base and target jurisdictions are the same, there is an average absolute

increase in MAPE across the three jurisdictions of just 2.3%, which corresponds to a relative increase of 8.7% (Table 7-4). If we consider this increase in error as a "penalty" for not having continuous monitoring in the target jurisdiction, it seems to be a relatively minor cost, likely of little practical significance, for estimating AADPT without the need for additional resources to deploy CC stations. This becomes even more apparent when comparing the spatially transferred Single Factor Method with the Ratio 24/8 Method (C7b), which is probably the method a practitioner would use to cost-effectively convert STCs to daily volumes. On average, the Single Factor Method performs 9.4% better than the Ratio 24/8 Method in absolute terms.

Table 7-4 Summary of AADPT estimation accuracy using the Single Factor Method – candidate days

C1: Target jurisdiction	C2: Observed factor group	C3: #Sites	C4: MAPE Ratio 24/8	C5: MAPE			C6: Avg. MAPE: Base ≠ Target	C7: Impact of spatially transferring EFs			
				Base jurisdiction				C7a: C6 vs. C5 (blue)		C7b: C6 vs. C4	
								Absolute difference	Relative difference	Absolute difference	Relative difference
				Waterloo	Overall	101		36.8	26.0	31.9	27.5
	#3: Schools	21	53.5	36.5	46.8	38.0	42.4	+5.9	+16.2%	-11.1	-20.7%
Toronto	Overall	34	26.0	20.0	19.1	19.3	19.7	+0.5	+2.7%	-6.3	-24.4%
	#3: Schools	6	47.0	27.7	32.4	25.8	26.8	-5.7	-17.4%	-20.3	-43.1%
Milton	Overall	33	48.8	31.8	36.3	31.2	34.0	+2.8	+9.1%	-14.8	-30.3%
	#3: Schools	15	71.3	41.5	51.8	40.8	46.7	+5.9	+14.3%	-24.7	-34.6%
Average: Overall								+2.3	+8.7%	-9.4	-24.7%
Average: #3: Schools								+2.0	+4.4%	-18.7	-32.8%

Note: MAPE = Mean absolute percent error; Avg. = Average; EF = Expansion factor. Cells highlighted in blue represent the cases where base and target jurisdictions are the same.

Regarding Factor Group #3: Schools, a higher magnitude of error is observed both when the base and target jurisdictions are the same and when they are different. Considering the specific characteristics of the expansion factors for sites in this group, it is possible that

applying models – and thereby using more accurate expansion factors compared to the Single Factor Method – could improve the performance. This possibility is explored in the next section.

To provide better context for Table 7-4, Figure 7-2 presents the expansion factors for each Ontario jurisdiction using the Single Factor Method. The expansion factors are notably similar, particularly for candidate days (highlighted in the pink background), which explains the relatively good performance observed when transferring the factors. Additionally, Table 7-4 shows that the factors from Waterloo and Milton transferred well to other jurisdictions, while those from Toronto produced slightly higher errors. This discrepancy can be attributed to the fact that Toronto's expansion factors for candidate days are slightly higher than those from Waterloo and Milton (Figure 7-2), which generates the issue with the MAPE calculation discussed in the previous section. Furthermore, the factors from Milton, when considering all days, differ more and are greater in magnitude than those from Waterloo and Toronto. This difference is reflected in Table 7-3: for example, MAPE when Waterloo is the target using the Single Factor Method: a) base is Waterloo = 30.9%, b) base is Toronto = 34.2%, and c) base is Milton = 45.5%.

To consolidate the discussion between all days and candidate days, Table 7-5 compares the MAPE – in absolute terms – of the Single Factor Method for the two configurations. A greater magnitude of error is anticipated when considering all days instead of candidate days due to the higher variation across the days included in all days. For example, daily pedestrian volumes are expected to vary more across Saturdays in January than across Tuesdays in April. The results for the From STC Site Method in Table 7-2 and Table 7-3 confirm this. The spatial transferability between Waterloo and Toronto produced reasonable results when using all days (Table 7-5), given that the expansion factors for both jurisdictions are relatively similar throughout the year (Figure 7-2). However, a greater magnitude of error is observed when Milton is the base jurisdiction, and all days are used. The difference in expansion factors depicted in Figure 7-2 explains this. Overall, this

suggests that employing candidate days when spatially transferring expansion factors is recommended.

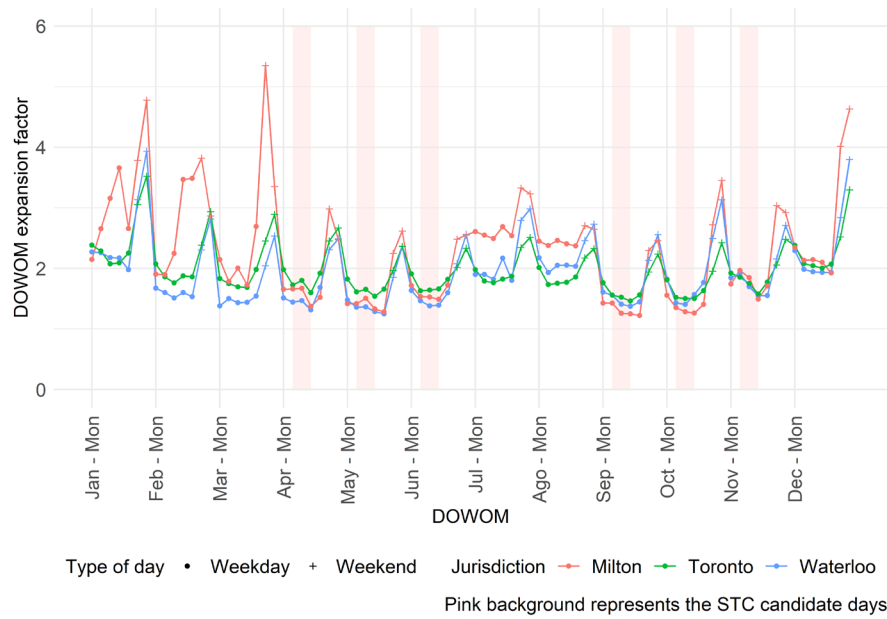


Figure 7-2 Single Factor Method expansion factors

Table 7-5 Change in AADPT estimation accuracy (MAPE) when using all days versus candidate days

Target jurisdiction	Base jurisdiction		
	MAPE Single Factor: All days – Candidate days		
	Waterloo	Toronto	Milton
Waterloo	+4.8	+2.3	+18.0
Toronto	+3.3	+1.5	+15.8
Milton	+5.7	+2.3	+9.9

7.6.2 Question 2: Does transferring models for associating STC sites with factor groups provide practical benefits over using a single average factor group when transferring expansion factors?

Short answer: The use of models yielded slightly better results than the single average factor group, but this trend was not consistent across all observations. Given the complexity of implementing this approach, the results do not suggest practical benefits.

The results discussed in the previous section highlighted that employing the Single Factor Method may lead to a higher magnitude of error for specific factor groups, such as those associated with school activity. A straightforward solution to this issue, when the base and target jurisdictions are the same, is to use models to assign a factor group to STC sites, thereby applying more accurate expansion factors for each group. However, when the base and target jurisdictions differ, these models must also be transferred from the base jurisdiction, introducing two sources of error: one related to factor group classification and the other to differences in expansion factors within the same factor group across jurisdictions. This section explores whether transferring these models to assign factor groups (Modeling Method) – and thus applying their associated expansion factors – provides practical benefits over using the Single Factor Method in the context of spatial transferability.

Figure 7-3 summarizes results from Table 7-2 and Table 7-3, comparing the performance of the Modeling Method with the Single Factor Method. The y-axis represents the difference in MAPE between the methods: if the value is above zero, the Single Factor Method performs better; if below zero, the Modeling Method is superior. The x-axis indicates the classification accuracy, as extracted from Table A3. Only scenarios where the base and target jurisdictions differ are included.

Figure 7-3a illustrates the overall performance and classification accuracy. It is observed that the differences in performance between all days and candidate days become

more pronounced as classification accuracy decreases. This outcome was expected, given that expansion factors for all days differ more across factor groups than those for candidate days. Furthermore, the graph includes 36 observations: 22 show improved performance with the Modeling Method, while 14 exhibit better results with the Single Factor Method. The 22 "improved" observations show an average absolute improvement in MAPE of 2.02%, whereas the cases where the Modeling Method generated worse results display an average worsening of 5.14%.

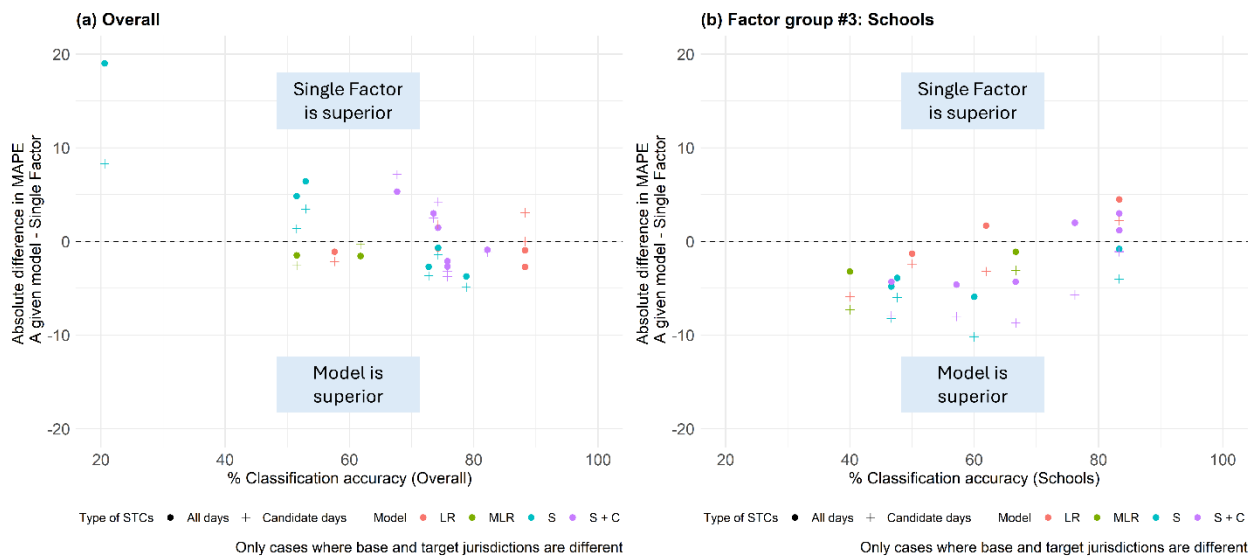


Figure 7-3 Single Factor Method versus Modeling Methods

Note: STC = Short-term count; MAPE = Mean absolute percent error; MLR = Multinomial logistic regression; LR = Logistic regression; S + C = School & Commercial; S = Schools

Considering Factor Group #3: Schools (Figure 7-3b), the Modeling Method showed improved performance in 28 observations, with an average absolute MAPE improvement of 4.53%. In contrast, the Single Factor Method was superior in 8 observations, where applying the model led to an average absolute MAPE increase of 2.45%. Compared to the overall results, the Modeling Method not only outperforms the Single Factor Method more

frequently but also with greater impact. This may be of particular interest to practitioners, given the importance of sites located near schools in terms of road safety.

In conclusion, while the Modeling Method generally tends to produce slightly better results than the Single Factor Method, particularly for sites related to school activities, we do not recommend its use for spatially transferring expansion factors. This recommendation is based on two reasons: a) the risk that the model could degrade the transferability performance to a greater extent than the potential benefits it might offer, and b) the added complexity of developing statistical models in the base jurisdiction.

7.6.3 Question 3: How does the accuracy of spatially transferring expansion factors compare to that of alternative methods when continuous monitoring is not available?

Short answer: Spatially transferred expansion factors performed significantly better than the alternative methods evaluated.

The spatial transferability of expansion factors using the Single Factor Method (using candidate days) is compared to three alternative methods described (Table 7-6): a) crowd-sourced data (StreetLight, 2023), Large Language Models and satellite imagery (Sobreira & Hellinga, 2025) – the topic of Chapter 5, and the local calibration of DD models (Sobreira & Hellinga, 2023b) – the topic of Chapter 4.

As demonstrated earlier in this work, the level of pedestrian activity at each site and within each jurisdiction directly influences the magnitude of the observed error. To address this and allow direct comparisons across different jurisdictions, the results are presented by volume category. The volume categories used were proposed by StreetLight (StreetLight, 2023). Only the spatial transferability to Waterloo is presented in Table 7-6 because all sites

in Toronto fall into the “Very high” category, and some categories in Milton lack a representative sample size. The error metric E (Equation 2.11) is also included in Table 7-6 .

Table 7-6 Single Factor Method compared to alternative methods

Accuracy metric	Volume category	Single Factor Method: Target jurisdiction = Waterloo			Alternative methods (applied to other jurisdictions)		
		#Sites	Base jurisdiction		Crowd sourced data	Large Language Models	Local calibration of DD models
			Toronto	Milton			
MAPE (%)	Low (< 150)	23	41.9	36.8	151	273	-
	Mid (150-500)	37	31.9	27.1	89	51	-
	High (500-1,000)	14	25.1	21.0	82	51	-
	Very high (>1,000)	27	26.9	23.4	71	49	-
E = MAE / $\bar{Y}$ Overall		101	0.275	0.247	-	0.50 - 0.75	0.52 - 0.71

The results in Table 7-6 suggest that the spatial transferability of expansion factors using the Single Factor Method is significantly superior to that of using crowd-sourced data, the ChatGPT approach, and the local calibration of spatially transferred DD models. This highlights the potential of the proposed method, particularly since it relies only on the availability of STCs with pedestrian volumes – commonly obtained through traffic signal design studies – and expansion factors from candidate days in a base jurisdiction with similar school-holiday months and weather seasonality.

A few considerations regarding the results in Table 7-6: a) our work used AADPT as the target measure, while the crowd source data and Large Language Model methods used Monthly Average Daily Pedestrian Traffic (MADPT). The omission of month-to-month seasonality in the MADPT measure might slightly improve the accuracy of these methods compared to if AADPT had been used. b) In our work, expansion factors and STCs are from the same period. In practical applications, it is common for some sites to only have STCs

from previous years, which can introduce a temporal element to the counts and potentially lead to a slightly higher magnitude of error.

7.6.4 Additional findings

The additional findings are related to cases where the base and target jurisdictions are the same. Table 7-7 summarizes the results of the Ratio 24/8, Empirical Factor Group, and Single Factor Methods in terms of overall performance and specifically for Factor Group #3: Schools on candidate days. It is observed that grouping all sites into a single factor group increased the MAPE by an average of 4.5% in absolute terms compared to allocating sites to their empirical factor groups. As expected – and consistent with previous literature (Sobreira & Hellinga, 2024 – Chapter 6) – this impact was more pronounced for sites in Factor Group #3 (average absolute increase in MAPE of 13.0%). Furthermore, the Single Factor Method consistently outperformed the Ratio 24/8 Method. It is important to note that the Empirical Factor Group Method is only applicable to sites with continuous monitoring and cannot be applied to STC sites, creating the need to develop models that associate STC sites with appropriate factor groups.

Table 7-8 presents the performance of the Modeling Methods compared to the Single Factor Method. Unlike the spatial transferability cases, where the models did not consistently yield improvements, when the base and target jurisdictions are the same, the Modeling Methods always show benefits. Modest improvements (0.3 to 5.1% lower MAPE) are observed in overall performance, while significant improvements (5.5 to 12.3% lower MAPE) are noted for Factor Group #3: Schools. These results align with the findings from (Sobreira & Hellinga, 2024 – Chapter 6).

Table 7-7 Summary of AADPT estimation accuracy when base and target jurisdictions are the same – candidate days

Jurisdiction Configuration #Sites			MAPE			Single Factor versus Emp. Factor Group	
			Ratio 24/8	Emp. Factor Group	Single Factor (Base = Target)	Absolute difference	Relative difference
Waterloo	Overall	101	36.8	23.6	26.0	+2.4	+10.1%
	#3: Schools	21	53.5	25.5	36.5	+11.0	+42.9%
Toronto	Overall	34	26.0	14.9	19.1	+4.2	+28.2%
	#3: Schools	6	47.0	21.2	32.4	+11.2	+52.6%
Milton	Overall	33	48.8	24.3	31.2	+6.9	+28.5%
	#3: Schools	15	71.3	23.9	40.8	+16.9	+70.9%
Average: Overall						+4.5	+22.2%
Average: #3: Schools						+13.0	+55.5%

Table 7-8 Impact of Modeling Methods on AADPT estimation accuracy when base and target jurisdictions are the same – candidate days

Base jurisdiction	Configuration	#Sites	MAPE	Absolute difference in MAPE Model i - Single Factor			
			Single Factor (Base = Target)	MLR	LR	S + C	S
Waterloo	Overall	101	26.0	-1.3	-0.6	-1.1	-1.5
	#3: Schools	21	36.5	-5.5	-5.1	-9.9	-7.8
Toronto	Overall	34	19.1	-	-	-1.0	-0.3
	#3: Schools	6	32.4	-	-	-10.3	-9.2
Milton	Overall	33	31.2	-	-3.4	-3.5	-5.1
	#3: Schools	15	40.8	-	-7.5	-10.1	-12.3

7.6.5 Conclusions and practical recommendations

Overall, the spatial transferability of expansion factors across jurisdictions with similar school-holiday months and weather seasonality has proven effective, making it a promising method for estimating long-term pedestrian volumes in jurisdictions that have

STCs but lack sites with continuous monitoring. However, when transferring expansion factors between jurisdictions with differing school-holiday schedules and weather patterns, the performance of the spatial transferability was poor. It remains unclear whether this issue arises from differences in factor group configurations or the lower pedestrian activity levels in Pima County that influence the magnitude of the expansion factors. Further research is needed to address this uncertainty.

The spatial transferability of models to assign STC sites to factor groups demonstrated good accuracy of classification. However, the application of expansion factors derived from these models produced inconsistent results: sometimes the Modeling Methods outperformed the Single Factor Method, and other times the opposite was true. Adding this inconsistency to the complexity of developing the models, we recommend using the Single Factor Method when spatially transferring expansion factors, particularly with STCs collected during periods of typical traffic (i.e., candidate days). Further research considering different jurisdictions is needed to support this recommendation.

The use of spatially transferred expansion factors through the Single Factor Method showed superior performance compared to alternative approaches available to jurisdictions lacking CC sites, such as crowd-sourced data, Large Language Models, or the local calibration of spatially transferred DD models. This further highlights the potential of this method for estimating long-term pedestrian volumes at intersections, as it only requires the availability of STCs – which are relatively easy and inexpensive to collect or may already be available from traffic signal design studies – and the “borrowing” of expansion factors from a similar jurisdiction.

For future research, we recommend:

1. Assessing the spatial transferability of expansion factors across a wider range of jurisdictions. The "similar" jurisdictions examined in this study are geographically close, potentially sharing habits and customs that may influence pedestrian activity patterns throughout the year. For example, it would be valuable to explore transferability in a

jurisdiction within Alberta, Canada, which has similar weather conditions and holiday periods to Ontario. A similar comparison could also be conducted across different USA states.

2. Further investigating modeling approaches to associate STC sites with factor groups when transferring expansion factors. Examining base jurisdictions with a larger number of CC stations may lead to the development of more robust models, potentially improving their performance and making them a more appealing alternative to the Single Factor Method.

To conclude this Chapter and encourage further assessment of the spatial transferability of expansion factors in other jurisdictions, Table 7-9 presents the DOWOM expansion factors on candidate days for the Ontario jurisdictions considered in this work, using the Single Factor Method. These expansion factors should be employed to expand 8-hour STCs (7am-9am, 11am-2pm, and 3pm-6pm) on candidate days.

Table 7-9 DOWOM expansion factors: Single Factor Method

Month	Day of week	Jurisdiction / Expansion factor		
		Waterloo	Toronto	Milton
April	Tuesday	1.444	1.725	1.662
	Wednesday	1.467	1.800	1.670
	Thursday	1.314	1.597	1.375
May	Tuesday	1.357	1.610	1.416
	Wednesday	1.366	1.653	1.504
	Thursday	1.286	1.540	1.331
June	Tuesday	1.465	1.630	1.537
	Wednesday	1.381	1.642	1.526
	Thursday	1.394	1.660	1.491
September	Tuesday	1.559	1.556	1.428
	Wednesday	1.408	1.524	1.259
	Thursday	1.373	1.461	1.249
October	Tuesday	1.429	1.521	1.352
	Wednesday	1.405	1.502	1.284
	Thursday	1.567	1.502	1.262
November	Tuesday	1.921	1.854	1.966
	Wednesday	1.698	1.751	1.846
	Thursday	1.548	1.576	1.493

Chapter 8

Estimating Pedestrian Exposure at Intersections for Different Data Availability Conditions

This Chapter is based on the following journal paper:

Sobreira, LTP; Hellinga, B. Estimating Pedestrian Exposure at Intersections for Different Data Availability Conditions. *Transportation Research Record: Journal of the Transportation Research Board*, under review.

8.1 Overall context

The methods discussed in the previous Chapters of this Thesis (i.e., the expansion factor method and the local development, local calibration, and naïve transferability of DD models) have been explored in the literature and in this Thesis, but primarily as standalone approaches. In Scenarios 2-B, 2-C, 3-B, and 3-C (Figure 1-1) – scenarios that likely reflect the reality of most U.S. and Canadian jurisdictions – these methods are applied simultaneously, raising unresolved questions. For instance, how does the performance of each method vary with the number of sites with available data? What is the threshold number of sites with available data to determine whether to pursue DD model development or local calibration? Lastly, how do spatially transferred expansion factors influence DD model development and local calibration compared to those derived directly from the target jurisdiction?

The objective of this Chapter is to assess the performance of different methods to estimate pedestrian exposure at intersections under varying conditions of data availability, providing clarity on the most suitable scenarios for each method. Additionally, this Chapter provides guidelines for practitioners seeking to estimate pedestrian exposure at intersections.

The remaining sections of this Chapter are organized as follows: the next section describes the study sites, followed by the framework and method. Subsequently, results and discussion are presented. Finally, guidelines for practitioners and conclusions are provided.

8.2 Study sites

Similarly to Chapter 5, this Chapter was conducted using sites in Cambridge, Kitchener, and Waterloo, the three main cities forming the Region of Waterloo, ON, Canada. It considered 145 sites equipped with camera-based CC stations, all located at signalized intersections. The analysis period spans a full year, from September 1, 2023, to August 31, 2024. Filters identical to those applied in previous Chapters were used to ensure proper data collection by CC stations and to remove outliers.

After applying the filters, the average number of days available for each site was 362, with a minimum of 343 days. This means that AADPT and DOWOM expansion factors could be computed for all sites.

Figure 8-1 shows the AADPT for each site and its corresponding percentile. Few sites display large volumes, mostly located near major universities. Additional statistics for the AADPT are as follows: average = 827, standard deviation = 1,074, minimum = 25, and maximum = 6,149.

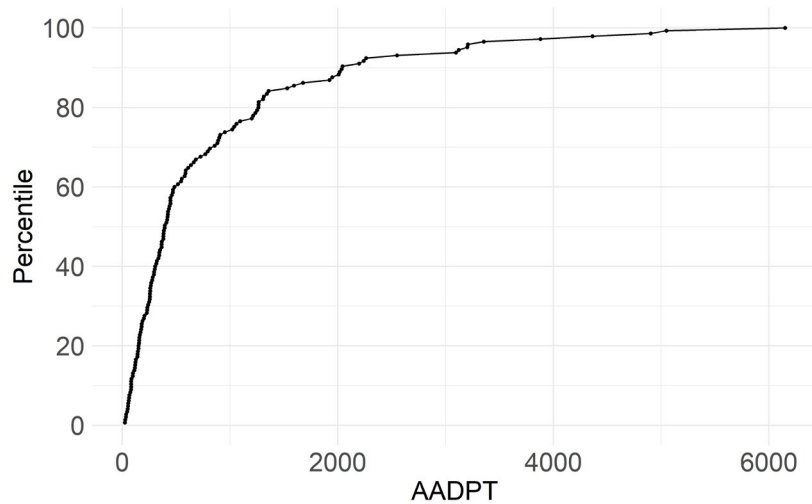


Figure 8-1 Cumulative relative frequency of intersection AADPT

8.3 Framework and method

The presence of CC stations at 145 sites in the Region of Waterloo enables the calculation of observed AADPT at all locations. Additionally, it allows each site to be treated as an STC site – by extracting STCs from the CC data – or as a site without any information. For this study, it was assumed that the target jurisdiction consists exclusively of these 145 sites, facilitating the simulation and evaluation of performance across different methods and data availability scenarios (e.g., 10 sites with CC stations, 55 sites with STCs, and 80 sites without any information).

This Chapter employs DOWOM expansion factors to expand 8-hour STCs collected during the following time periods: 7am – 9 am, 11 am – 2 pm, and 3pm – 6 pm. These are the typical time periods for which 8-hour turning movement counts are taken in Canadian jurisdictions. Additionally, this Chapter focuses exclusively on STCs collected on candidate days (i.e., periods of typical activity). For the Region of Waterloo, the following periods were defined as typical: non-holiday Tuesdays, Wednesdays, and Thursdays in April, May,

June, September, October, and November. These months were chosen to avoid periods of inclement weather and school holidays.

It was also decided to implement the expansion factor method using a single factor group (i.e., expansion factors are calculated as the average across all sites, without considering factor groups). This decision is based on: most STCs available for jurisdictions are derived from traffic signal studies conducted on typical days; b) applying the factor group approach adds a layer of complexity to the expansion factor method; and c) only marginal benefits are achieved when using factor groups with typical days, as shown in Chapter 6 and Chapter 7.

This Chapter employs two performance metrics: E (Equation 2.11) and MAPE (Equation 2.10). MAPE is also presented disaggregated by daily pedestrian volume category: low (< 150), mid (150-500), high (500-1,000), and very high ($> 1,000$).

The next subsection introduces the methods for applying the naïve transferability, local calibration, and model development of DD models are described. Subsequently, and most importantly, the process implemented to simulate different scenarios of data availability is discussed.

8.3.1 Naïve transferability and local calibration of DD models

Figure 8-2 illustrates the method for the naïve transferability and local calibration of DD models developed in another jurisdiction. The process begins with selecting potential DD models from the literature and gathering relevant details, including explanatory variables, model coefficients, and descriptive statistics of pedestrian activity and explanatory variables used during model development.

Next, the explanatory variables for sites in the target jurisdiction are collected. An investigative step follows, comparing the magnitude of these variables (e.g., average, minimum, and maximum) between the jurisdiction where the model was developed

(information often available in the documentation of the model) and the target jurisdiction. Significant differences in magnitude of the explanatory variables may suggest that the DD model is unsuitable for transfer to the target jurisdiction. For example, if employment density is an explanatory variable in the model, with average and maximum values of 500 and 1,000 employees per sq-km, respectively, in the jurisdiction where the model was developed, but 2,500 and 5,000 employees per sq-km in the target jurisdiction, it indicates that the model was not calibrated for the target jurisdiction's conditions and is unlikely to be transferrable. Finally, AADPT is predicted for each site and compared to observed AADPT using scatterplots, E (Equation 2.11) and MAPE (Equation 2.10).

The local calibration is an extension of the naïve transferability by using sites with available pedestrian volume data to adjust the naively predicted AADPT. This study employs the local calibration Model C (Equation 4.4).

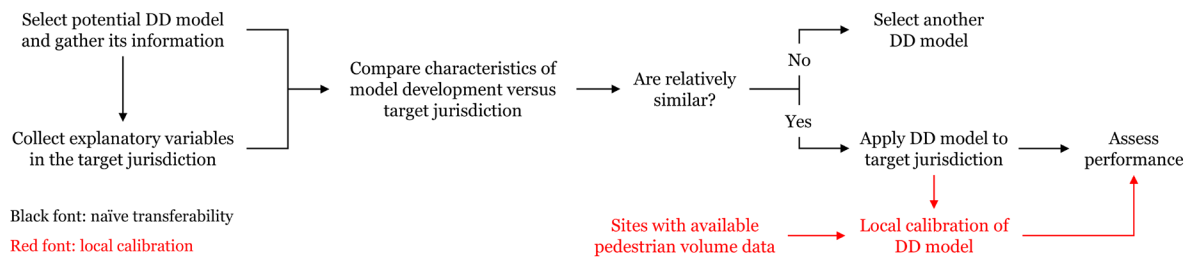


Figure 8-2 Method: Naïve transferability and local calibration of DD models

8.3.2 Local development of DD models

The initial step in developing DD models involves identifying potential explanatory variables to associate with the dependent variable. Based on a literature review, 15 variables were considered: a) variables within a 400m or 800m radius: commercial area, employment density, household income, household size, population density, residential area, schools, walk commute mode share, and worship places; b) variables within a 100m, 200m, 400m, or

800m radius: transit stops; and c) non-radius-based variables: whether the intersection is four-way and whether the intersection is on a major street.

The variable representing schools was tested in two forms: a) all schools and b) primary schools, high and post-secondary schools, and major universities. A more detailed description of the statistically significant variables is provided in the next section.

The explanatory variable dataset consisted of either 12 variables (all schools combined) or 14 variables (schools disaggregated by type). Additionally, each variable can be assigned to different radii, and continuous variables were also tested with and without logarithmic ($\ln$) transformed form. Consequently, there are millions of potential sets of variables, making exhaustive testing impractical. To address this, a Monte Carlo simulation process was applied:

1. Randomly select the form of the school variable (aggregate or disaggregate).
2. Randomly assign a radius to each variable and, if the variable is continuous, randomly determine whether it is $\ln$ -transformed.
3. Examine the linear correlation of every pair of variables, randomly removing one variable of the pair if the correlation exceeds 0.50.
4. Calibrate the model using the remaining variables and a stepwise procedure.

We initially considered log-linear and Negative Binomial models for Step 4, with log-linear models outperforming Negative Binomial models. However, we observed overfitting, particularly when simulating scenarios with small samples of sites with pedestrian volume data, as only a few sites were available for model calibration. To address this issue, we applied Least Absolute Shrinkage and Selection Operator (LASSO) regularization to the log-linear models (Tibshirani, 1996).

LASSO modifies the traditional objective function of linear (and log-linear) models which minimizes the sum of squared residuals, by introducing a penalty parameter (λ) proportional to the absolute values of the model coefficients (Equation 8.1). In essence,

LASSO regularization: a) aids in the selection of explanatory variables by reducing some coefficients to zero, and b) penalizes large coefficients. This approach results in more conservative models with a reduced risk of overfitting.

LASSO regularization objective function:

$$SSR + \lambda \times \sum_{j=1}^k |\beta_j| \quad (8.1)$$

Where:

SSR = Sum of squared residuals

λ = Penalty parameter that varies from 0 to infinity

k = Total number of explanatory variables

β_j = Model coefficient of explanatory variable j

One step in the LASSO regularization process is the selection of the parameter λ , which controls the magnitude of the penalty applied to the model coefficients. Larger values of λ result in more conservative models by shrinking coefficients more aggressively. Two values of λ are often selected: a) minimum λ : the value of λ that minimizes error when applied to validation datasets; and b) 1-standard error: the largest value of λ that produces an error within one standard error of the minimum error found in option (a). The 1-standard error λ provides a slightly more conservative model and was the approach adopted in this work.

The Monte Carlo simulation process can be repeated N times to identify an optimal model. In this study, N was set to 1,000, as discussed in the next subsection. Among the 1,000 simulations, the best model was selected based on performance metrics E , MAPE, and the number of statistically significant variables (simpler models were preferable).

8.3.3 Simulating scenarios of data availability

Figure 8-3 illustrates an example of data availability and the candidate methods to estimate AADPT for each type of site in a fictional jurisdiction with 30 sites. In this example, 10 sites have CCs (depicted in blue), allowing direct calculation of AADPT. These sites are also used to develop DOWOM expansion factors, which are applied to expand the STCs from the 10 STC sites (depicted in yellow). As a result, there are $k = 20$ sites with available AADPT. There are 10 sites with no pedestrian volume information available (depicted in grey). The k sites are used to locally calibrate and develop DD models, enabling two different estimates of AADPT for these sites.

As this procedure is performed in a controlled environment (i.e., observed AADPT is known for all sites), the performance of the expansion factor method for STC sites, as well as the local calibration and development of DD models for sites without information, can be evaluated.

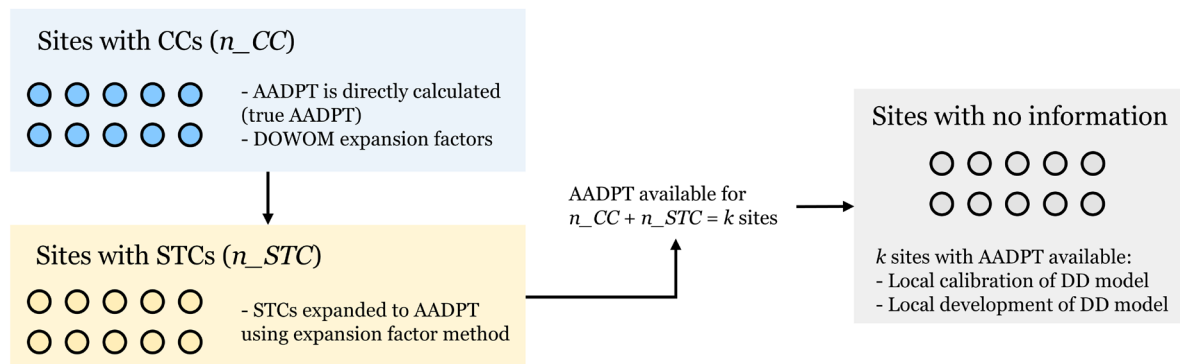


Figure 8-3 Example of scenario of data availability

There are 145 sites with CC stations in the Region of Waterloo. The simulation of different scenarios of data availability ranged from 0 to 145 sites with available information, in increments of 5 sites. The first step was to define the number of CC stations (n_{CC}) necessary to adequately implement the expansion factor method; this represents the blue sites

in Scenarios 2-A, 2-B, and 2-C in Figure 1-1. Nordback et al. (2019) recommend at least five CC stations per factor group when applying the expansion factor method. For our application, which uses a single factor group, we tested n_{CC} values between 1 and 50. $n_{CC} = 10$ provided results similar to $n_{CC} = 50$. $n_{CC} = 5$ presented similar average performance to $n_{CC} = 10$ but with much higher variability. Therefore, we selected $n_{CC} = 10$.

When $k = 0$ (number of continuous count sites, $n_{CC} = 0$ and number of STC sites, $n_{STC} = 0$), then Scenario 3-D in Figure 1-1 is represented. When $k = 145$ ($n_{CC} = 10$ and $n_{STC} = 135$), Scenario 2-A is represented. For the remaining values of k (10, 15, ..., 140), there is varying availability of STC sites and sites with no information. The performance assessment of locally calibrated and developed DD models for these cases allows for the proper identification of Scenarios 2-B and 2-C, indicating the appropriate threshold for selecting local calibration or development of DD models (dashed line in Figure 1-1). A similar rationale applies to Scenarios 3-A, 3-B, and 3-C, but with the absence of CC stations and the use of expansion factors from another jurisdiction. The simulation process is presented below:

1. Out of the $n = 145$ sites, randomly select 10 CC sites (n_{CC})
 - a. Calculate observed AADPT for the selected CC sites
 - b. Develop DOWOM expansion factors from the selected CC sites
2. Out of the remaining sites ($n - n_{CC} = 135$), randomly select n_{STC} sites and one STC from each STC site. $n_{STC} \in \{0, 5, 10, \dots, 135\}$. In the first iteration, $n_{STC} = 0$
 - a. Expand STCs to AADPT using the appropriate DOWOM expansion factor
3. Determine the total number of sites (k) with available AADPT, where $k = n_{CC} + n_{STC}$
4. Estimate AADPT for the remaining sites ($n - k$), for which no pedestrian count information is available using the following methods:
 - a. Naïve transferability of a given DD model
 - b. Local calibration of a given DD model

- c. If $k \geq 30$, Local development of DD model. This step uses the process described in the previous sub-section for optimal model selection
- d. If $k = n = 145$, only the expansion factor method is applied
- 5. Measure overall performance across all sites, as well as separately for STC sites and sites with no information
- 6. Return to step 2 and proceed to the next iteration with $n_STC = 5$

There are countless combinations of CC sites, STC sites, and specific STCs from each site (each STC site has 78 candidate days available). Running the simulation process once from 10 to 145 sites constitutes a single “round of simulation”. Within each round and for each value of k , there is another iterative process in step 4c to select the optimal locally developed model. The number of replications to find the optimal model is referred to as “loops of simulation”. This approach is computationally intensive, requiring a balance between the number of rounds and loops of simulation. After analyzing model performance as a function of the number of loops, we set the number of rounds at 60 and the number of loops at 1,000. The median and the 85th percentile of the performance across rounds for each value of k were analyzed.

8.4 Results and discussion

This section is organized into two subsections. The first presents results for the naïvely transferred, locally calibrated, and locally developed DD models, assuming that observed AADPT for all sites is used in the local calibration and development of the models. While this condition does not reflect any of the data availability scenarios tested in the simulation process, it provides valuable insights into the underlying aspects of each method. The second section presents the results from the simulation of the data availability scenarios.

8.4.1 Observed AADPT for all sites

This subsection is organized by first presenting the results for the naïve transfer and location calibration of DD models and then second, presenting the results for locally developed DD models.

8.4.1.1 Naïve transferability and local calibration of DD models

Table 8-1 presents the three DD models developed in other jurisdictions, which were considered for naïve transferability and local calibration. These models were selected because they demonstrated adequate transferability to Milton, Ontario (Sobreira & Hellinga, 2023a – Chapter 3), a jurisdiction that is relatively similar to the Region of Waterloo. Detailed information about these DD models, including their structure and coefficients, can be found in the model documentation (Griswold et al., 2019; Schneider et al., 2021; Singleton et al., 2021) and in Chapter 3.

The transfer compatibility of these models to the Region of Waterloo was assessed by comparing the distribution of each explanatory variable across the 145 sites in the Region of Waterloo with the average, minimum, and maximum values observed during the original model development. Overall, all three models demonstrated adequate compatibility with the Region of Waterloo (i.e., none of the variables exhibited a completely different magnitude). However, a few variables had maximum values in the 145 sites that exceeded those from the model development. For instance, the model developed by Schneider et al. (2021) observed a maximum of 7 transit stops, whereas the maximum in the Region of Waterloo sites was 11. To prevent applying the model to conditions outside its calibration range, the maximum value of each variable was constrained to the corresponding maximum observed during model development.

Table 8-1 Summary of DD models transferred from other jurisdictions

Model	Jurisdiction	Sample size	Explanatory variables
Griswold et al.	State of California, USA	1,270	Number of employees, Population, Street segments, Walk commute mode share, Schools, Intersection is on a principal arterial, Intersection is on a minor arterial, Four-way intersection
Schneider et al.	Milwaukee, USA	260	Population density, Employment density, Transit stops, Retail places, Restaurants and bars, Presence of schools, Household with zero vehicle
Singleton et al.	State of Utah, USA	1,494	Population density, Employment density, Household size, Household income, Vehicle ownership, Residential places, Commercial places, Intersection density, 4-way intersections, Schools, Worship places, Transit stops, Park acreage, Intersection is on a major road

Table 8-2 summarizes the performance for naïve transferability and local calibration of each transferred DD model. The local calibration coefficients were calculated assuming observed AADPT for all 145 sites. This is an exercise to provide the reader with a sense of the magnitude of these coefficients and does not reflect any of the data availability scenarios further tested in the simulation process.

Table 8-2 Summary of naïve transferability and local calibration of DD models

DD model	Naïve transferability		Local calibration			
	MAPE	E	γ^1	α^1	MAPE	E
Griswold et al.	70.8	0.610	1.108	0.832	77.0	0.551
Schneider et al.	89.1	0.927	2.211	0.657	81.2	0.603
Singleton et al.	95.4	0.711	2.551	0.598	97.2	0.619

¹Refer to model coefficients in Equation 4.4: applied to ln-transformed AADPTs

Overall, all three DD models demonstrated reasonable transferability, with performance metrics falling within expectations. For context, Sobreira & Hellinga (2023a)

(Chapter 3) reported E values exceeding three and MAPEs greater than 300% when transferring the same models to Downtown Toronto, Canada, and Pima County, USA.

The transferability of the Griswold et al. (2019) exhibited the best performance among the three models. Its E value of 0.61 is comparable to the accuracy reported during the development of other models in the literature (Munira et al., 2020; Schneider et al., 2021). For the other two models, E and MAPE suggest different conclusions: Schneider et al. (2021) achieved a better MAPE, while Singleton et al. (2021) had a better E. This indicates that Schneider's model performs better at sites with low pedestrian activity, whereas Singleton's model is better at sites with high activity.

Local calibration improved the performance of all models in terms of E, achieving values indicative of practical adequacy in all cases. However, only Schneider et al. (2021) showed improvement in MAPE. This was due to local calibration leading to overestimation at sites with low pedestrian activity, which significantly impacted MAPE.

8.4.1.2 Local development of DD models

Table 8-3 summarizes information about the optimal locally developed DD model, which was calibrated assuming observed AADPT for all 145 intersections. Again, this exercise aims to provide the reader with insight into the types of variables and the magnitude of coefficients in the developed DD models. It does not reflect any of the data availability scenarios tested further in the simulation process.

Overall, the coefficients align with expectations and are consistent with those reported in the literature. A key finding is the distinction between secondary/post-secondary schools and major universities. The coefficient for major universities is significantly larger, suggesting that universities such as the University of Waterloo and Laurier University generate more pedestrian trips than secondary and post-secondary schools. The LASSO regularization procedure performs variable selection without testing hypotheses about the

relationships between independent and dependent variables; that is why p-values are not reported in Table 8-3.

Table 8-3 Summary of DD model development: statistics and model coefficients

Variable	Radius (m)	Statistics				Model coefficients
		Avg.	SD	Min.	Max.	
Intercept	na	na	na	na	na	3.8893
#Transit stops	200	4.66	2.95	0	15	0.1101
Commercial area (%) ^{1,2}	400	21.1	15.6	0	65.6	0.0207
#Schools: secondary & post-secondary ¹	400	0.61	1.02	0	4	0.0249
Population density ^{1,3}	400	2419	1207	4	6161	0.0913
Walk commute to work (%)	800	10.6	7.9	0.2	28.4	0.0684
#Major universities	800	0.11	0.39	0	2	0.6194
Intersection is four-way (0=No;1=Yes)	na	0.77	0.43	0	1	0.2864
MAE	348.5					
E	0.421					
MAPE	63.9					
Lambda (1-Standard error) ⁴	0.059					

Note: na = not applicable; Avg. = average; SD = standard deviation. ¹Ln-transformed variables; null variables were replaced by 0.001. ²Area of parcels that are commercial / total buffer area × 100. ³Population per sq-km. ⁴Refer to Equation 8.1

8.4.2 Simulation of data availability scenarios

This section begins with a detailed evaluation of the performance of locally developed DD models as a function of the number of sites with available pedestrian volume data. Following this, the simulation process is presented, considering the expansion factor method and the naïve transferability, local calibration, and local development of DD models. The evaluation of naïve transferability and local calibration is limited to the models of Griswold et al. (2019) and Schneider et al. (2021), which demonstrated the best and worst performance, respectively, for the naïve transferability (Table 8-2).

8.4.2.1 Local development of DD models as a function of data availability

Figure 8-4 summarizes the performance of locally developed DD models in terms of E across 60 rounds of simulation for each value of k. The results are shown in three categories: blue, green, and red. For instance, when $k = 30$, the model is developed using data from 30 sites: 10 CC sites with observed AADPTs and 20 STC sites with AADPTs estimated by applying an expansion factor to a STC. The blue category represents the comparison between model predictions and the AADPTs used in model development (which is a combination of observed AADPT and estimated AADPT). The green category compares model predictions to observed AADPTs for all sites used in model development. Lastly, the red category represents the comparison between model predictions and the observed AADPTs but only for the 115 sites for which it is assumed no information is available (i.e., model application). The solid line represents the median E across the 60 simulation rounds, and the shaded background depicts the 15th and 85th percentiles. Figure 8-4 presents some interesting findings:

1. As expected, the performance of the model application (red category) improves as the number of sites used for model development (k) increases. When $k \geq 130$ the background range (85th – 15th percentile) becomes larger due to the smaller sample size for the model application, amplifying the impact of variability across sites.
2. The performance of the locally developed DD model (blue and green categories) becomes worse as the number of sites increases up to approximately $k = 60$, then stabilizes beyond this point. This trend, combined with the poor application performance (red category) for small values of k , indicates potential overfitting in the models. Additionally, as k increases, the variability in model performance decreases, reflected in the narrower width of the background.
3. The difference between the blue and green categories represents the “compromise” of using expanded AADPTs instead of observed AADPTs for STC sites. This compromise is relatively minor, with medians differing by an average of only 0.02 points of E. Given

that E is approximately 0.45 for the blue curve when $k > 60$, this corresponds to a relative difference of 4.4%.

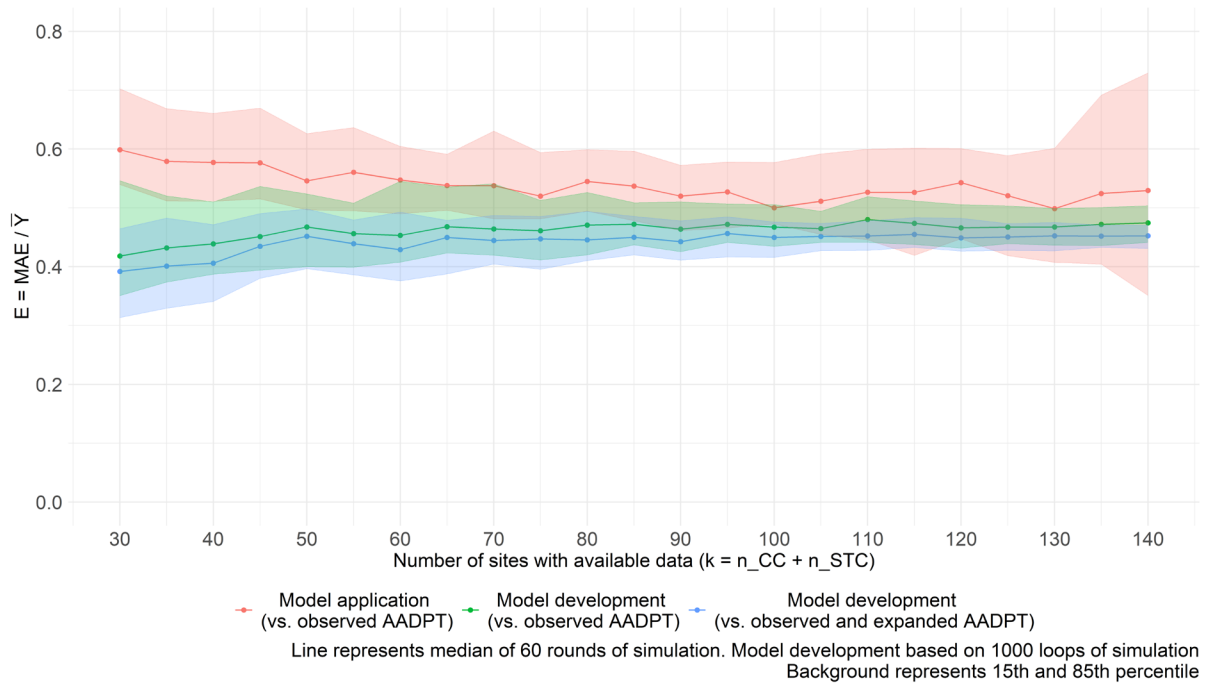


Figure 8-4 Simulation process: summary of model development

This subsection detailed the performance of local DD model development and application as a function of data availability. The next subsection further examines these data availability scenarios by incorporating naïve transferability and local calibration of DD models as additional candidate methods for estimating AADPT at sites with no pedestrian volume counts.

8.4.2.2 Comparison between methods

Figure 8-5 summarizes the performance of the assessed methods in terms of E across 60 simulation rounds for each value of k and using the Griswold et al. (2019) model for the naïve transferability and locally calibrated DD models. The figure combines performance results for all three types of sites (i.e., CC sites, STC sites, and sites with no information), simulating the overall performance a practitioner would encounter. For instance, when $k = 50$, there are 10 CC sites from which we can compute AADPT with no error, 40 STC sites where the expansion factor method is applied, and 95 sites with no information, for which predictions from the naïvely transferred, locally calibrated, and locally developed DD models are included. The solid line represents the median E across the 60 simulation rounds, while the shaded background corresponds to the 85th percentile. Additionally, the top of the figure highlights the best-performing method for each value of k, based on both the median and the 85th percentile.

As k increases, the number of sites with available STCs where the expansion factor method is applied also increases. Consequently, as expected, the overall performance improves, and the three solid lines converge toward the same value as k approaches 145 sites.

Overall, as discussed in Table 8-2, the naïvely transferred Griswold et al. model demonstrated reasonably good accuracy ($E = 0.61$). In Figure 8-5, it produced comparable results to those from the locally calibrated and locally developed DD models. When $k < 30$, the locally calibrated DD model results exhibit high variability and performs worse than the naïvely transferred DD model. However, for $k \geq 30$, the locally calibrated DD model performs better than the naïvely transferred DD model (for both the median and 85th percentile).

The locally developed DD model was evaluated for $k \geq 30$. Even with a small sample size ($k = 30$), the generated models remain relatively stable, with the 85th percentile only 0.10 points higher than the median. This stability is attributed to the application of the LASSO regularization. In contrast, we observed much greater variability in performance for small values of k when traditional Ordinary Least Squares (OLS) log-linear models were

employed (i.e., median and 85th percentile of E around 0.55-0.60 and 0.90, respectively). The locally developed DD models perform better than locally calibrated models for $k \geq 60$.

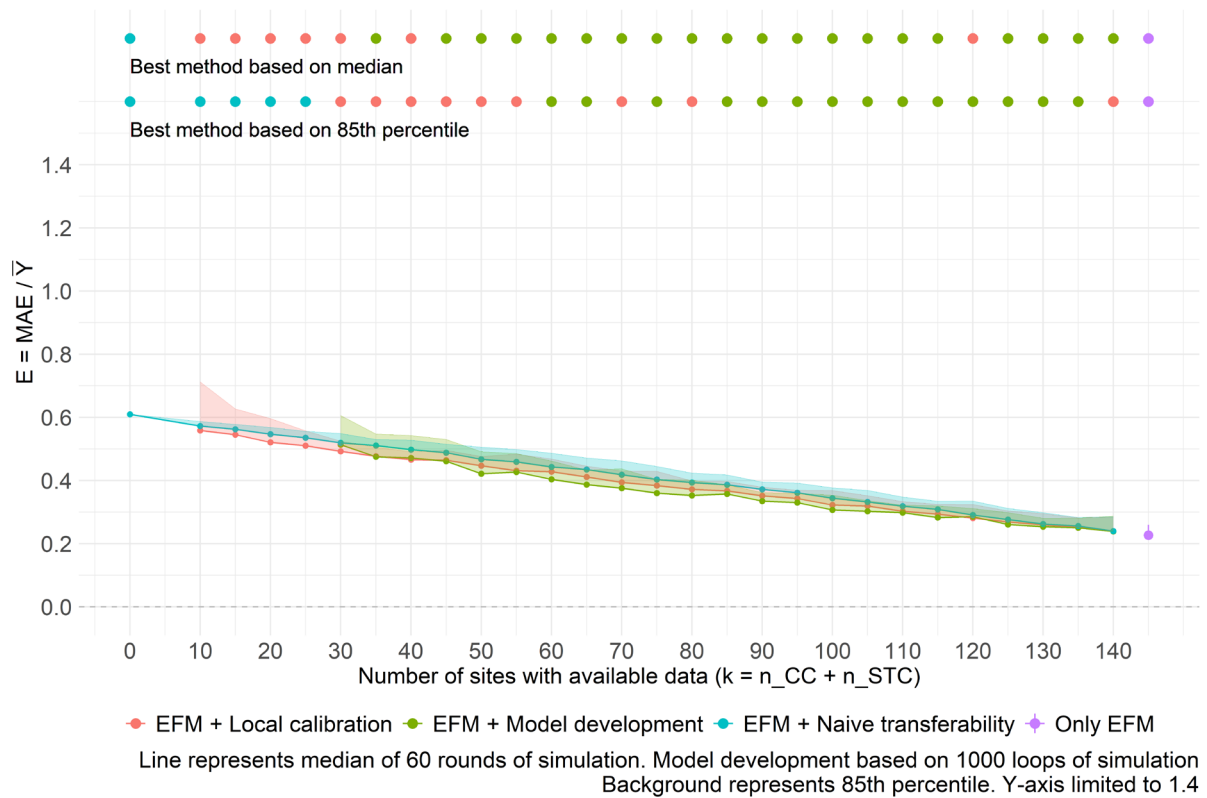


Figure 8-5 Simulation process: performance across all sites. DD model transferred: Griswold et al.

Note: n_{CC} = number of sites with continuous count stations; n_{STC} = number of sites with short-term counts; EFM = expansion factor method. The number of continuous count stations is limited to ten.

Figure 8-6 summarizes the overall performance when the Schneider et al. model is transferred. Among the three DD models tested, the Schneider et al. model exhibited the poorest naïve transferability to the Region of Waterloo (Table 8-2), a trend that is also reflected in Figure 8-6. Consequently, the benefits of applying local calibration are more

pronounced, with performance surpassing the naïve transferability when $k \geq 10$, a smaller value of k than observed in Figure 8-5. Additionally, small values of k lead again to high variability in performance, reinforcing the need for caution when applying local calibration under such conditions. The curve for locally developed model remains the same as in Figure 8-5, demonstrating superior performance when $k \geq 50$.

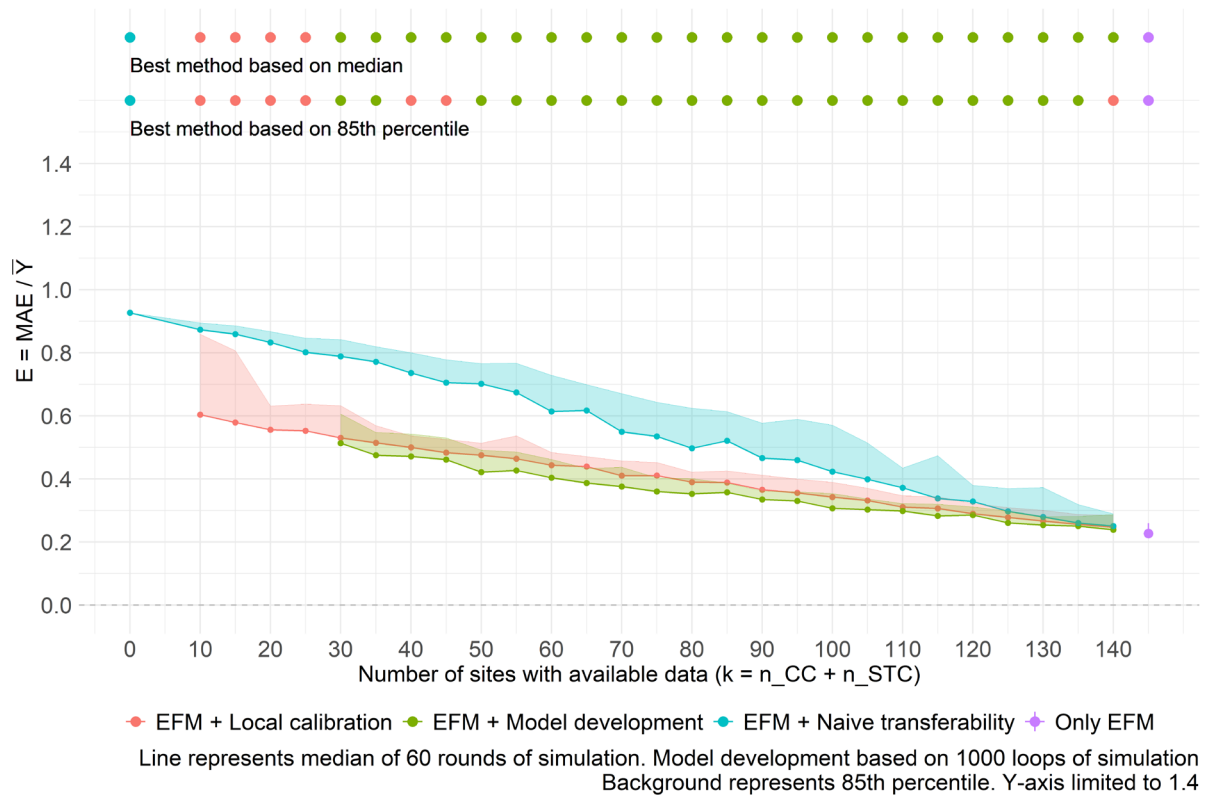


Figure 8-6 Simulation process: performance across all sites. DD model transferred: Schneider et al.

Note: n_{CC} = number of sites with continuous count stations; n_{STC} = number of sites with short-term counts; EFM = expansion factor method. The number of continuous count stations is limited to ten.

Figure 8-7 is similar to Figure 8-6 (i.e., Schneider et al. model is transferred) but quantifies model performance across only those sites with no available information. The following conclusions can be made from this figure:

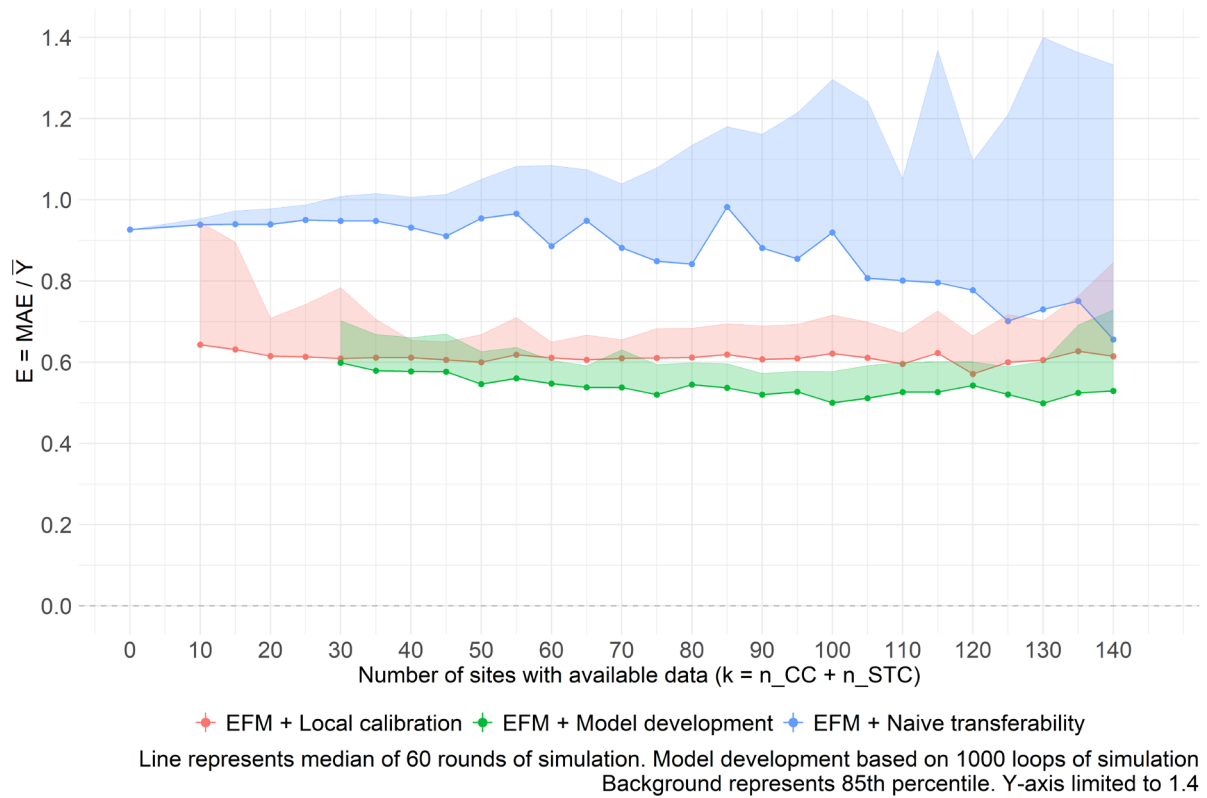


Figure 8-7 Simulation process: performance across sites with no information. DD model transferred: Schneider et al.

Note: n_{CC} = number of sites with continuous count stations; n_{STC} = number of sites with short-term counts; EFM = expansion factor method. The number of continuous count stations is limited to ten.

1. The improvement in the median performance for the naïve transferability when k increases is linked to the presence of certain sites in which the naïve transferability

performs poorly. As k increases, the naïve transferability is applied to fewer sites, while more sites receive the expansion factor method.

2. The local calibration performance stabilizes when $k \geq 20$. This trend is also observed in similar figures for the models of Griswold et al. (2019) and Singleton et al. (2021), as well as in the findings of Sobreira & Hellinga (2023b) (Chapter 4); highlighting the potential of this method even with small sample sizes.
3. The median performance for the local calibration ($E \sim 0.60$ - 0.62) is similar to the results in Table 8-2 ($E = 0.603$), where the local calibration was performed using observed AADPT for all 145 sites, suggesting that the “compromise” of using expanded AADPTs instead of observed AADPTs is minimal.
4. Locally developing the DD model improves performance by reducing E by approximately 0.05 points compared to the local calibration. A similar trend was observed in similar figures for the Singleton et al. model, while it showed smaller improvements for the Griswold et al. model.

Overall, the results from the simulation of different data availability scenarios as presented in Figure 8-5, Figure 8-6, and Figure 8-7 (as well as in similar figures for the Singleton et al. model, which are not included in this document) produced relatively consistent results in terms of the preferred AADPT estimation method based on data availability. This consistency is important considering that DD models with different accuracy of naïve transferability were considered. The recommended AADPT estimation method to use as a function of the number of sites for which pedestrian counts are available is:

- 0 to 10 sites: naïve transferability of DD model
- 10 to 50-70 sites: local calibration of DD model
- More than 50-70 sites: locally developed DD model

This indicates that the threshold separating Scenarios 2-B and 2-C, represented by the dashed line in Figure 1-1, is approximately 50 to 70 sites with available pedestrian counts. An additional scenario could have been introduced to illustrate cases where very few counts are available (fewer than ten), in which local calibration is not recommended over naïve transferability.

We acknowledge that these findings are based on a single target jurisdiction and that the thresholds may vary when applying these methods to jurisdictions with simpler or more complex characteristics. However, these results align with previous findings in the literature, which demonstrated that the local calibration of DD models can significantly improve the performance of poorly transferred models (i.e., reducing E from ~ 2.00 to $0.50\text{--}0.70$) even with small sample sizes (Sobreira & Hellinga, 2023b – Chapter 4). Additionally, prior research has shown that DD models with adequate performance can be developed when 50–70 sites with pedestrian volume data are available (Hankey et al., 2017; Munira et al., 2020; Sanders et al., 2017).

At this stage, the simulation process has been assessed exclusively using the performance metric E and assuming that CC stations are available in the target jurisdiction. Table 8-4 expands on this analysis by incorporating additional performance metrics (MAPE by AADPT category, MAPE and E by type of site, and MAPE and E for all sites). It also explores Scenarios 3-A, 3-B, and 3-C in which CC stations are not available, and expansion factors are transferred from a jurisdiction with similar weather and school holiday periods (City of Toronto, in this case). The results are presented for key values of k and consider the transferability of the Griswold et al. DD model. The main findings are compiled below:

Table 8-4 Summary: simulation process

k	Methods employed	Continuous counts are from:	Type of site	n	MAPE - AADPT by category				By type of site		All sites	
					Low (< 150)	Mid (150-500)	High (500-1000)	Very high (>1000)	MAPE	E	MAPE	E
0	Naïve transferability (Griswold et al.)	-	CC	0	-	-	-	-	-	-	70.8	0.61
			STC	0	-	-	-	-	-	-		
			No info	145	140.4	54.9	48.4	56.2	70.8	0.61		
30	Expansion factor method & Local calibration (Griswold et al.)	Waterloo	CC	10	0.0	0.0	0.0	0.0	0.0	0.00	69.4	0.49
			STC	20	26.2	21.9	19.1	21.7	23.8	0.21		
			No info	115	221.6	57.0	39.8	49.9	83.4	0.57		
		Toronto	STC	30	32.0	27.0	19.4	24.8	26.6	0.24	73.2	0.52
			No info	115	221.6	60.0	40.6	49.9	85.7	0.58		
100	Expansion factor method & Model development	Waterloo	CC	10	0.0	0.0	0.0	0.0	0.0	0.00	41.5	0.31
			STC	90	30.1	22.8	19.7	23.5	23.8	0.24		
			No info	45	249.3	49.6	44.7	44.8	88.5	0.50		
		Toronto	STC	100	33.1	27.8	22.8	25.2	27.6	0.27	46.6	0.34
			No info	45	262.7	47.3	43.8	45.0	89.7	0.52		
145	Expansion factor method	Waterloo	CC	10	0.0	0.0	0.0	0.0	0.0	0.00	22.3	0.23
			STC	135	30.2	22.8	20.5	23.5	24.0	0.25		
			No info	0	-	-	-	-	-	-		
		Toronto	STC	145	36.3	26.8	23.8	24.1	27.5	0.26	27.5	0.26
			No info	0	-	-	-	-	-	-		
Benchmark approaches												
Large Language Model and satellite imagery					273	51	51	49	-	-	-	0.50-0.71
StreetLight					151	89	72	71	-	-	-	-

Note: k = number of sites with data available; n = number of sites

1. As expected, the expansion factor method significantly outperforms all other methods, particularly for the Low (<150) category.
2. When assessing sites with no information, a reduction in error, in terms of E, is observed as we progress from the naïve transferability ($E = 0.61$) to local calibration ($E = 0.57$), and further to model development ($E = 0.50$). However, this trend does not hold for MAPE, as the Low (<150) category biases the analysis. For instance, while naïve transferability and model development yield overall MAPEs of 70.8% and 88.5%, respectively, naïve transferability is only superior to model development within the Low (<150) category. This outcome is attributed to the nature of the LASSO regularization. By generating more conservative coefficients for the explanatory variables, LASSO compensates by estimating larger intercepts, which results in overestimation at sites with low volumes.
3. Transferring expansion factors from Toronto results in only a minor degradation in absolute terms, though the percentage impact varies across methods:
 - a. Expansion factor method ($k = 30, 100$, and 145): MAPE increased by 2.8 to 3.6 points and E by 0.01 to 0.02 points. On average, these represent percentage increases of 14.2% in MAPE and 9.0% in E.
 - b. Local calibration ($k = 30$) and local development ($k = 100$): the impact of using expansion factors from another jurisdiction was smaller, with MAPE increasing by 2.3 points and E by 0.01 points for local calibration, and by 1.2 points and 0.02 points for local development. On average, these correspond to percentage increases of 2.0% in MAPE and 2.4% in E.
 - c. Given the higher costs associated with operating CC stations compared to collecting STCs, transferring expansion factors from a similar jurisdiction emerges as a cost-effective alternative for target jurisdictions without CC stations. This aligns with the findings in Chapter 7 (Sobreira & Hellinga, n.d.).
4. Table 8-4 includes two benchmark methods for estimating pedestrian exposure at intersections: a) Large Language Models (LLMs) combined with satellite imagery (Sobreira & Hellinga, n.d., 2025) and b) third-party crowd-sourced data (StreetLight, 2023). The locally calibrated and locally developed DD model methods perform better

than the latter and are relatively comparable to the former, while the expansion factor method remains the most accurate approach.

- a. This highlights the potential of the LLM method, which relies solely on satellite imagery, interactions with a given LLM, and the collection of pedestrian volumes at a few key sites. Unlike the local calibration and local development of DD models, it does not depend on land use, socioeconomic, or transportation features for its application.

8.5 Guidelines for practitioners

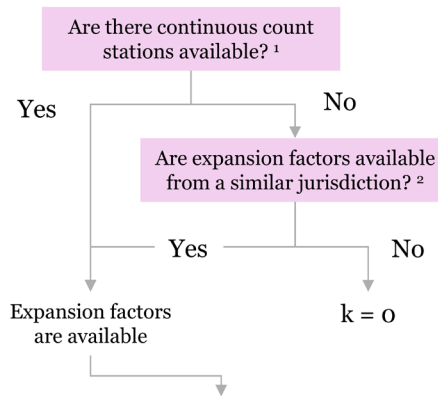
This section provides guidelines for practitioners aiming to estimate pedestrian exposure at intersections. The recommendations are drawn from the findings of this study, the authors' expertise of several years on this topic, and relevant literature. Figure 8-8 provides a flowchart outlining the key steps in the procedure for estimating AADPT at intersections at a jurisdiction-wide level.

The first step is to determine whether CC stations are available in the target jurisdiction. If they are, expansion factors can be calculated. If not, the practitioner should identify a jurisdiction with similar weather seasonality and school holiday periods to transfer the expansion factors. If no such jurisdiction exists, alternative methods – such as the naïve transferability of DD models or third-party crowd-sourced data – can be applied.

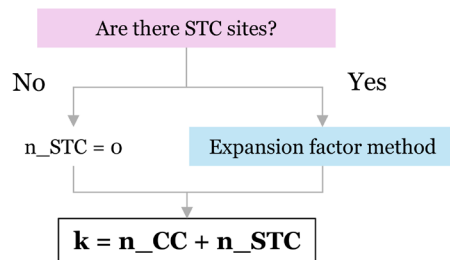
The second step is to assess the availability of STCs in the target jurisdiction. If STCs are available, the expansion factor method should be used to expand STCs to AADPT.

Although the primary focus of this Chapter was not on the expansion factor method, practitioners must make two decisions when implementing it. The first decision is the type of expansion factor to use: DOWOM or day-of-year factors. The former – used in this Chapter – is more straightforward to apply. The latter likely cannot be transferred from another jurisdiction and requires that STCs and CCs be from the same year but may yield improved performance (Hankey et al., 2014).

1. Continuous counts (n_CC):



2. Short-term counts (n_STC):



3. Sites with no information:

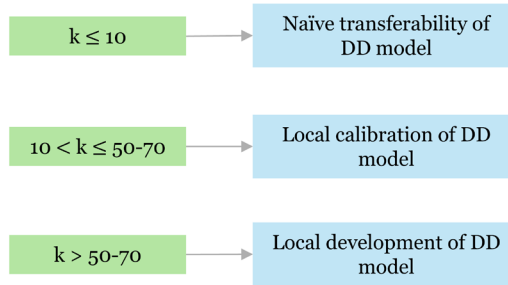


Figure 8-8 Guidelines for estimating AADPT at intersections

Note: ¹Nordback et al. (2019) recommended a minimum of five continuous count stations per factor group for applying the expansion factor method. In our study – where a single factor group was used – we suggest a minimum of 5-10 stations. ²A similar jurisdiction refers to one with comparable weather seasonality and school holiday periods (Sobreira & Hellinga, n.d. – Chapter 7).

The second decision is whether to combine expansion factors into factor groups. This decision depends on two factors: a) the availability of STCs on typical days versus all days and b) the number of CC stations and STCs available per site. The benefits of factor groups are more pronounced when using STCs from all days rather than only typical days (Sobreira & Hellinga, 2024 – Chapter 6). However, employing factor groups introduces additional complexity, as it requires associating STC sites (where pedestrian activity patterns are unknown) with a specific factor group. This task demands either modeling to link factor groups to land use attributes (Humagain & Singleton, 2021; Medury et al., 2019; Sobreira &

Hellinga, 2024) or conducting multiple STCs per site, including both weekdays and weekends (Hankey et al., 2014; Johnstone et al., 2018; L. Miranda-Moreno et al., 2013; Nordback et al., 2019).

As described in this Chapter and in the literature, the expansion factor method produces the most accurate estimates of AADPT. For sites without STCs, several alternative methods can be used. This study specifically examined spatial transferability, local calibration, and local development of DD models. The data availability thresholds recommended for each method are illustrated in Figure 8-8. General guidelines for the application of each method are presented below.

Naïve transferability of DD models:

1. Consider DD models from the literature that:
 - a. Include a sufficient number of explanatory variables (at least five or more).
 - b. Were developed using AADPT (observed or expanded) rather than STCs.
 - c. Avoid DD models with context-specific variables, such as proximity to a specific metro station or arena.
2. Assess pedestrian volume characteristics in the model's original jurisdiction and compare them to the target jurisdiction:
 - a. Look for models developed in areas with similar pedestrian activity levels (statistics provided in the original study may offer insight).
 - b. Consider differences in motorization levels; for example, a DD model calibrated in Amsterdam may not transfer well to Miami.
3. Evaluate the independent variables in the model:
 - a. Collect independent variables for the target jurisdiction and compare their average, minimum, and maximum values to those reported in the original model.
 - b. If several variables fall outside the range observed in the model development, this may indicate poor transferability. For instance, in Chapter 3 (Sobreira & Hellinga,

2023a) it was observed unrealistic AADPT predictions when transferring the Schneider et al. (2021) model to Downtown Toronto due to discrepancies in employment density.

- c. If only a few variables are out of range, you may consider constraining them at the minimums or maximums from the model development.
4. Assess the predicted AADPTs:
- a. As a practitioner familiar with the target jurisdiction, you may have a sense of the expected pedestrian activity levels for some regions.
 - b. Inspect the estimates for extremely low or high values.

Local calibration of DD models:

1. Assess the naively predicted AADPTs at sites with known observed/expanded AADPT:
 - a. Estimate accuracy (E, MAPE).
 - b. Scatterplot with predicted AADPT versus observed/expanded AADPT.
2. When applying the local calibration:
 - a. Ln-transform AADPTs before application (Equation 4.4).
 - b. If enough sites are available (i.e., > 40), consider performing a cross-validation procedure to assess the stability of the local calibration (e.g., a 4-fold procedure having 30 sites for training and 10 sites for testing).

Local development of DD models:

1. Consider using explanatory variables commonly found in other DD models.
2. Test different radii for each variable (e.g., 400 and 800m).
3. Test continuous variables in their ln-transformed form.
4. Compare multiple modeling approaches, including log-linear and Negative Binomial models. If possible, explore machine learning methods and geographically weighted models.

5. Mitigate overfitting for small datasets (<100-150 sites) by using techniques such as LASSO regularization.
6. Apply cross-validation to assess model stability. If models are unstable, this may indicate an insufficient sample size for capturing the complexity of the data. In such cases, local calibration of DD models may be a better alternative.
7. Spatially inspect the residuals of the model. This may give insights about important features that are not being accounted for (e.g., proximity to universities or located in central business district).

8.6 Conclusions

This Chapter explored various methods for estimating pedestrian exposure at intersections under different cases of data availability. The methods examined include the expansion factor method, as well as the naïve transferability, local calibration, and local development of DD models. A total of 145 signalized intersections in the Region of Waterloo with continuous monitoring were assessed, enabling the simulation of different data availability scenarios based on the number of sites with CC stations, STCs, and no available information. This assessment provided answers to the three research questions posed in the initial section:

How does the performance of each method vary with the number of sites with available data?

Regardless of the number of sites available for the other methods, the expansion factor method consistently demonstrated the best performance and is the recommended approach when STCs are available.

Local calibration was performed using three different DD models, all originally developed in different jurisdictions, which were naively transferred and locally calibrated to the Region of Waterloo. While different performances were observed across the naïve

transferability of each model, all locally calibrated models converged to a performance level of practical use ($E \sim 0.55-0.65$) when more than 10 sites were used for local calibration – this performance remained stable if more sites are used. Using fewer than 10 sites for local calibration resulted in variable and sometimes poor performances, and therefore performing local calibration with fewer than 10 sites is not recommended.

Model development was assessed using samples of at least 30 sites with available data. Traditional calibration techniques led to overfitting, necessitating the use of strategies such as LASSO regularization to mitigate this issue. As expected, increasing the sample size for model development reduced the error during application, with performances of E around 0.50-0.60 being achieved. This leads to the next question.

What is the threshold number of sites with available data to determine whether to pursue DD model development or local calibration?

The findings suggest that the threshold is around 50-70 sites, with model development applied when this threshold is surpassed and local calibration used when it is not. This aligns with conclusions from other studies in the literature, which show that small sample sizes (10-20 sites) are sufficient for achieving adequate local calibration performance (Sobreira & Hellings, 2023b – Chapter 4), while 50-70 sites are sufficient for developing DD models (Hankey et al., 2017; Munira et al., 2020; Sanders et al., 2017).

How do spatially transferred expansion factors influence DD model development and local calibration compared to those derived directly from the target jurisdiction?

The results indicate that transferring expansion factors from Toronto – a similar jurisdiction in terms of weather and school holiday periods – leads to only a slight degradation in performance (average percentage increases of 2.0% in MAPE and 2.4% in E) in DD model development and local calibration, compared to using factors from the Region of Waterloo. Considering the significant costs of operating CC stations compared to

collecting STCs, using expansion factors from a similar jurisdiction proves to be a cost-effective alternative for target jurisdictions lacking CC stations.

We acknowledge that these findings, although in line with general trends observed in the literature, are based on a single target jurisdiction. Jurisdictions with different levels of complexity (e.g., more heterogeneity across sites) may have higher thresholds for adequate model development, or jurisdictions where naively transferred models perform worse may require different guidelines for locally calibrated DD models. These uncertainties should be resolved through similar studies in different jurisdictions, which would help establish a stronger foundation for guidelines on this topic.

Chapter 9

Estimating Pedestrian Volumes at Each Crosswalk of Intersections: Comparison of Land Use Models and Short-Term Count Methods

This Chapter is based on the following journal paper:

Sobreira, LTP; Hellinga, B. Estimating Pedestrian Volumes at Each Crosswalk of Intersections: Comparison of Land-Use Models and Short-Term Count Methods. *Journal of Transportation Engineering, Part A: Systems*, vol. 150, issue 2, 2023.
<https://doi.org/10.1061/JTEPBS.TEENG-8141>

9.1 Overall context

The great majority of studies developed on estimating pedestrian volume at intersections focus on the exposure for the entire intersection. In other words, the ultimate goal is normally to find the best AADPT for the intersection, regardless of the volume distribution in each crosswalk. However, there are some existing engineering applications that require the pedestrian volume to be disaggregated per crosswalk. For instance, the 7th edition of the Highway Capacity Manual presents a novel methodology to estimate pedestrian level of service at signalized intersections in which one of the inputs is the pedestrian volume in each crosswalk (NASEM, 2022). Another case is the development of safety analyses that are disaggregated per movement. For example, to evaluate the conflicting movement between left-turn vehicles and pedestrians, pedestrian volumes in each crosswalk would be necessary. Lastly, there are traffic signal interventions that are based on each crossing instead of the entire intersection, such as leading pedestrian intervals and exclusive pedestrian phases. Disaggregated pedestrian volume per crosswalk would help in the selection of crossings for these types of interventions. In addition to the existing applications noted above, there are many traffic operational and road safety processes (e.g., network

screening) that currently do not rely on pedestrian volume data by crossing, largely because these data have not been widely available. There is the possibility that these processes can be more effective if they are extended to make use of pedestrian volume data by crossing.

The opportunity for a jurisdiction to estimate the pedestrian volume in each crosswalk of a given intersection depends on the type of information available for the intersection. When a CC station is available, AADPT can be directly calculated for each crosswalk. This is possible because CC stations conventionally collect the pedestrian volume for each crossing. When one or a few STCs are available, point estimates of the pedestrian volume and the percentage share of each crosswalk are known. In this case, the STC can be expanded to estimate the AADPT for the entire intersection, and the shares computed from the STC can be used to estimate the AADPT for each crosswalk. When no pedestrian count information is available, a DD model could be used to estimate the AADPT. All the DD models reported in the literature to date estimate the volume for the entire intersections; they cannot directly provide AADPT by crossing.

Considering the relevancy of the engineering applications that require pedestrian volume disaggregated per crosswalk and the challenges and uncertainties associated with estimating it when a CC station is not available at the intersection, the objectives of this Chapter are: a) evaluate the accuracy of estimating the AADPT in each crosswalk based on point estimates of percentage shares per crosswalk obtained from STCs; and b) develop models to estimate the percentage share of intersection AADPT per crosswalk as a function of attributes of the intersection that are commonly available for jurisdictions (e.g., land use, socioeconomic, geometric, and operational attributes). This Chapter focuses on four-way signalized intersections.

This Chapter is organized as follows: the next section presents the methods for estimating the percentage shares per crosswalk using the proposed Land Use (LU) model and STC method. After that, the results are presented and discussed. The final section provides the conclusions of this work and recommendations for practitioners and for further research.

9.2 Method

This section is divided into three sub-sections. The first presents the proposed method for modeling the percentage share of pedestrian volume per crosswalk as a function of intersection and land use characteristics in the local vicinity of the intersection. This model will be referred to as the Land Use (LU) model. The second shows the simulation process used for the STC method, and the third summarizes the procedure for evaluating the performance of both the LU model and STC method.

A common requirement for model development and STC performance evaluation is the calculation of the observed (or true) AADPT in each crosswalk. Signalized intersections where CC stations are available were used in this study. An adaptation of the AASHTO formulation for calculating average annual traffic volumes for the entire intersection was used, considering each crosswalk individually (Equation 9.1). This approach requires valid pedestrian data for at least one day-of-week-of-month to estimate the AADPT at a given intersection. Daily volumes that were outside the limits of a) 3rd quartile plus 1.5 times interquartile range or b) 1st quartile minus 1.5 times interquartile range were considered outliers and removed from the analysis. The AADPT for the entire intersection is computed by Equation 9.2.

$$AADPT_{i,u} = \frac{1}{12} \sum_{m=1}^{12} \left[\frac{1}{7} \sum_{d=1}^7 \left(\frac{1}{n_{i,u,m,d}} \sum_{k=1}^{n_{i,u,m,d}} V_{i,u,k,m,d} \right) \right] \quad (9.1)$$

$$AADPT_i = \sum_{u=N,E,S,W} AADPT_{i,u} \quad (9.2)$$

Where:

$AADPT_{i,u}$ = AADPT at crossing u (u = north, east, south, or west) of intersection i

m = months (1-12)

d = day of week (1-7)

$n_{i,u,m,d}$ = number of times a specific day of week occurs in a specified month

$V_{i,u,k,m,d}$ = daily (24 hour) pedestrian volume at crossing u of intersection i on the k^{th} occurrence of month (m) and day of week (d)

9.2.1 Modeling the percentage share of pedestrian volume per crosswalk

To avoid repetition of the long term “percentage share of pedestrian volume per crosswalk”, from now on it will be referred to as “share per crosswalk”.

9.2.1.1 Conceptual framework

The premise of DD models is that the level of pedestrian activity at an intersection (i.e., AADPT) can be explained by the characteristics of the intersection and the immediately surrounding land uses. A similar approach was adopted, and it was hypothesized that the share of the overall pedestrian activity at the intersection that uses each available crossing can be explained by the characteristics of the crossings, the characteristics of the roads that the crossings traverse, and the land uses that the crossings connect. To capture the land use, socioeconomic, and some operational attributes linked to each crosswalk, a circular buffer with a given radius generated from the centre of the intersection is segmented into four quadrants based on the roadway orientation (Figure 9-1). Quadrants Q1 and Q2, for example, represent the north crosswalk. Some geometric and operational features (e.g., crossing width and functional classification of the road where the crosswalk is located) are specific to each crossing and, therefore, do not require the use of the quadrant approach.

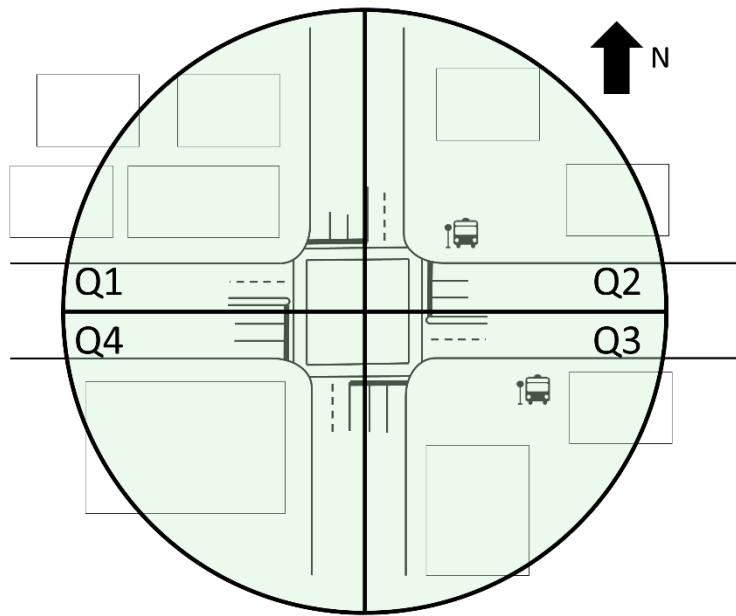


Figure 9-1 Quadrant representation

It was hypothesized that the shares per crosswalk can be modelled as a function of the potential trip interaction between quadrants that the crossing connects and that different land uses and attributes are associated with different pedestrian trip attraction and production potential. For example, consider the interaction between transit stops and retail land use. Transit stops are locations for pedestrian trip origins and destinations as pedestrians board and alight transit vehicles at this location. Retail properties can represent origins or destinations for pedestrian trips. The existence of a transit stop in quadrant Q2 directly impacts the likelihood of use of the north and east crosswalks. If the amount of retail land use is high in quadrant Q1 and minimal in quadrant Q3, then the interaction between these two attributes is expected to suggest a higher fraction of AADPT on the north crosswalk than on the east crosswalk. The magnitude of the interaction between a given pair of attributes at a given crosswalk is named the “force of the interaction” and is calculated using Equation 9.3.

$$Force_{u,x,y} = Q_{a,x} \times Q_{b,y} + Q_{a,y} \times Q_{b,x} \quad (9.3)$$

Where:

a and b = quadrant ID numbers for the two quadrants connected by crossing u . (i.e., for the north crosswalk: $a = 1$ and $b = 2$)

x and y = land use, socioeconomics, or operational attributes

$Force_{u,x,y}$ = force of the interaction between attributes x and y at crossing u

$Q_{a,x}$ = value of the attribute x (e.g., number of commercial places) collected at quadrant a for a given buffer radius

The next section presents a discussion on the candidate variables to represent the different interactions between quadrants as well as the geometric and operational variables associated with each crossing.

9.2.1.2 Explanatory variables

This is the first study that attempts to create a model for the shares per crosswalk; therefore, no prior reference is available. Explanatory variables that commonly appear in DD models for the entire intersection were considered, but in terms of the interaction between quadrants. A detailed review of these variables is available in the literature: Schneider et al. (2021), Singleton et al. (2021), and Sobreira and Hellinga (2023 – Chapter 3).

Two types of variables were considered for the modeling: radius-based and crosswalk-based. The radius-based variables are usually counted or averaged within some area. Based on the literature review, four categories were selected for this type of variable: population land use, commercial land use, the presence of schools, and the presence of transit stops. Table 9-1 presents the candidate attributes considered to represent each of the four

categories. The attributes were collected in the following radii: 50, 100, 200, 400, and 800 meters. The crosswalk-based variables are inherent to each crosswalk and do not depend on a radius. The following attributes were considered: crossing width, the presence of any type of physical median, the presence of a refuge island, road classification, the number of paved sidewalks in the corners of the crosswalk, and the presence of slip lanes. Refuge islands were defined based on the National Association of City Transportation Officials (NACTO) Design Guide criteria: the width of the median must be at least 6 feet (1.8m) and it must reach the crosswalk (NACTO, 2014).

Table 9-1 Candidate attributes to represent each category of radius-based variables

Category	Candidate attributes
Population land use	Residential area (RA) Residential places (RP) Population density (PD)
Commercial land use	Commercial Area (CA) Commercial places (CP) Employment density (ED)
Schools	All types (ATS) High and post secondary (HPS)
Transit stops	Transit stops (TS)

The radius-based variables were included in the models in the form of their interactions between quadrants. Table 9-2 presents the candidate interactions between the categories of variables that were considered. Not all possible interactions were considered, rather engineering judgement was used to identify candidate interactions. For example, it is expected that the presence of transit stops and commercial land use in adjacent quadrants will impact the use of a certain crosswalk. Similarly, transit stops in adjacent quadrants may serve different bus routes, representing potential origins and destinations at the intersection. Conversely, it is not likely that the existence of schools in adjacent quadrants will be a “trip generator” to a given crosswalk. In total, eight interactions were considered. As no prior knowledge is available regarding the best attribute and radius to represent each category and

interaction, all the attributes presented in Table 9-1 were tested with different values of radii. For instance, the interaction between population and commercial land uses was examined considering the pairs of residential area (r=100m) & commercial area (r=200m), residential places (r=100m) & employment density (r=400m), and so on. A programming code was developed to automatically define quadrants based on the roadway orientation and to collect the attributes within different radii from GIS data sources.

Table 9-2 Candidate interactions between categories

Category	Population	Commercial	Transit stops	Schools
Population	No	Yes	Yes	Yes
Commercial	-	Yes	Yes	Yes
Transit stops	-	-	Yes	Yes
Schools	-	-	-	No

After preliminary analysis, the interaction between transit stops and schools was identified as an important component in explaining the shares per crosswalk. We noted that the existence of a desire line between transit stops and schools that passes through a given crosswalk is associated with increased pedestrian volume shares. Figure 9-2 illustrates this situation, in which the presence of a transit stop in quadrant Q3 and of a school in quadrant Q4 creates a desire line that passes through the south crosswalk of the highlighted signalized intersection. Although commercial and residential land uses are also observed in the other quadrants, 70% of the crossings were made at the south crosswalk, likely because of the interaction between the school and the transit stop. There is also a second transit stop in quadrant Q4, but pedestrians travelling between this transit stop and the school do not need to traverse any of the crossings at the subject intersection and therefore there is no associated desire line. Considering the importance of this interaction, the existence of any desire lines was identified considering transit stops and schools within 800 meters from the intersection, instead of the interaction formulation considered for the other categories. A categorical variable is assigned to each crosswalk (i.e., the presence or absence of a desire line).

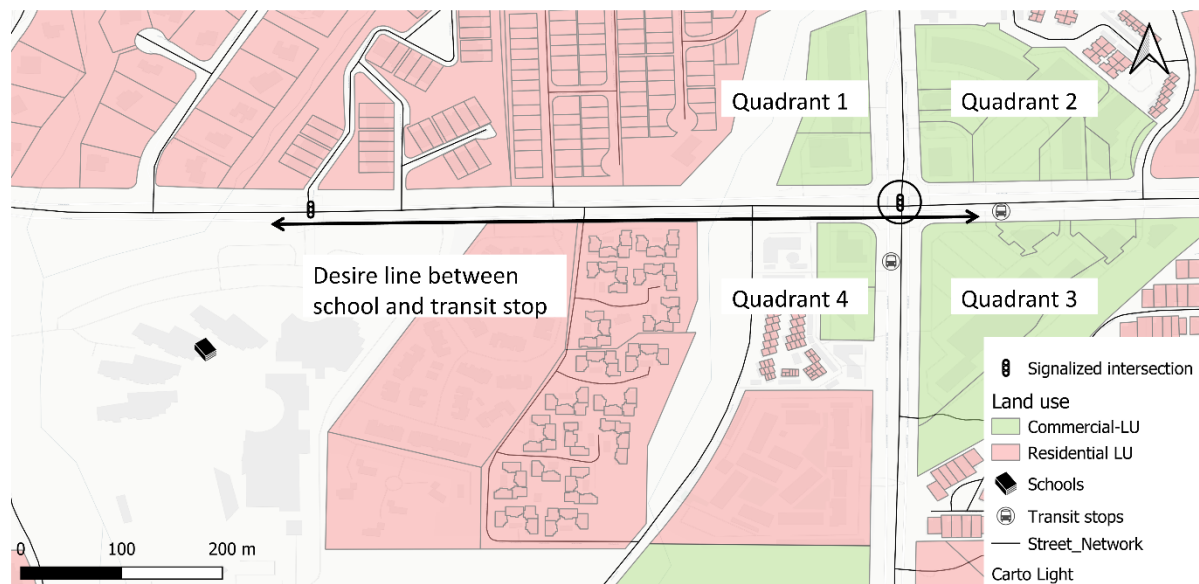


Figure 9-2 Example of desire line between transit stops and schools

9.2.1.3 Modeling shares per crosswalk

Analogous to the modeling of mode choice for urban planning processes, the goal is to estimate the “utility” associated with each crosswalk as a function of the interactions between quadrants and of crosswalk-based variables. Then, the shares per crosswalk can be calculated.

The Berkson-Theil method was used in this Chapter to model the shares per crosswalk (Aitchison, 1982; Wulansari & Astari, 2018). This approach is particularly useful when the dependent variables are shares or percentages. While the calibration was based on a dataset of four-way intersections, the method could, with some adjustments, also be applied to intersections with a different number of legs. The first step is the designation of one of the crosswalks as the reference crosswalk (in this work, the north crosswalk was arbitrarily selected as the reference). Then, the dependent variable (i.e., shares per crosswalk) and the explanatory variables are transformed to be represented in terms of the comparison to the reference crosswalk (Equations 9.4 and 9.5). Each intersection in the dataset generates three

observations (i.e., east vs. north, south vs. north, and west vs. north). After that, a linear regression is calibrated using the transformed variables (Equations 9.4 and 9.5) to represent the utility associated with each crosswalk (Equation 9.6). Note that the model coefficients obtained from the calibration ($\beta_0, \beta_1, \dots, \beta_k$) are applied to non-transformed explanatory variables to estimate shares for each crosswalk (Equation 9.7).

$$Y_{i,z-ref}^* = \frac{\ln [S(z_i)]}{\ln [S(ref_i)]} \quad (9.4)$$

$$X_{k,i,z-ref}^* = X_{k,i,z} - X_{k,i,ref} \quad (9.5)$$

$$U_{i,u} = \beta_0 + \beta_1 \times X_{1,i,u} + \dots + \beta_k \times X_{k,i,u} \quad (9.6)$$

$$P_{i,u} = \frac{e^{U_{i,u}}}{\sum_{u=N,E,S,W} e^{U_{i,u}}} \quad (9.7)$$

Where:

S = observed share for crosswalk z (east, south, or west) or reference crosswalk ref (north) at site i

$Y_{i,z-ref}^*$ = transformed dependent variable at site i

$X_{k,i,-}$ = value of explanatory variable k for crosswalk z (east, south, or west) or reference crosswalk ref (north) at site i

$X_{k,i,z-ref}^*$ = transformed explanatory variable k at site i

β = model coefficients

$X_{k,i,u}$ = value of explanatory variable k at crosswalk u (north, east, south, or west) at site i

$U_{i,u}$ = utility associated with crosswalk u at site i

$P_{i,u}$ = predicted share at crosswalk u at site i

As discussed in the previous section, the interaction between eight different pairs of categories was considered in the modeling. In addition, different attributes can represent a given category (Table 9-1), and five different radii are also tested. As a result, millions of different sets of variables can be generated, but testing all of them is not practical. Consequently, a simulation process was adopted to identify the attributes and radii that are of most interest. Firstly, the attributes and radii to represent each category (Table 9-1) are randomly selected. Then, the model is calibrated, and the significant explanatory variables (p -value < 0.05) from a backwards stepwise regression calibration process are identified. The simulation is repeated 30,000 times, and the three most frequent combinations of attributes and radii for each of the eight category interactions considered were taken as candidates (e.g., for the interaction between commercial land use and transit stops the three most frequent statistically significant combinations of attributes were: a) CP ($r=400m$) & TS ($r=200m$); b) CP ($r=400m$) & TS ($r=100m$); and (3) CA ($r=200m$) & TS ($r=200m$)). With only three alternatives to represent each category interaction between quadrants, all the possible combinations were tested, and the model with the best performance was found – as discussed in the Results section of the Chapter.

9.2.2 Estimating the percentage share of pedestrian volume per crosswalk from STCs

The modeling approach described in the previous section assumes that no information on the distribution of pedestrian volumes across crosswalks is available at a given intersection; and therefore, the LU model relies solely on attributes of the intersection, crosswalk, and nearby land use. In the STC method, it is assumed that counts from one or a few days are available at a given intersection permitting the calculation of point estimates of the shares per crosswalk.

The first step of the method to obtain the shares per crosswalk from STCs is defining “candidate” days to be polled. Normally, jurisdictions obtain STCs on “typical” days in terms of vehicular and pedestrian activity which usually means collecting STCs on: a) Tuesdays, Wednesdays, or Thursdays; b) non-holidays; and c) months that are not affected by severe weather conditions or school holidays. The last condition varies according to the jurisdiction and is detailed in the results section. Another factor is that STCs are frequently collected over just 6, 8 or 12 hours of the day. In this work, we consider only 8-hour counts covering the AM peak (7am-9am), mid-day period (11am-2pm), and PM peak (3pm-6pm) periods.

This study makes use of empirical data from 4-legged signalized intersections where CCs are available. The observed AADPT per crosswalk was calculated from a time series of one year (Equation 9.1) of recorded pedestrian volumes. Since CC are available for all sites, STCs are also available for all sites. Candidate STCs were identified by randomly sampling from the year of counts for each site considering the criteria defined in the previous paragraph and used for the calculation of point estimates of shares per crosswalk. We considered scenarios in which (k), the number of STCs available at each site, ranges from one to ten. When $k = 1$, all the candidate STCs were used for estimating the shares per crosswalk (i.e., each day represents one observation). When $k > 1$, k candidate STCs were randomly sampled from the set of candidate STCs, then the pedestrian volume at each crosswalk was summed for the k days, and the shares per crosswalk were calculated. For each k , 100 random samples were generated.

It is worth noting that the candidate STCs are from the same period that the observed AADPT per crosswalk was averaged. In practice, STCs are rarely collected at each site on an annual basis, and therefore it is common to use STCs that were collected one or more years ago to estimate the shares for the current (or future) year(s). Therefore, our evaluation represents an upper bound on the accuracy of using STCs to estimate the shares per crosswalk and the accuracy in a real-world application would be worse.

9.2.3 Evaluation method

As discussed in the initial section, a two-step procedure is required for estimating AADPT per crosswalk at sites where CC stations are not available. The first step is the estimation of AADPT for the entire intersection by expanding STCs or using DD models. The second step is the estimation of shares per crosswalk using the proposed LU model or STC method. Each step introduces error to the final AADPT per crosswalk estimate. The focus of this work is to compare the performance accuracy of the proposed LU model and the STC method with a naïve allocation of the pedestrian volume per crosswalk assuming equal percentage shares (i.e., 25% each). In practice, the method used to estimate AADPT for the intersection will vary depending on the data available in the jurisdiction (i.e., CCs vs STCs vs no counts) and consequently, the accuracy of the AADPT for the intersection will also vary. To avoid including error associated with the estimation of AADPT, observed AADPT for the entire intersection (Equation 9.2) was used as the “starting point” for the evaluation of the predicted shares using the proposed LU model and STC method.

The predicted AADPT per crosswalk is calculated using Equation 9.8. The performance of the predictions is measured using the indicator presented in Equation 9.9. It represents the percentage of pedestrian volume that is wrongly assigned to the crosswalks. The proposed LU model yields a single value of $E_{i,r}$ for each site (i.e., $r=1$), whereas the STC method generates 100 different values, one for each randomly selected set of STCs (i.e., $r=1,2,3,\dots,100$).

$$AADPT'_{i,u} = P_{i,u} \times AADPT_i \quad (9.8)$$

$$E_{i,r} = \frac{\sum_{u=N,E,S,W} |AADPT'_{i,u,r} - AADPT_{i,u}|}{AADPT_i} \quad (9.9)$$

Where:

$AADPT'_{i,u,r}$ = estimated AADPT at crosswalk u at site i

$AADPT_{i,u}$ = AADPT at crossing u (u = north, east, south, or west) of intersection i
calculated using Equation 9.1

$AADPT_i$ = AADPT of intersection i calculated using Equation 9.2

$P_{i,u}$ = predicted share at crosswalk u at site i

$E_{i,r}$ = error metric at each site i for the replication r (r is equal to 1 for the naïve method and LU model and ranges from 1 to 100 for the STC method)

Two tests were conducted to inspect whether the LU model and STC method is statistically superior to the naïve predictions. The first is a one-tailed paired hypothesis test to examine if $E_{i,r}$ decreases by using the LU model or STC method. The average of $E_{i,r}$ across replications was used for the STC method. The second test is called likelihood of improvement and measures the percentage of sites (or replications for the STC method) in which an improvement in $E_{i,r}$ is observed from either the LU model or the STC method versus the naïve method. Both tests are complementary and have their unique strengths and limitations. The hypothesis test can measure the magnitude of average improvement but fails to represent the practical reality of the STC method. The test is performed in terms of the average of $E_{i,r}$ per site. However, in practice, a single observation (i.e., one or a few STCs) would be available for each site – we do not have the luxury of repeated samples of different STCs. Consequently, we are interested in how likely our estimate from a random STC will be more accurate than the naïve model estimate. The likelihood of improvement test considers each observation individually but does not consider the amount by which the model estimate is better or worse than the naïve method.

9.3 Results and discussion

Pedestrian volume data from October 2021 to September 2022 from three jurisdictions were used in this work: Cities of Milton and Toronto in Ontario Canada; and Pima County in Arizona USA. Pedestrian count data were collected by permanently deployed camera-based counting systems that use image processing techniques to automatically record traffic counts by turning movement and pedestrian counts by crossing. Disaggregate raw count data were aggregated to 24 hour counts for input to Equations 9.1 and 9.2 and to 8 hours counts to be used as candidate STCs.

Information from different jurisdictions was combined to increase our sample size and capture a broader range of AADPTs and intersection and crosswalk characteristics. For the LU model, however, this decision requires the relationship between shares per crosswalk and explanatory variables (i.e., model coefficients) to be similar across jurisdictions. We believe this assumption is valid for two reasons: a) although it has been demonstrated that the association between the AADPT for the entire intersection and land use/socioeconomic features (i.e., DD model coefficients) varies across jurisdictions (Sobreira & Hellinga, 2023a – Chapter 3); phenomenologically, it is not expected that the distribution of the AADPT per crosswalk will be specific for a given jurisdiction. Take the example of an intersection with the same surroundings in two different jurisdictions with different overall levels of pedestrian activity (e.g., modal share associated with walking). It is expected that the AADPT for the entire intersection will be different across the two jurisdictions, but that the proportion of AADPT per crosswalk (i.e., shares) will be similar. b) The jurisdiction was included as a categorical explanatory variable in our model, and it was not significant.

Table 9-3 presents a summary of the average and standard deviation associated with pedestrian volume features (AADPT and shares per crosswalk) and radius- and crosswalk-based explanatory variables for the empirical dataset. For the radius-based variables, the statistics were collected considering quadrants generated from a 400m radius buffer. Some points in Table 9-3 are worth highlighting:

Table 9-3 Descriptive statistics

Variable	Unit	Milton (n = 18)		Pima County (n = 13)		Toronto (n = 35)		Source
		Avg.	SD	Avg.	SD	Avg.	SD	
Pedestrian volume								
AADPT for the entire intersection	Volume per day	377	230	98	35	11,127	7,241	All: Jurisdiction
Share per crosswalk (observed value)	%	25.0	12.7	25.0	17.0	25.0	7.6	All: Jurisdiction
Radius-based (quadrant – 400m)								
Commercial area	Sq-meters	5,534	9,494	19,011	20,519	35,881	22,284	M: Environics (2020) PC and T: ODP - Parcels LU
Commercial places	Count	12.4	19.9	5.4	6.2	62.9	57.3	M: Environics (2020) PC and T: ODP - Parcels LU
Employment density	Employment per sq-km	1,686	2,594	818	1,427	35,499	82,278	M and T: Environics (2020) PC: LEHDS (2019)
Residential area	Sq-meters	64,258	45,365	46,561	34,594	40,898	26,170	M: Open Street Maps (2022) PC and T: ODP - Parcels LU
Residential places ¹	Count	117.8	99.4	39.2	43.9	119.3	105.5	M: Open Street Maps (2022) PC and T: ODP - Parcels LU
Population density	Population per sq-km	2,939	2,089	1,308	761	11,054	7,510	M and T: Census (2016) PC: Census (2015-2019)
Schools - All types	Count	0.11	0.32	0.08	0.27	0.52	0.73	All: ODP - Schools
Schools - High and post secondary	Count	0.04	0.20	0.02	0.14	0.27	0.57	All: ODP - Schools
Transit stops	Count	1.93	1.57	1.15	0.89	4.35	2.44	All: ODP - Transit stops
Crosswalk-based								
Crossing width	Meters	20.59	5.46	27.76	8.05	19.68	6.42	All: Google Maps
Presence of any physical median	0 = No; 1 = Yes	0.44	0.50	0.60	0.49	0.26	0.44	All: Google Maps
Presence of refuge island	0 = No; 1 = Yes	0.01	0.12	0.51	0.50	0.04	0.19	All: Google Maps
Number of paved sidewalks at the crosswalk corners	From 0 to 4	3.78	0.45	1.84	1.50	4.00	0.00	All: Google Maps
Presence of slip lane	From 0 to 2	0.00	0.00	0.06	0.23	0.01	0.12	All: Google Maps
Road class: Major arterial, minor arterial, local or collector	Class as a binary variable	na	na	na	na	na	na	All: ODP - Road network
Desire line: transit stop and school	0 = No; 1 = Yes	0.07	0.26	0.02	0.15	0.16	0.37	All: ODP - Transit stops & Schools
Desire line: transit stop and high/postsecondary school	0 = No; 1 = Yes	0.01	0.12	0.01	0.11	0.08	0.27	All: ODP - Transit stops & Schools

Note: Avg. = average; SD = standard deviation; ODP = Open Data Portal; LU = land use; LEHDS = Longitudinal Employer-Household Dynamics Survey; na = not applicable; M = Milton; PC = Pima County; T = Toronto. ¹Multi-unit residential buildings were counted as one place. ODP data was collected in 2022

1. The AADPT for the entire intersection varies significantly across jurisdictions. Some sites in Toronto are in Downtown, where pedestrian activity is extremely high (AADPT ~ 20,000-30,000), resulting in an increased AADPT average for the jurisdiction. Conversely, some sites in Pima County present very low pedestrian activity (AADPT ~ 10-30). Sites with an AADPT lower than 50 were removed from the analysis to avoid using sites where the daily variability of the shares per crosswalk is high due to very low pedestrian activity.
2. The standard deviation of the shares per crosswalk varies across jurisdictions. The distribution of pedestrian volume per crosswalk in Toronto is generally closer to an equal division than in Pima County. Given this information, it would be expected that – on average – the variability of the attributes across quadrants of a given intersection would be higher in Pima County than in Toronto. This was checked by taking the coefficient of variation (COV) of each variable at each intersection (i.e., the standard deviation of a given attribute in the four quadrants divided by the average of the attribute in the four quadrants). After averaging the COVs across each variable, COVs of 1.08 and 0.76 were obtained for Pima County and Toronto, respectively, confirming the aforementioned expectation
3. The magnitude of the radius-based variables is different across jurisdictions. To address this, the radius-based variables were scaled from 0 to 1 within each jurisdiction so that compatible magnitudes are used in the modeling process.

Regarding the months used for determining candidate STCs for the sites in Ontario, December, January, February, and March were removed due to the high likelihood of inclement (snow and freezing temperatures) weather, and July and August were removed due to school holidays. For the sites in Pima County, August was removed due to inclement (hot) weather, June and July due to school holidays, and December because STCs are not typically collected in this month.

9.3.1 Optimal land use model

Table 9-4 presents the significant explanatory variables and their model coefficients and p-value for the LU model that exhibited the best performance in terms of the average of the error metric ($E_{i,r}$). The F-test showed that the model fits the data better than the model with no explanatory variables (p-value = 4.32×10^{-16}). The variance inflation factor (VIF) was lower than 1.15 for all the variables, and the linear correlation among the explanatory variables in both transformed and non-transformed forms was in general close to zero, presenting a greatest absolute value of 0.42. These suggest that no multicollinearity issues are observed in the model. The performance of the model is discussed in the next section.

Table 9-4 Significant explanatory variables

Variable	Coefficients		
	Estimate	Std. error	p-value
Intercept	0.0066	0.0565	0.908
Force commercial & transit stops: Commercial places (400m) & Transit stops (200m)	2.9293	0.4271	< 0.000
Force population & schools: Residential area (400m) & high/postsecondary schools (800m)	1.0076	0.2335	< 0.000
Force population & transit stops: Population density (400m) & transit stops (800m)	1.0155	0.2927	0.001
Desire line: transit stop and high/postsecondary school	0.6938	0.1997	0.001
Number of paved sidewalks at the crosswalk corners	0.2972	0.1093	0.007
Crosswalk is on a collector road	0.4080	0.1393	0.004
Presence of slip lane	-1.1804	0.5373	0.029
Crosswalk is on a local road	0.7733	0.3832	0.045
Model diagnostics			
Sample size	66		
Null log-likelihood	-265.46		
Full model log-likelihood	-218.94		
McFadden pseudo- R^2	0.175		
VIF for all variables < 1.15			

Four variables that represent interactions between quadrants are present in the model. This means, for example, that the interaction between commercial places and transit stops in adjacent quadrants is an important feature to explain the use of a given crosswalk. It is hard to interpret the magnitude of the "force" variables because they were scaled before being added to the model. However, the variable that represents the existence of a desire line between a transit stop and a high/postsecondary school is categorical (i.e., 0 = does not exist; 1 = exists); therefore, the coefficient of 0.6938 means that – holding the other variables constant – the existence of the desire line increases the odds of using a crosswalk by two times (this is calculated from $e^{0.6938}$). The signs of the coefficients are aligned with expectations: the utility of a given crosswalk increases as the variables increase.

An interesting finding is that creating a variable with high and postsecondary schools (instead of all types of schools) was beneficial for explaining the pedestrian activity. Phenomenologically, it is expected that these types of schools will generate more pedestrian trips than elementary schools due to the age of the students (students attending high school and post-secondary school are older than students at junior and middle (elementary) schools and are more likely to walk to/from school and/or to leave the school property during breaks than younger students. However, it is not common to observe this distinction in the DD models for entire intersections available in the literature.

Regarding the crosswalk-based variables, it is observed that crosswalks on collector or local roads are associated with increased pedestrian activity compared to crosswalks on major or minor arterials. One hypothesis for this is that arterials are often the "natural path" for pedestrian trips; consequently, crossing collector and local roads that intercept the main road is an inherent part of the trip. Another variable that, when increased, leads to an increase in the utility of a crosswalk is the number of paved sidewalks at the crosswalk corners. This variable captures particularities in Pima County given that almost all the crosswalks in Milton and Toronto present four paved sidewalks (Table 9-3). The presence of a slip lane is the only variable with a negative sign, meaning that the presence of one slip lane reduces the odds of using a crosswalk by 0.31 times.

9.3.2 Evaluation: LU model and STC method

Figure 9-3 compiles the performance of the three approaches considered in this study: naïve method (i.e., equal shares of 25%), LU model, and STC method with between 1 to 10 days of STCs. The y-axis represents the error metric ($E_{i,r}$ – Equation 9.9). The boxplots were constructed based on the information of every replication of every site, meaning that the number of observations is 66 for the naïve method and LU model (a single observation for each site), 5,102 for the 1-day STC method (all candidate STCs were considered once), and 6,600 for the 2- to 10-day STCs (number of sites times 100 replications). The lower and upper whiskers of the boxplots represent, respectively, the minimum and maximum values observed in the dataset that are not classified as outliers based on the ± 1.5 times interquartile criteria. Table 9-5 summarizes the data presented in Figure 9-3 in terms of the average and 95th percentile for each method.

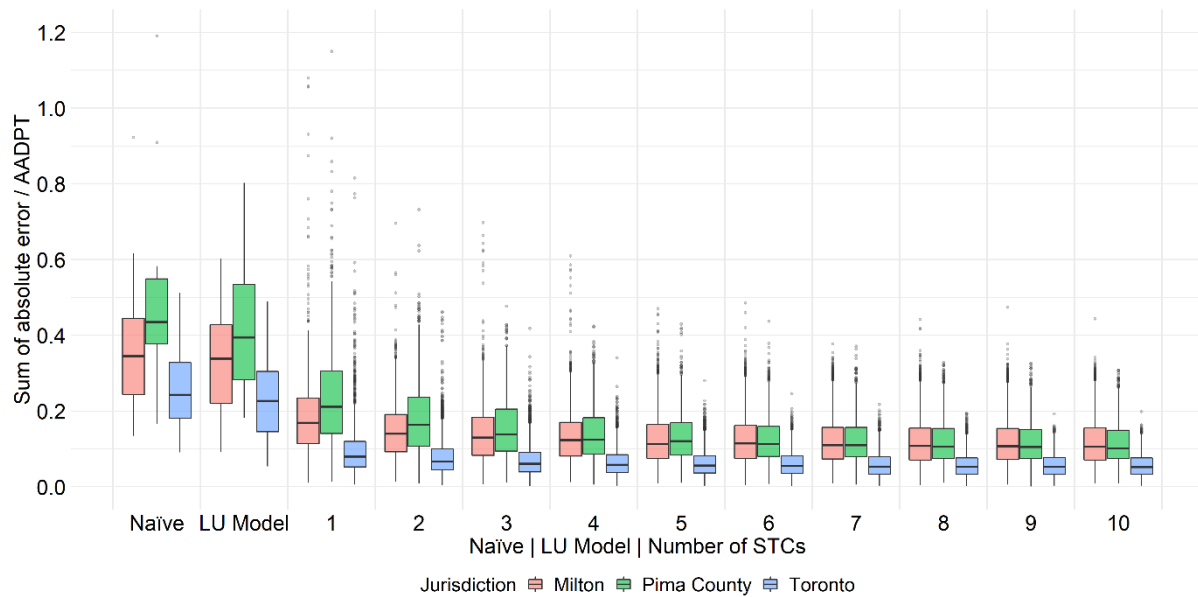


Figure 9-3 Comparison of the performance of each method

Table 9-5 Summary of the performance of each method

Jurisdiction	Number of sites	Method – Statistics of error indicator (<i>E</i>)									
		Naïve		Model		STC: 1 day		STC: 2 days		STC: 3 days	
		Avg.	95th perc.	Avg.	95th perc.	Avg.	95th perc.	Avg.	95th perc.	Avg.	95th perc.
Milton	18	0.385	0.661	0.331	0.571	0.185	0.356	0.149	0.293	0.140	0.280
Pima County	13	0.500	1.022	0.442	0.758	0.236	0.460	0.179	0.350	0.154	0.298
Toronto	35	0.257	0.439	0.233	0.410	0.098	0.224	0.078	0.170	0.070	0.148
All	66	0.340	0.610	0.301	0.592	0.153	0.368	0.117	0.277	0.106	0.249

Figure 9-3 and Table 9-5 show that the LU model is slightly better than the naïve method. Visually, all three main elements of the boxplots (1st and 3rd quartiles and median) of the three jurisdictions are improved by using the LU model. The proposed LU model provides an improvement of 11.4% compared to the naïve method in the average error across all sites. A paired hypothesis test was conducted and showed that the average improvement is statistically significant ($p\text{-value} = 0.0495$). The average error of 0.301 obtained from the LU model means that, on average, 30% of the AADPT of the entire intersection is wrongly allocated to crosswalks.

It is observed that estimating the shares per crosswalk using a single STC provides error considerably smaller than both the naïve method and the LU model, providing average improvements of 54.9% and 49.1%, respectively. Both improvements are statistically significant ($p\text{-value} < 0.000$). Concerning the number of STCs used, it is seen that both the magnitude and variability of the estimation errors are reduced when increasing from one to two STCs and from two to three STCs. The benefits of having estimates from more than three STCs are minimal (the average of the error indicator is 0.106 for three STCs and 0.085 for ten STCs).

Figure 9-4 presents the likelihood of improvement test conducted to measure the percentage of sites/replications in which applying the model or the STC method is beneficial (i.e., the error is reduced) compared to the naïve method. The results indicate there is a 62%

likelihood that the application of the LU model provides more accurate estimates than the naïve method. Similarly, there is an 88% likelihood that the one-day STC method provides more accurate estimates than the naïve method. Using STCs from multiple days increases the likelihood that the estimate will be more accurate than the naïve model but the marginal benefit decreases. Using STCs from 3 days provides a likelihood of 95.2% that the estimate will be more accurate than the naïve method. Using STCs from 10 days only increases the likelihood to 97.5%.

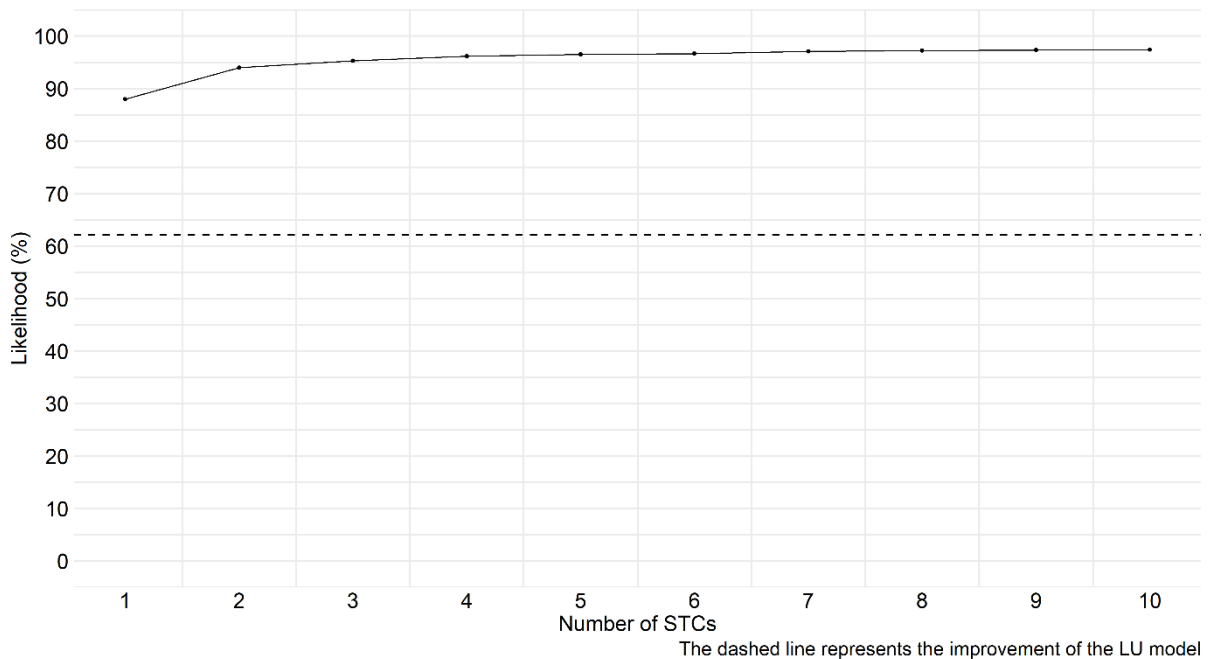


Figure 9-4 Likelihood that STC method estimate is more accurate than naïve method estimate

The performance of the LU model is examined in more detail. Figure 9-5 shows the observed and predicted shares per crosswalk in each jurisdiction. As expected, the observed shares in Toronto are less dispersed and more concentrated around 25% than in Milton and Pima County. In general, the model performs poorly for extreme cases of observed shares: it

overestimates when observed shares are lower than 10% and underestimates when observed shares are greater than 60%. Some of these cases were visually inspected, allowing the identification of the attributes responsible for generating such low or high shares (e.g., the absence of commercial land use or the presence of a desire line between school and transit stop); however, the model was not able to predict very low or high shares even though these attributes were captured by it. One hypothesis for this is that most of the sites used in the calibration are from Toronto (35 out of 66), and the dispersion of the shares in Toronto is considerably lower than in Milton and Pima County. Therefore, although we initially assumed (and expected) that the model coefficients should be similar across jurisdictions, small differences between jurisdictions may exist and could have biased the model towards Toronto particularities. In addition, some of the sites in Toronto are in Downtown, an area with extremely high pedestrian activity where the allocation per crosswalk may follow a different pattern than in the rest of the city or other jurisdictions. The jurisdiction and the AADPT for the entire intersection were considered as explanatory variables, but they turned out to be non-significant.

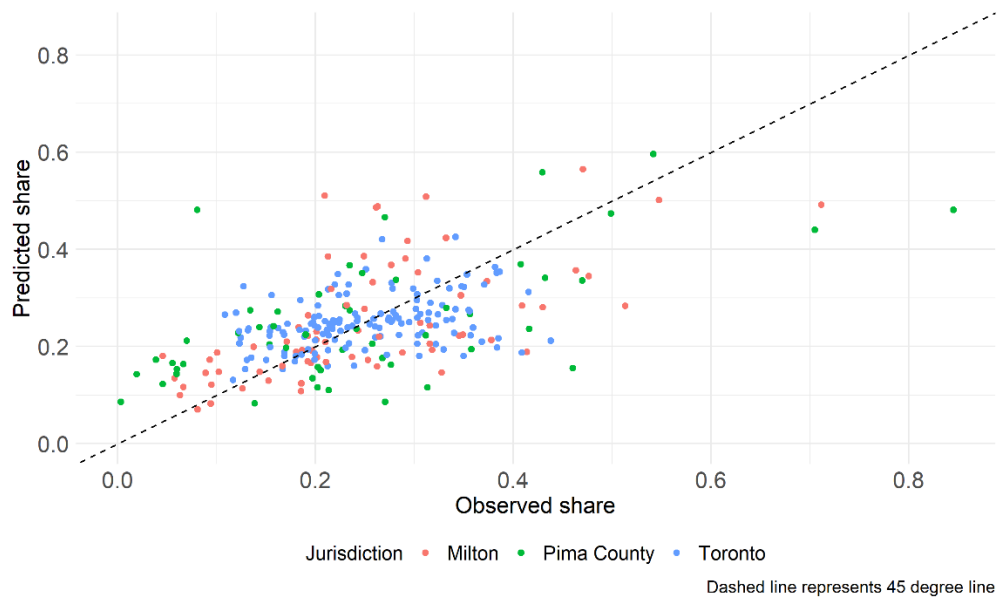


Figure 9-5 Scatterplot: observed and predicted shares per crosswalk

The small (but statistically significant) benefits of using the proposed LU model instead of the naïve method may initially discourage practitioners and researchers from investing time to try to model the shares per crosswalk. However, as the model would be applied to intersections where no pedestrian data is available and a DD model would be calibrated or transferred from another jurisdiction to estimate the AADPT for the entire intersection, almost all the variables used in this study would have already been compiled. Therefore, efforts to develop a similar model would be minimized.

Two final points are worth noting about the amount of error associated with each method. The first is that some additional error may have been introduced in the LU model due to the use of pedestrian volume data and explanatory variables from different periods of time (Table 9-3). Outdated variables – especially those related to land use and transit stops – would impact the quality of the model. In our work, these variables were collected from Open Data Portals that are maintained by each jurisdiction, and no information on the quality of the data or the frequency with which it is updated is available. The second point is that the amount of error obtained from the STC method is likely to be underestimated. The predicted shares per crosswalk were estimated using STCs from the same one-year period in which the observed (true) share was averaged. In practice, STCs are unlikely to be available for the current year and would typically be between one to five years old, which would increase estimation errors if the shares per crosswalk changed over time.

The use of STCs from previous years to estimate the shares per crosswalk was investigated. The City of Toronto provides historical STCs in their Open Data Portal, so STCs from 2015 to 2020 were collected for the same sites used in the original evaluation (i.e., CCs are available). 86 STCs from 32 sites were available. The error indicator (E) was computed using the shares per crosswalk obtained from the previous year STCs. For comparison purposes, each observation from the previous year STC was paired with an observation from the current year (i.e., Oct 2021-Sep 2022) considering the same day-of-week-of-month, so any seasonality present in the crosswalk shares could be controlled for. For example, an STC collected at site i on a Tuesday in April of 2018 was paired with an

STC extracted from the CC dataset at site i on a Tuesday in April of 2022. The average error of the previous and current year STCs is 0.174 and 0.087, respectively; representing an average increase of 100% in the error when STCs from previous years are used. The LU model provided an average error of 0.233 in Toronto, meaning that the STC method is still superior to it even with the use of previous year STCs.

We hypothesized that the “oldness” of the STC affected the estimates, as it would be expected that older STCs would provide worse performance because of changes in pedestrian activity patterns, but we were not able to identify a statistically significant relationship between age of the STC and the crosswalk share estimation accuracy. Our ability to carry out this analysis was limited by the small sample size for each year, and it is possible that statistically significant relationships could be developed with a larger data set. Furthermore, the small sample size precluded the examination of error for more than one STC.

One challenge with our analysis is that the COVID pandemic occurred between the previous and current year periods. We know this significantly impacted pedestrian activity levels (i.e., AADPT), but we do not know the extent to which the pandemic impacted the shares per crosswalk. Most importantly, we do not know the relative contribution of the pandemic versus other influences (e.g., changes in land use, etc.) on the shares per crosswalk. This makes it somewhat difficult to predict the level of degradation in the accuracy of estimating shares per crosswalk associated with using previous years’ STCs. It seems likely that the degradation found in this analysis (i.e., error increases by 100% when using STCs between 1- and 5-year-old) represents an upper bound. Furthermore, if we assume that these results from the City of Toronto data can be applied to other jurisdictions, then we can estimate the lower and upper bounds on the average error from the STC method using 1 day of counts (Table 9-6). The results in Table 9-6 indicate that the level of degradation in estimation accuracy of the STC method when using STCs from previous years can be quite large, such that the performance becomes worse than the proposed LU model.

Table 9-6 Average estimation error of each method considering upper and lower bounds on STC method accuracy

Jurisdiction	Average Error Indicator				% Improvement vs Naïve			% Improvement vs LU	
	Naïve	LU Model	STC LB	STC UB	LU Model	STC LB	STC UB	STC LB	STC UB
Milton	0.385	0.331	0.185	0.370	13.9%	51.8%	3.7%	44.1%	-11.8%
Pima County	0.500	0.442	0.236	0.471	11.5%	52.8%	5.6%	46.7%	-6.7%
Toronto	0.257	0.233	0.098	0.196	9.4%	61.9%	23.8%	58.0%	15.9%
All	0.340	0.301	0.153	0.306	11.4%	54.9%	9.8%	49.1%	-1.8%

Note: STC LB = lower bound estimate of error for STC 1 day method (results from Table 9-5); STC UB = upper bound estimate of error for STC 1 day method (results from applying 100% increase in error found from City of Toronto analysis using STCs from previous years)

9.4 Conclusions and recommendations

This Chapter examined the estimation of the percentage shares of pedestrian volume in each crosswalk of an intersection. Two different methods were considered. The first method assumed that STCs were available at the intersection, allowing the point estimation of shares per crosswalk. The second method presumed that no pedestrian volume data is known at the intersection, and a LU model to estimate the shares per crosswalk was developed. Sites where CC stations were available were used in the analysis.

The rationale for the LU model is that the shares per crosswalk can be estimated based on the interaction between land use, socioeconomic, and operational information in adjacent quadrants and on the geometric and operational attributes of each crosswalk. The interaction between commercial places & transit stops, residential area & high/postsecondary schools, population density & transit stops, and transit stops & high/postsecondary schools showed to be important for explaining the shares per crosswalk and were statistically significant in the model. In addition, the number of paved sidewalks at the crosswalk corners, crosswalks located on local or collector roads, and the presence of slip lanes were also significant in the model.

The performance of the LU model and STC method was measured in terms of the fraction of the observed (or true) AADPT of the entire intersection that was wrongly allocated to the crosswalks. A naïve estimate assuming equal shares per crosswalk was used as a benchmark. The use of the LU model generated an average wrong allocation of 0.301, representing a statistically significant improvement of 11.4% when compared to the naïve method. The use of STCs for estimating the shares per crosswalk was more accurate than the LU model when using STCs from the current analysis period. The use of an STC from a single day provided an average wrong allocation of 0.153, meaning an improvement of 54.9% from the naïve estimate. Using STCs from two or three days resulted in average performance indicators of 0.117 and 0.106, respectively. The marginal benefits of using STCs from more than three days are minimal. However, this analysis is based on using STCs from the current analysis time period. In practice, STCs are likely to be between 1 and 5 years old. Analysis using a limited set of STCs from previous years showed that estimation errors in practice may be as much as twice as large as the errors reported above.

In summary, this research demonstrates that there are methods available for practitioners to estimate the pedestrian exposure per crosswalk and that different levels of estimation accuracy of shares per crosswalk are observed depending on the type of information available. When STCs are available from one or more days then the recommendation is to compute crossing shares from these STCs. The marginal benefit of using STCs from more than three days is marginal so it is recommended that when STCs are available from more than three days, that the three most recent days be used. When no count information is available at the intersection, but the jurisdiction does have data from a number of continuous count sites at other intersections, then it is recommended the jurisdiction develop a direct demand model following the approach used in this Chapter to estimate the crossing shares.

Naturally, this raises the question of what a jurisdiction should do if it does not have enough continuous count sites or does not have the resources to develop their own direct demand model. A potential solution is to apply a direct demand model that has been

developed in another jurisdiction but at present the spatial transferability of the LU model proposed in this Chapter has not been examined. Consequently, a proposed area for future research is to examine the spatial transferability of the proposed LU model. Other areas requiring further research include quantifying the impact of “oldness” of STCs on the accuracy of the STC method on a larger data set.

There are several limitations to the research in this Chapter. One important limitation is that the LU model was calibrated using information from three different jurisdictions. Phenomenologically, it would be expected that the association between shares per crosswalk and explanatory variables (i.e., model coefficients) would be relatively similar across jurisdictions, so that combining different jurisdictions would not affect the quality of the model. However, this expectation could not be verified due to the small sample size when jurisdictions are considered individually. This may have impacted the performance of the model and requires further investigation. A minor limitation is the use of pedestrian count data automatically extracted from video recordings. The level of error associated with these counts is not known; however, assuming that the amount of error is similar across the crosswalks of the same intersection, it is expected that this error will have minimal effects on the shares per crosswalk.

Chapter 10

Conclusions and Recommendations

Estimating pedestrian exposure at intersections is essential for developing pedestrian-centric interventions. Pedestrian exposure is typically used as an input for road safety analyses and site prioritization. A jurisdiction's ability to estimate pedestrian exposure at all intersections depends on the quantity and type of pedestrian volume data available (i.e., CC stations, STCs, or no data). This Thesis explored the application of methods under different data availability conditions, developed approaches for jurisdictions with limited pedestrian volume data, and proposed methods to disaggregate pedestrian volume from the intersection level to the crosswalk level. This Chapter highlights the main contributions of this Thesis, followed by directions for future research.

10.1 Contributions

The main contributions of this Thesis are presented below:

1. The spatial transferability of DD models, especially when accompanied by local calibration, proved to be a useful method for jurisdictions seeking to estimate pedestrian exposure but lacking sufficient sites to develop their own DD model. Chapter 3 demonstrated that direct (or naïve) transferability leads to inconsistent performance, varying based on the characteristics of both the jurisdiction where the model was developed and where it is applied. Chapter 4 and Chapter 8 showed that local calibration using limited, readily available pedestrian volume data can improve the transferability of DD models, achieving performance levels relatively similar to those of locally developed DD models.
2. Continuing on jurisdictions that lack sufficient data to develop their own DD model, a novel method leveraging AI Large Language Models (such as ChatGPT), satellite

imagery, and pedestrian volume data from a few key sites was proposed (Chapter 5). The performance of this method rivaled – and in some cases surpassed – existing conventional methods like DD models, all without requiring complex statistical models or extensive datasets. This simplicity makes the proposed method a promising and cost-effective approach for pedestrian exposure estimation.

3. This Thesis explored in detail the application of the expansion factor method using DOWOM factors to expand 8-hour STCs. Temporal indicators accounting for different types of seasonality were developed to assist in identifying factor groups. Additionally, thresholds for these indicators are provided for practitioners seeking a simpler and more straightforward approach to identifying factor groups (Chapter 6). Furthermore, this Thesis confirmed the findings of Medury et al. (2019) that employing models to associate STC sites with factor groups yields only marginal improvements compared to using a single factor group (i.e., the average across all sites) (Chapter 6 and Chapter 7). This is particularly true when STCs from typical days (i.e., candidate days) are available.
4. The spatial transferability of expansion factors across jurisdictions with similar characteristics in terms of weather and school holiday periods proved to be of practical value, with only minor performance degradation compared to expansion factors developed within the local jurisdiction (Chapter 7). This is a significant finding, given that many jurisdictions lack CC stations to calculate “local” expansion factors (Grossman et al., 2019), and implementing CC stations is relatively costly – especially when compared to the deployment of STCs.
5. Chapter 8 investigated the application of different methods for estimating pedestrian exposure under various scenarios of data availability. The performance of these methods was assessed based on the number of sites with available data, leading to practical guidelines for practitioners. More specifically, it provides recommendations on the minimum number of sites with pedestrian volume data required for the local development of DD models, as opposed to alternative approaches such as spatial transferability and local calibration of DD models.

6. The final major contribution was the novel exploration of disaggregating pedestrian volumes from the intersection level to the crosswalk level. Land use models were proposed and compared to methods that utilize crosswalk-specific volume information from STCs. These approaches address scenarios where STCs are available as well as cases where no information is available.

10.2 Recommendations for future research

Recommendations for future research specific to each objective are presented at the end of each Chapter. Additional potential topics for further investigation are suggested below:

- Pedestrian volume is often not the "final product" of traffic engineering analyses and is frequently used as an input for other investigations, such as the development of safety performance functions (SPFs). Therefore, it is recommended to investigate how the accuracy of pedestrian volume estimates impacts the quality of its outputs, like the site ranking provided by SPFs.
- This Thesis employed expansion factors and STCs retrieved from the same time period (i.e., expansion factors were developed using a full year of data, and STCs were from within that year). In practice, however, this may not always be the case, as STCs for some sites may be from different years than those used to calculate the expansion factors. This raises the question of how the "age" of STCs would impact the performance of the expansion factor method. This concern may also extend to the application of temporal factors to adjust for potential temporal changes.
- Chapter 3 explored the spatial transferability of DD models. In general, it was observed that the models transferred relatively well to Milton, ON, but poorly to Pima County, USA, and Toronto, ON. This difference appears to be linked to the level of similarity between the jurisdiction where the model was developed and the jurisdiction where the model was applied. Milton's land use and socioeconomic characteristics fall within the

range observed in the jurisdictions where the models were developed (i.e., within the minimum and maximum values, and with relatively similar central tendencies), whereas Pima County and Toronto do not share these characteristics. In theory, DD models would adequately transfer to another jurisdiction if the relationships between explanatory variables and the dependent variable (i.e., model coefficients) are relatively similar. However, this similarity cannot be defined solely by assessing the distribution of explanatory variables between the two jurisdictions. Further exploration is needed to identify the key components (e.g., motorization rate, the intensity of pedestrian activity, and cultural attitudes toward active transportation) that can be used to assess whether a DD model developed in a given jurisdiction is suitable for transfer to another.

- This Thesis focused on methods for intersections. Similar approaches could be applied to midblock locations or segments, but these have been less studied. Midblock crossings add complexity due to factors like jaywalking and varied pedestrian behavior, which may pose additional challenges for accurate modeling and transferability of models of across different jurisdictions.

References

- AASHTO. (2010). *Highway Safety Manual*.
- Aitchison, J. (1982). The Statistical Analysis of Compositional Data. In *Source: Journal of the Royal Statistical Society. Series B (Methodological)* (Vol. 44, Issue 2).
- Allen, B. (2021). *Estimating Intersection Annual Average Daily Bicycle Traffic from 8-hour Turning Movement Counts* [Master Thesis]. University of Waterloo.
- Chen, L., Lu, Y., Sheng, Q., Ye, Y., Wang, R., & Liu, Y. (2020). Estimating pedestrian volume using Street View images: A large-scale validation test. *Computers, Environment and Urban Systems*, 81(March).
<https://doi.org/10.1016/j.compenvurbsys.2020.101481>
- Driessen, T., Dodou, D., Bazilinskyy, P., & De Winter, J. (2023). Putting ChatGPT Vision (GPT-4V) to the test: Risk perception in traffic images. *Preprint*.
<https://www.researchgate.net/publication/375238184>
- Elvik, R. (2009). The non-linearity of risk and the promotion of environmentally sustainable transport. *Accident Analysis and Prevention*, 41(4), 849–855.
<https://doi.org/10.1016/j.aap.2009.04.009>
- Ewing, R., & Cervero, R. (2010). Travel and the Built Environment. *Journal of the American Planning Association*, 76(3), 265–294. <https://doi.org/10.1080/01944361003766766>
- FHWA. (2016). *Traffic Monitoring Guide*.
- Figliozzi, M., Johnson, P., Monsere, C., & Nordback, K. (2014). Methodology to characterize ideal short-term counting conditions and improve AADT estimation accuracy using a regression-based correcting function. *Journal of Transportation Engineering*, 140(5). [https://doi.org/10.1061/\(ASCE\)TE.1943-5436.0000663](https://doi.org/10.1061/(ASCE)TE.1943-5436.0000663)
- Google. (n.d.). *Google Maps Static API*. Retrieved January 3, 2024, from <https://developers.google.com/maps/documentation/maps-static/start>

- Google Maps. (2024). *Set of images showing aerial views from different intersection in the Region of Waterloo, Canada [Google Maps images]. Retrieved from <https://maps.googleapis.com/maps/api/staticmap> in January, 2024 using the package `httr` from R.*
- Government of Canada. (2021). *Government of Canada announces the country's first-ever federal strategy and fund dedicated to building active transportation trails and pathways.* <https://www.canada.ca/en/office-infrastructure/news/2021/07/government-of-canada-announces-the-countrys-first-ever-federal-strategy-and-fund-dedicated-to-building-active-transportation-trails-and-pathways.html>
- Griswold, J. B., Medury, A., Schneider, R. J., Amos, D., Li, A., & Grembek, O. (2019). A Pedestrian Exposure Model for the California State Highway System. *Transportation Research Record*, 2673(4), 941–950. <https://doi.org/10.1177/0361198119837235>
- Grossman, A., Rodgers, M. O., Xu, Y., Guensler, R., & Watkins, K. (2019). If Safety Matters, Let's Measure It: Nationwide Survey Results for Bicycle and Pedestrian Treatment Prioritization. *Journal of Transportation Engineering, Part A: Systems*, 145(1). <https://doi.org/10.1061/jtepbs.0000205>
- Hankey, S., Lindsey, G., & Marshall, J. (2014). Day-of-year scaling factors and design considerations for nonmotorized traffic monitoring programs. *Transportation Research Record*, 2468, 64–73. <https://doi.org/10.3141/2468-08>
- Hankey, S., Lu, T., Mondschein, A., & Buehler, R. (2017). Spatial models of active travel in small communities: Merging the goals of traffic monitoring and direct-demand modeling. *Journal of Transport and Health*, 7(January), 149–159. <https://doi.org/10.1016/j.jth.2017.08.009>
- Hankey, S., Zhang, W., Le, H. T. K., Hystad, P., & James, P. (2021). Predicting bicycling and walking traffic using street view imagery and destination data. *Transportation Research Part D: Transport and Environment*, 90. <https://doi.org/10.1016/j.trd.2020.102651>

- Hernández-Vega, H., & Matamoros-Jiménez, C. (2021). Clustering Approach to Generate Pedestrian Traffic Pattern Groups. *Ciencia e Ingeniería Neogranadina*, 31(2), 41–60. <https://doi.org/10.18359/rcin.4403>
- Huang, J., Jiang, P., Gautam, A., & Saripalli, S. (2024, November 24). GPT-4V Takes the Wheel: Promises and Challenges for Pedestrian Behavior Prediction. *Association for the Advancement of Artificial Intelligence*. <http://arxiv.org/abs/2311.14786>
- 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/10.2307/48646235>
- 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). <https://doi.org/10.1061/jtepbs.0000737>
- Johnstone, D., Nordback, K., & Kothuri, S. (2018). Annual Average Nonmotorized Traffic Estimates from Manual Counts: Quantifying Error. *Transportation Research Record*, 2672(43), 134–144. <https://doi.org/10.1177/0361198118792338>
- Kim, S., Park, S., & Jang, K. (2019). Spatially-varying effects of built environment determinants on walking. *Transportation Research Part A: Policy and Practice*, 123(February), 188–199. <https://doi.org/10.1016/j.tra.2019.02.003>
- Kuzmyak, J. R., Walters, J., Bradley, M., & Kockelman, K. M. (2014). *Estimating Bicycling and Walking for Planning and Project Development: A Guidebook*. Transportation Research Board. <https://doi.org/10.17226/22330>
- Lian, T., Loo, B. P. Y., & Fan, Z. (2024). Advances in estimating pedestrian measures through artificial intelligence: From data sources, computer vision, video analytics to the prediction of crash frequency. *Computers, Environment and Urban Systems*, 107. <https://doi.org/10.1016/j.compenvurbsys.2023.102057>

- Lindsey, G., Petesch, M., & Hankey, S. (2015). *The Minnesota Bicycle and Pedestrian Counting Initiative: Implementation Study*. <http://www.lrrb.org/pdf/201534.pdf>
- Martin, A., Suhrcke, M., & Ogilvie, D. (2012). Financial incentives to promote active travel: An evidence review and economic framework. *American Journal of Preventive Medicine*, 43(6). <https://doi.org/10.1016/j.amepre.2012.09.001>
- 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*, 2673(5), 720–730. <https://doi.org/10.1177/0361198119838266>
- Miranda-Moreno, L. F., & Fernandes, D. (2011). Modeling of pedestrian activity at signalized intersections: Land use, urban form, weather, and spatiotemporal patterns. *Transportation Research Record*, 2264, 74–82. <https://doi.org/10.3141/2264-09>
- Miranda-Moreno, L., Nosal, T., Schneider, R., & Proulx, F. (2013). Classification of bicycle traffic patterns in five North American cities. *Transportation Research Record*, 2339, 68–79. <https://doi.org/10.3141/2339-08>
- Munira, S., & Sener, I. N. (2017). *Use of Direct-Demand Modeling in Estimating Nonmotorized Activity: A Meta-analysis* (Issue October). https://www.vtti.vt.edu/utc/safe-d/wp-content/uploads/2018/04/UTC-Safe-D_Direct-Demand-Model-for-PedBike_TTI-Report_12Oct17_Final.pdf
- Munira, S., Sener, I. N., & Dai, B. (2020). A Bayesian spatial Poisson-lognormal model to examine pedestrian crash severity at signalized intersections. *Accident Analysis and Prevention*, 144(December 2019), 105679. <https://doi.org/10.1016/j.aap.2020.105679>
- NACTO. (2014). *Urban Bikeway Design Guide* (2nd ed.).
- NASEM. (2022). *Highway Capacity Manual 7th Edition: A Guide for Multimodal Mobility Analysis*. The National Academies Press. <https://doi.org/10.17226/26432>

- NHTSA. (2010). *Highlights of 2009 Motor Vehicle Crashes*. <http://www-nrd.nhtsa.dot.gov/cats/index.aspx>.
- NHTSA. (2020). *Overview of the 2019 Crash Investigation Sampling System*. <https://crashstats.nhtsa.dot.gov/Api/Public/>
- Nordback, K., Kothuri, S., Johnstone, D., Lindsey, G., Ryan, S., & Raw, J. (2019). Minimizing Annual Average Daily Nonmotorized Traffic Estimation Errors: How Many Counters Are Needed per Factor Group? *Transportation Research Record*, 2673(10), 295–310. <https://doi.org/10.1177/0361198119848699>
- Nordback, K., Marshall, W. . E., & Janson, B. N. (2013). *Development of Estimation Methodology for Bicycle and Pedestrian Volumes Based on Existing Counts*. <http://www.coloradodot.info/programs/research/pdfs>
- Nosal, T., Miranda-Moreno, L. F., & Krstulic, Z. (2014). Incorporating weather: Comparative analysis of annual average daily bicyclist traffic estimation methods. In *Transportation Research Record* (Vol. 2468, pp. 100–110). National Research Council. <https://doi.org/10.3141/2468-12>
- Open AI. (2024). *ChatGPT-4 Vision*.
- Orr, A. (2021). *Estimating Average Annual Daily Pedestrian Volumes at Intersections based on Turning Movement Counts* [Master Thesis]. University of Waterloo.
- Pulugurtha, S. S., & Sambhara, V. R. (2011). Pedestrian crash estimation models for signalized intersections. *Accident Analysis and Prevention*, 43(1), 439–446. <https://doi.org/10.1016/j.aap.2010.09.014>
- Sanders, R. L., Frackelton, A., Gardner, S., Schneider, R., & Hintze, M. (2017). Ballpark method for estimating pedestrian and bicyclist exposure in Seattle, Washington: Potential option for resource-constrained cities in an age of big data. *Transportation Research Record*, 2605(1), 32–44. <https://doi.org/10.3141/2605-03>

- Schneider, R. J., Schmitz, A., & Qin, X. (2021). Development and validation of a seven-county regional pedestrian volume model. *Transportation Research Record*, 2675(6), 352–368. <https://doi.org/10.1177/0361198121992360>
- Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., ... Resnik, P. (2024). *The Prompt Report: A Systematic Survey of Prompting Techniques*. <http://arxiv.org/abs/2406.06608>
- 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(January), 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*, 2675(6), 429–440. <https://doi.org/10.1177/0361198121994126>
- Sobreira, L. T. P., & Hellinga, B. (n.d.). Spatial Transferability of Expansion Factors for Estimating Pedestrian Volume at Intersections. *Transportation Research Record, Under review*.
- Sobreira, L. T. P., & Hellinga, B. (2023a). 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*, 036119812311610. <https://doi.org/10.1177/03611981231161061>
- Sobreira, L. T. P., & Hellinga, B. (2023b). Enhancing the Spatial Transferability of Direct Demand Models for Estimating Pedestrian Volumes at Intersections. *Transportation Research Record*, 2678(7). <https://doi.org/10.1177/03611981231205881>
- Sobreira, L. T. P., & Hellinga, B. (2024). Expanding Short-Term Pedestrian Volume Counts to Average Annual Daily Counts at Intersections: Temporal Pattern Indices and a Comparison of Statistical Land Use and Simple Threshold Methods. *Transportation*

Research Record: Journal of the Transportation Research Board.

<https://doi.org/10.1177/03611981241281736>

Sobreira, L. T. P., & Hellinga, B. (2025). Estimating Pedestrian Volumes at Intersections using Artificial Intelligence: A ChatGPT Vision Approach. *Transportation Research Record: Journal of the Transportation Research Board.*

<https://doi.org/https://doi.org/10.1177/03611981241308873>

Stipancic, J., Miranda-Moreno, L., Strauss, J., & Labbe, A. (2020). Pedestrian safety at signalized intersections: Modelling spatial effects of exposure, geometry and signalization on a large urban network. *Accident Analysis and Prevention*, 134(August 2019), 105265. <https://doi.org/10.1016/j.aap.2019.105265>

StreetLight. (2023). *StreetLight Pedestrian Volume Methodology and Validation White Paper.*

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.

Transport Canada. (2009). *Canadian Motor Vehicle Traffic Collision Statistics: 2009.*

Transport Canada. (2019). *Canadian Motor Vehicle Traffic Collision Statistics: 2019.*

Turner, S., Qu, T., & Lasley, P. (2012). *Strategic Plan for Non-Motorized Traffic Monitoring in Colorado.*

Wang, L., & Currans, K. M. (2018). Detransformation Bias in Nonlinear Trip Generation Models. *Journal of Urban Planning and Development*, 144(3).

[https://doi.org/10.1061/\(asce\)up.1943-5444.0000455](https://doi.org/10.1061/(asce)up.1943-5444.0000455)

Weather Spark. (2023, December 5). *Climate and Average Weather Year Round in Waterloo.*
<https://Weatherspark.Com/y/19227/Average-Weather-in-Waterloo-Canada-Year-Round>.

WHO. (2021). Global Plan - Decade of Action for Road Safety 2021-2030. In *2021.*

- Wickham, H. (2022). *httr: Tools for Working with URLs and HTTP*. R package version 1.4.4. <https://CRAN.R-project.org/package=httr>.
- Wulansari, D. N., & Astari, M. D. (2018). Mode choice analysis using discrete choice model from transport user (Case study: Jakarta LRT, Indonesia). *MATEC Web of Conferences*, 181. <https://doi.org/10.1051/mateconf/201818103001>

Appendix A

Steps 1, 2, and 3 from Chapter 7

This appendix provides detailed information on the methods and results of Steps 1, 2, and 3 from Chapter 7.

Step 1: Identifying factor groups in the base jurisdiction

Method

The factor groups in the base jurisdiction are identified using three traffic indicators proposed in Chapter 6 (Sobreira & Hellinga, 2024). These indicators, calculated based on the DOWOM expansion factors for each site, aim to distinguish different traffic patterns:

1. Adapted Weekend-Weekday Index (WWI*): It evaluates the similarity between weekdays and weekends, as detailed in Equations 6.1 and 6.2. Equation 6.2 excludes months affected by school holidays – July and August in Ontario, and June and July in Pima County. These months are omitted because, at sites heavily influenced by school activity, weekday-weekend relationships during school holidays are likely not representative of typical patterns.
2. School Holiday-Typical Index (SHTI): this index identifies sites where pedestrian activity is heavily influenced by school schedules. It measures the pedestrian activity between school holiday periods and typical periods (i.e., months not impacted by school holidays or adverse weather), as described in Equation 6.3. In Ontario, the typical period encompasses months not affected by winter, from April to November. In Pima County, the typical period excludes months affected by extreme hot weather: August and September. Note that June and July are also affected by hot weather but are the school holiday months. The typical period excludes months with adverse weather, as pedestrian activity may be reduced during these months, making it difficult to isolate the influence of school holidays on typical pedestrian activity.

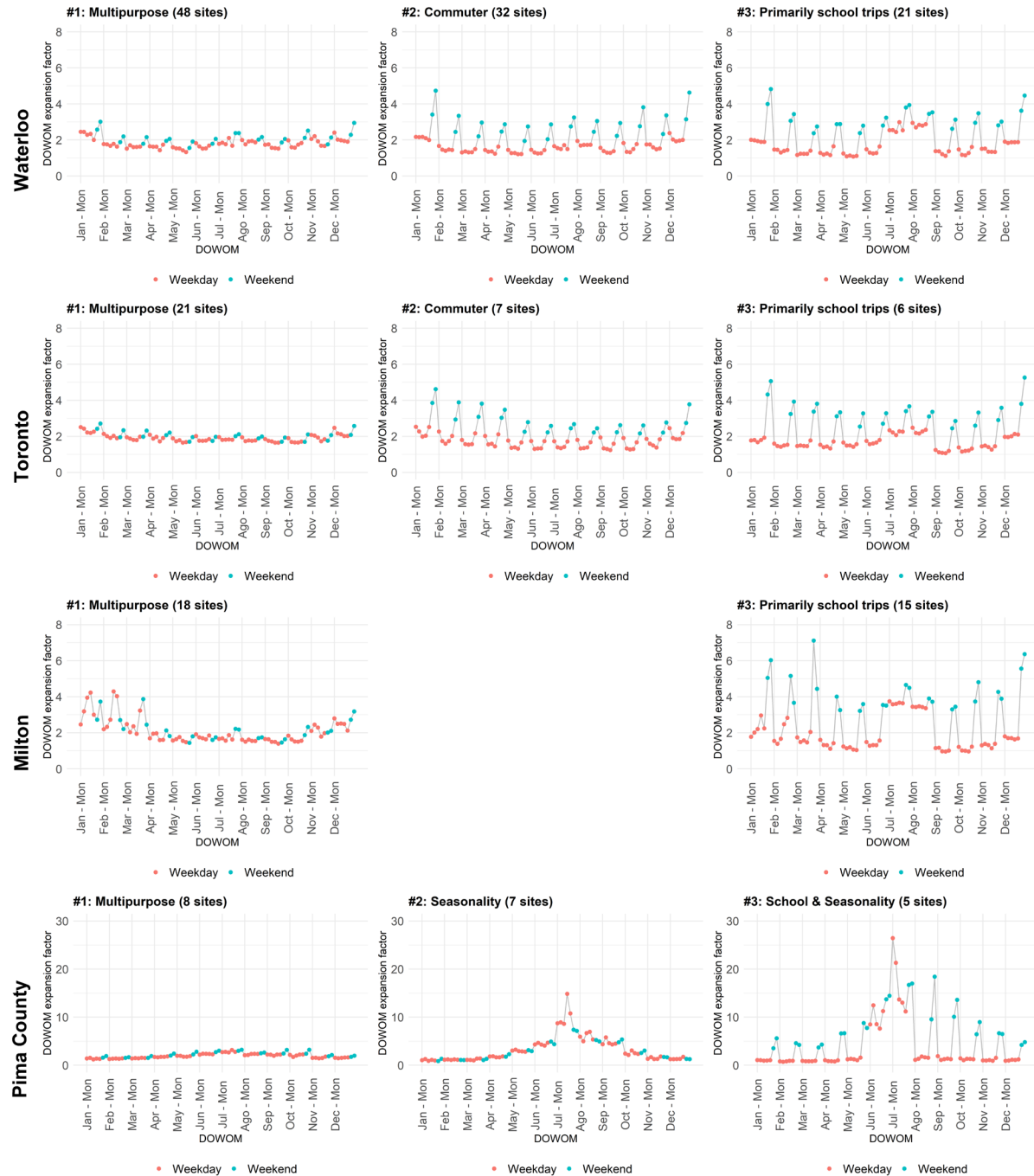
3. Deviation Across Months (DAM): this indicator measures the standard deviation across months to identify seasonality at each site (Equations 6.4 and 6.5). School-holiday months are excluded; therefore, DAM captures only seasonality unrelated to school activity. For example, it can identify sites with reduced pedestrian volume during adverse weather conditions.

The indicators are calculated for each site, and the k-means clustering algorithm is then applied to group sites with similar indicators. The optimal number of clusters (or factor groups) is determined based on two criteria: a) clustering metrics: the total within-cluster sum of squared errors, the silhouette test, and the gap statistic; and b) engineering judgment: the resulting factor groups must make phenomenological sense. In other words, the resulting groups should be – in theory – identifiable through land use, socioeconomic, and transportation attributes, allowing for the development of models in the next step.

Before applying the k-means algorithm, all three indicators were constrained to a maximum value of two to avoid outliers from biasing the clustering procedure due to large indicator values. Additionally, WWI* and SHTI were scaled from 0 to 1. We chose not to scale DAM for two reasons: a) it already ranged from 0 to 1 at almost all sites; and b) it exhibited small variations in some jurisdictions (e.g., from 0.15 to 0.35 in Toronto), so scaling it from 0 to 1 would distort its effect in the clustering process.

Results

Figure A1 illustrates the factor groups identified in each jurisdiction. Each plot in the figure represents the average (across all sites in the factor group) of the 84 DOWOM expansion factors, with dots arranged chronologically from Monday to Sunday. Vertical gridlines mark the start of each week, aligning with Mondays for each month. The color of the dots differentiates between weekdays and weekends. The y-axis is fixed from 0 to 8 in the Ontario jurisdictions and from 0 to 30 in Pima County.



Lines and points represent the average DOWOM expansion factors within factor groups

Figure A1 Resulting factor groups in each jurisdiction

Three factor groups are observed in Ontario: a) #1 Multipurpose: characterized by relatively consistent pedestrian volume across different months and between weekends and weekdays; b) #2 Commuter: marked by reduced volume on weekends; and c) #3 Primarily School Trips: showing reduced volume on weekends and during school-holiday months. Waterloo and Toronto have all three factor groups, while Milton does not include the #2 Commuter group. Visually, the expansion factors within each factor group appear to have similar magnitudes across the three Ontario jurisdiction, suggesting that these factors may be transferable across the Ontario jurisdictions.

As a comparison, in Chapter 6 five factor groups were found when performing a similar analysis in Waterloo. In addition to the three groups observed in the present work, the authors found two additional groups with characteristics similar to #1 Multipurpose and #2 Commuter but with an added seasonality effect: reduced pedestrian volume during the winter months (December-February). This difference may be attributed to the milder winter in 2023-2024 (the period used in this work) compared to the harsher winter in 2022-2023 (the period used in Chapter 6).

As expected, different configurations of factor groups are observed in Pima County. The first factor group, Multipurpose, is similar to the one observed in Ontario. The second factor group also represents a Multipurpose case (i.e., similar volume on weekends and weekdays) but with an added seasonality effect due to the hot summer temperatures. The third factor group combines both school and seasonality effects. In addition to these differences in group configuration, the magnitude of the factors is significantly higher in Pima County than in Ontario. This discrepancy is likely due to the very low pedestrian activity at the sites considered in this work (Table 7-1), rather than a characteristic of general sites in Arizona.

Figure A2 exemplifies the application of the clustering technique in Waterloo. The DAM indicator was also included in the procedure but is not shown in the figure. The clusters are well-defined, with a “dividing threshold” around 1.5 for both WWI* and SHTI.

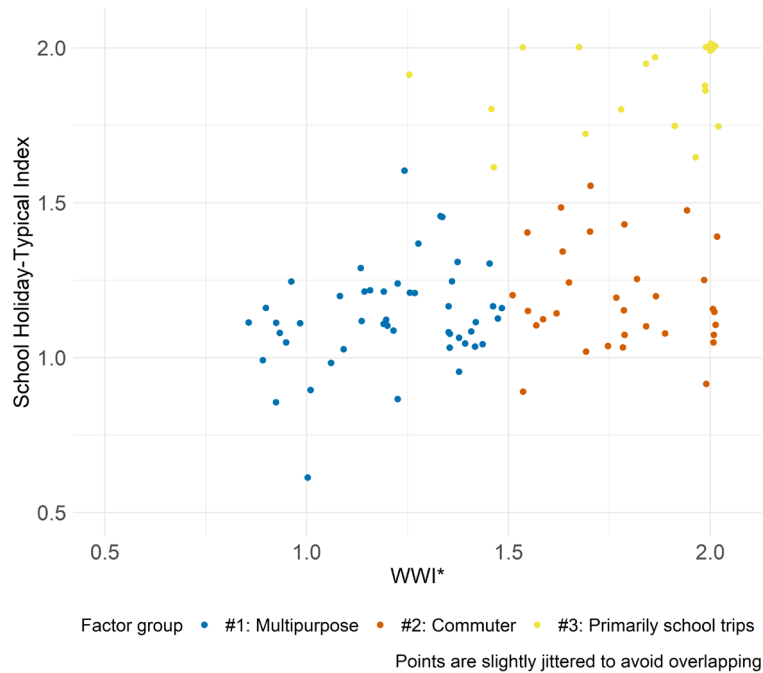


Figure A2 WWI*, SHTI and factor groups in Waterloo

The factor groups assigned to each site in this section are used as the dependent variable in the next section, where models are developed to associate sites with factor groups.

Step 2: Modeling factor groups in the base jurisdiction

Method

Method identical to the one implemented in Chapter 6 was also employed here. MLR, LR, and the school-based approaches were considered.

It was decided not to develop models for Pima County for two reasons: a) the small sample sizes for each factor group (between 5 and 8 sites) and b) the differing pedestrian activity patterns compared to Ontario sites, as illustrated by the distinct factor group patterns (Figure A1). Given these differences, associating STC sites with factor groups across jurisdictions would not be meaningful.

Regarding the Ontario jurisdictions, Figure A1 shows three factor groups for Waterloo and Toronto, which suggests the development of multinomial logistic regressions (MLRs). However, the sample sizes for Factor Groups #2 and #3 in Toronto are only 7 and 6 sites, respectively; numbers that are insufficient for proper model calibration. As a result, an MLR was only developed for Waterloo and not for Toronto. In Milton, two factor groups were observed, leading to the use of a binary logistic regression (LR) instead.

Results

Table A1 presents a summary of the MLR and LR models. The coefficients are calculated relative to a reference group, indicated in the table. Overall, all coefficients align with theoretical expectations and are consistent with the literature. The MLR achieved a general accuracy of classification of 74.3%, which is slightly higher than the performance reported in previous studies: 67% by Medury et al. (2019) and 69% by Sobreira & Hellinga (2024) – Chapter 6. The combination of Factor Groups #1 and #2 into a single group in the LR improved the general classification, as anticipated for having a simpler model. However, the LR model in Waterloo shows poorer accuracy (61.9%) in classifying sites from Factor Group #3. The 5-fold cross-validation procedure, repeated 100 times, indicated no signs of overfitting. Finally, all variables had a variance inflation factor (VIF) lower than two, suggesting the absence of multicollinearity.

Table A2 summarizes the school-based models. The performance of the School & Commercial model is comparable to that of the LR model and significantly better than the School model. This highlights the importance of incorporating the commercial component into the school model. For the Waterloo results, the School & Commercial model should be interpreted as follows: a site is classified into Factor Group “#3: Schools” if a) there is any type of school within 400 meters and b) the commercial area within the 400-meter buffer represents less than 20% of the total buffer area.

Table A1 Summary of MLR and LR models

Variable	MLR			LR			
	Waterloo (ref.: #1)			Waterloo (ref.: #1 + #2)		Milton (ref.: #1)	
	Radius (m)	#2	#3	Radius (m)	#3	Radius (m)	#3
		Coef. (SD)	Coef. (SD)		Coef. (SD)		Coef. (SD)
Intercept	-	-0.506 (0.568)	1.284 (0.645)	-	0.373 (0.466)	-	-0.754 (0.850)
Commercial area ¹	400	-0.022 (0.019)	-0.214 (0.053)	400	-0.176 (0.044)		
Employment density ²						400	-1.270 (0.532)
Park area ³	100	-1.869 (0.826)	-0.266 (0.190)				
Schools - Sec/Post-sec ⁴	200	3.484 (0.940)	4.652 (1.125)	200	2.233 (0.701)	800	2.569 (1.000)
Trails ⁵	100	0.594 (0.225)	0.017 (0.322)				
Accuracy of classification - General (all data)			74.3%		86.1%		84.9%
Accuracy of classification - Schools (all data)			81.0%		61.9%		80.0%
Accuracy of classification - General (5-fold train dataset)			71.1%		85.8%		84.8%
Accuracy of classification - General (5-fold test dataset)			62.5%		84.1%		84.0%
McFadden pseudo-R ²			0.36		0.37		0.36
Sample size			101		101		33

Ref. = Reference; Coef. = Coefficient. SD = Standard deviation. Statistical significance: bold for p-value < 0.05.

Factor Groups: #1= Multipurpose; #2 = Commuter; #3 = Primarily school trips

¹Percentage variable: area of parcels that are commercial / total buffer area × 100. ²Employment density per km² / 1,000.

³Park area in m² / 1,000. ⁴Count of secondary and postsecondary schools. ⁵Length of multi use trails and paths in km × 10.

Adjustments were made to variables to avoid very small or large coefficients

Table A2 Summary of school-based models

Variable	School & Commercial (Radius / % commercial)			School (Radius)		
	Waterloo	Toronto	Milton	Waterloo	Toronto	Milton
Commercial area ¹	400m	200m	400m			
Commercial area % threshold	20%	25%	25%			
Schools - All types	400m			200m		
Schools - Secondary/Post-secondary		400m	800m		200m	800m
Accuracy of classification - General	80.2%	94.1%	81.8%	74.3%	73.5%	66.7%
Accuracy of classification - Schools	81.0%	83.3%	80%	66.7%	66.7%	86.7%
Sample size	101	34	33	101	34	33

¹Percentage variable: area of parcels that are commercial / total buffer area × 100

Step 3: Assigning STC sites with a given factor group in the target jurisdiction

Step 3 switches to the target jurisdiction, where only STC sites are available. The models developed in Step 2, based on the base jurisdiction, are transferred and applied to the sites in the target jurisdiction to assign each site to a specific factor group.

Table A3 presents the confusion matrices for each model and each base-target pair. To provide a reference, we also included cases where the base and target jurisdictions are the same. On average, the models transferred reasonably well. When the base and target jurisdictions were the same, the average classification accuracy was 78% for Factor Groups #1, #2, or #1/2, and 76% for Factor Group #3. When the base and target jurisdictions differed, these accuracies were 70% and 64%, respectively. However, there were specific cases of poor transferability: the MLR model (base: Waterloo) performed poorly when transferred to Toronto (Factor Group #2) and Milton (Factor Group #3), and the School model (base: Milton) struggled when transferred to Toronto (Factor Group #1/2).

Table A3 Confusion matrices: transferring models to assign factor groups

Target jurisdiction	Observed factor group	# Sites	Base jurisdiction (predicted factor group)																		
			Waterloo									Toronto					Milton				
			MLR			LR		S + C		S		S + C		S		LR		S + C		S	
			#1	#2	#3	#1/2	#3	#1/2	#3	#1/2	#3	#1/2	#3	#1/2	#3	#1/2	#3	#1/2	#3	#1/2	#3
Waterloo	#1: Multipurpose	48	40	3	5	74	6	64	16	61	19	71	9	65	15	62	18	59	21	36	44
	#2: Commuter	32	12	18	2																
	#3: Schools	21	4	0	17	8	13	4	17	7	14	9	12	11	10	8	13	5	16	5	16
	Classification accuracy		83%	56%	81%	93%	62%	80%	81%	76%	67%	89%	57%	81%	48%	78%	62%	74%	76%	45%	76%
Toronto	#1: Multipurpose	21	16	4	1	27	1	20	8	13	15	27	1	21	7	25	3	18	10	2	26
	#2: Commuter	7	6	1	0																
	#3: Schools	6	1	1	4	3	3	1	5	1	5	1	5	2	4	1	5	1	5	1	5
	Classification accuracy		76%	14%	67%	96%	50%	71%	83%	46%	83%	96%	83%	75%	67%	89%	83%	64%	83%	7%	83%
Milton	#1: Multipurpose	18	11	6	1	13	5	15	3	17	1	18	0	17	1	16	2	15	3	9	9
	#3: Schools	15	2	7	6	9	6	5	10	6	9	8	7	8	7	3	12	3	12	2	13
	Classification accuracy		61%	-	40%	72%	40%	83%	67%	94%	60%	100%	47%	94%	47%	89%	80%	83%	80%	50%	87%

Note: MLR = Multinomial logistic regression; LR = Logistic regression; S + C = School & Commercial; S = Schools