Estimating Average Annual Daily Pedestrian Volumes at Intersections based on Turning Movement Counts

by

Andrew Orr

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Master of Applied Science

in

Civil Engineering

Waterloo, Ontario, Canada, 2021

Author's Declaration

I hereby declare that I am the sole author of this 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.

Abstract

There is a focus on increasing the use of active transportation and, consequently, a need to have pedestrian traffic volumes such as Annual Average Daily Pedestrian Traffic (AADPT) for infrastructure planning and safety analysis. Traditional methods rely on the deployment of dedicated sensors to count pedestrians, but this limits the number of locations at which counts can be obtained and therefore does not permit estimation of AADPT for all intersections in the urban area. The focus of this thesis is to propose and evaluate methods for addressing this limitation.

The proposed methods assume that (i) dedicated sensors that provide continuous pedestrian volume counts are deployed at a small number of intersections within the urban area, and (ii) 8-hour turning movement counts (TMCs) are available for intersections for which AADPT are to be estimated. These two assumptions are normally met in practice. Within this context, the problem of estimating AADPT can be divided into five sub-problems, namely:

1. Calculating AADPT with missing counts in a dataset
2. Selecting and implementing a set of count data filters
3. Associating specific continuous count sites with each other
4. Finding suitable factors groups for short-term count sites
5. Converting short-term counts to AADPT estimates

This thesis examines the existing methods in the literature for solving each of these sub-problems and proposes several extensions. By solving all the subproblems, there is a hope that reliable average daily estimates from pedestrian data collected alongside turning movement counts can be obtained. It is recommended to use the AASHTO method for determining continuous count site AADPT values or solving sub-problem 1. For the data filters, it was determined that using pre-exiting filters from the literature with some adjustments was appropriate. However, a new null count filter was needed for the dataset. For grouping specific continuous count sites, existing solutions from the literature were incorporated into this work along with a proposed k-means clustering approach. Specific land uses and temporal metrics were incorporated into linear regression models for the purposes of predicting specific temporal trends and placing a short-term count site in a factor group. Lastly, the AADPT estimation methods were all taken from the literature and are mathematically adjusted to handle 8hr to 24hr conversions.

The methods are applied to a set of field data from Milton, Ontario and Pima County, Arizona. The results indicate that the AADPT estimation error metrics still are much larger for count sites located within 1km of a high school and, consequently, a modified factor grouping method is proposed for sub-problems 3 and 4.

Acknowledgements

I would like to express my thanks to my MASc supervisor Dr. Bruce Hellinga. His knowledge in transportation engineering and his commitment to helping me improve upon my research has been truly valuable to me. I am extremely honored that I had the chance to be part of his research group.

To the members of Dr. Hellinga's research group that I had the opportunity to meet and get to know. Alan Xaykongsa, Ben Allen, Mohammad Zarei, Wenfu Wang, Trevor Vanderwoerd, and Lucas Sobreira I want say how great they all are and how much I appreciate all the help they gave me whether it was related to my research, other tasks I have worked on as a master's student or even if it was on a more personal level.

I would also like to acknowledge Miovision for providing the count data for my thesis and Matthew Muresan for preparing the Miovision data for me to use in my research.

I also would like to recognize Dr. Chris Lee from the University of Windsor who enhanced my interest in transportation engineering and convinced me to enroll in a master's program.

Lastly, I owe as much gratitude that I can possibly give to my parents Kevin Orr and Elsa Orr. They have always given me the support I needed so that I can strive to be successful and the encouragement for me to become greatest individual I can possibly be. I thank both of them so much for always having my back and believing in me.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction and Background

1.1 Introduction and Research Motivation

Recently, various municipalities have been considering to implement policies that promote environmental awareness and sustainable development. One of the ways that different municipalities are attempting to promote more environmentally friendly polices includes encouraging their respective citizens to consider alternative travel modes. One of the more predominant alternatives, is making the choice to walk to a given destination. Within a municipality, there may be different types of existing infrastructure that is useable for pedestrians and one prevalent aspect in municipal infrastructure often used by pedestrians includes roadway intersections. Therefore, if a municipality is trying to improve the design of select intersections and hopefully increase pedestrian usage, then it would be necessary for municipal officials to evaluate the mobility levels and safety risks at those select intersections. However, specific sources of information are required to complete a capacity/safety analysis at a given intersection.

One of the more important temporal characteristics that can be utilized for the purposes of infrastructure improvement and safety screening is the actual traffic volume (U.S. Department of Transportation Federal Highway Administration, 2016). For the most part, traffic data collected at an intersection usually corresponds to vehicular traffic. However, traffic volumes for active transportation modes such as cyclists and pedestrians can be collected at intersections as well. Traffic volumes can be expressed in terms of different time periods, including hourly counts, peak period counts, and daily counts (U.S. Department of Transportation Federal Highway Administration, 2016). Traffic volumes can be obtained for a specific date/day or can be averaged across multiple days where these days can be in a specific month or season, or specific days of the week, or across all days in a year (U.S. Department of Transportation Federal Highway Administration, 2016). Although the most appropriate count period and averaging period can vary depending on the application. It is also common to express traffic volume as either the Annual Average Daily Traffic (AADT), which reflects the daily traffic volume averaged across all days in the year, or the Seasonal Average Daily Traffic (SADT), which is the daily traffic volume averaged over all days within a specified portion of the year (U.S. Department of Transportation Federal Highway Administration, 2016). If a specific individual wanted to associate traffic volumes with pedestrians, AADT and SADT would be respectively known as Annual Average Daily Pedestrian Traffic (AADPT) or Seasonal Average Daily Pedestrian Traffic (SADPT).

1.2 Overview of Research Problem

1.2.1 Introduction to Research Problem

With the need for obtaining traffic volumes such as AADT has been established, it important to understand what datasets AADT can be calculated from and what steps are needed to calculate AADT based on the dataset used. More specifically, data used to obtain AADT can either come from a continuous count site (CCS) or a short-term count site (STCS) (U.S. Department of Transportation Federal Highway Administration, 2016). By definition, a continuous count site has some type of data collecting device at its specific location and the collection device is set up to continuously receive data for the entire analysis period (i.e., year or season) (U.S. Department of Transportation Federal Highway Administration, 2016). On the other hand, for a short-term count site, count data are only collected for

a short period of time (ranging from several hours to up to several weeks) and therefore count data are available for only a short portion of the analysis period (U.S. Department of Transportation Federal Highway Administration, 2016). However, it is critical to understand some of the challenges associated with the datasets coming from either type of count site.

At a continuous count site, data recording devices are deployed for the entire analysis period. It is also common for some data records to be unavailable because of sensor hardware malfunctions, data communication failures, and/or data storage errors. It is also possible that recorded data may contain unacceptably large errors (U.S. Department of Transportation Federal Highway Administration, 2016). Therefore, it is necessary to address **(1)** missing count data and **(2)** methods to ensure that only reliable count data are used.

At a short-term count site, count data are available for only a small portion of the entire analysis period which means that an AADT value cannot be calculated directly from these data. However, it is still possible to estimate AADT from a STCS, but it is necessary to make use of information obtained from nearby continuous count sites that experience similar temporal patterns of traffic volumes (U.S. Department of Transportation Federal Highway Administration, 2016). This is illustrated in Figure 1.1.

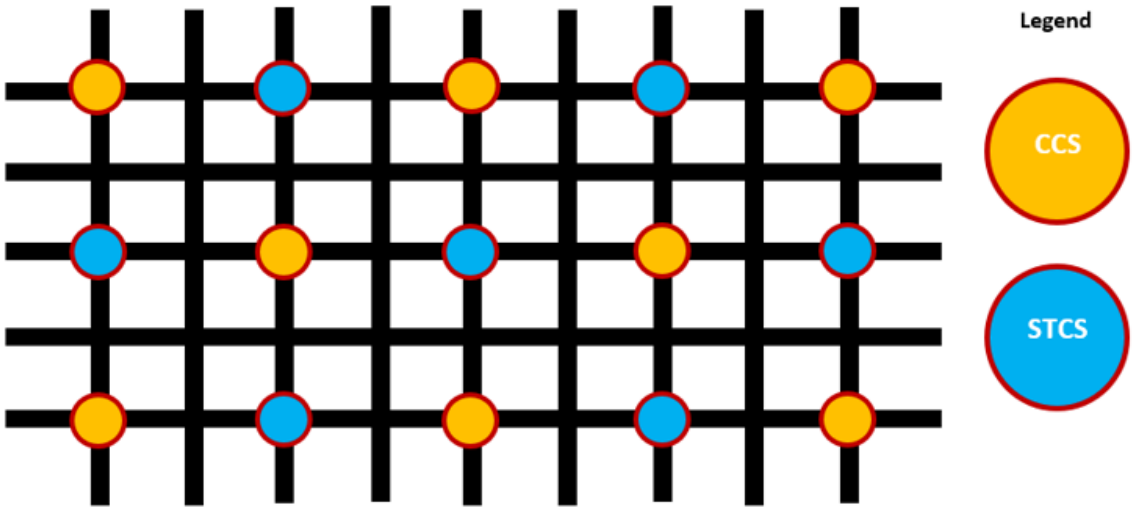

Figure 1.1. A Hypothetical Network with CCSs and STCs

Continuous count sites always have data that closely resembles a complete count data profile as implied in the previous paragraph. When a group of short-term count sites are matched up with a continuous count site, it is generally assumed that short-term counts would have the same count data profile as the continuous count site (U.S. Department of Transportation Federal Highway Administration, 2016). However, the process of matching different count sites can be challenging especially if the short-term count data are extremely limited. Therefore, this leads to two additional challenges which closely deal with the grouping of different types of count sites: (**3a**) how to divide the available continuous count sites into appropriate groups (called factor groups); and (**3b**) how to determine which group of continuous count site should be used as the appropriate reference for a given short term count site.

Once the association between short-term and continuous count sites is determined, the final challenge (**4**) is to determine methods by which AADT can be estimated based on the short-term counts.

Addressing these 5 technical challenges (illustrated in Figure 1.2) form the focus of this thesis. Each of these sub-problems is discussed in more detail in the following sections.

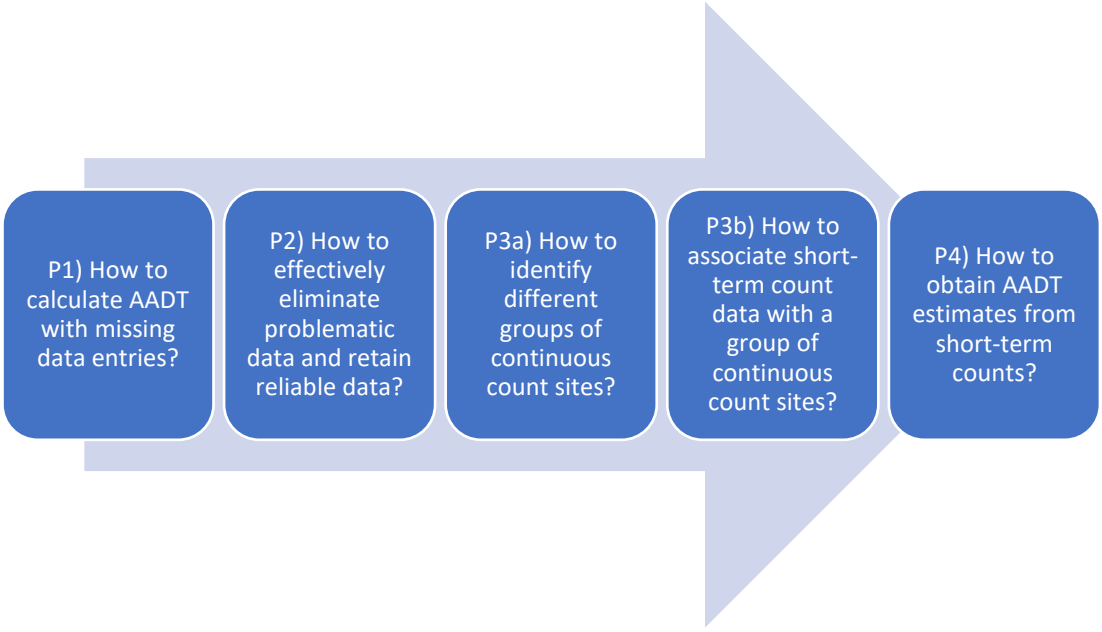

Figure 1.2. Technical Challenges Addressed in this Research

1.2.2 Challenges in Calculating AADT with Missing Counts (P1)

In a perfect world, a continuous count site would have a counter that acquires daily counts for every day of the year to calculate a value such as AADT. However, it is more than likely that not all the daily counts for a given study year is present before starting any computational procedures. Currently, there are existing methods for computing true AADT values with missing counts. However, the methods that currently exist have specific requirements for count totals showing up in every month-of-year (MOY), day-of-week (DOW) and, in some cases, hour-of-day (HOD) (U.S. Department of Transportation Federal Highway Administration, 2016). The reason why alternate AADT calculation methods for continuous count sites have specific count requirements is that, within a given study period, there are multiple counts that occur in different MOYs, DOWs and HODs. It is generally unreasonable to assume that a count occurring within a specific MOY on a specific DOW to be similar to another count occurring within a different MOY and on a different DOW. If the count values within a series of data are more abundant within a specific MOY and/or frequently occurring on a specific DOW, the AADT value becomes more biased towards that MOY and DOW. However, if the distribution of counts associated with all MOYs and DOWs is roughly around the same quantity, the issue of bias is less of a concern. Figure 1.3 and Figure 1.4 illustrate average traffic ratios (averaged across multiple sites in Milton, Ontario) for monthly to yearly pedestrian volumes and daily to weekly pedestrian volumes. Figure 1.3 clearly shows that the monthly traffic portions are quite different throughout the year and Figure 1.4 shows that, though pedestrian volumes are quite similar across different weekdays, the pedestrian volumes are lower on weekends. Therefore, having different distributions of counts corresponding to all the MOYs and DOWs could present challenges in calculating AADT.

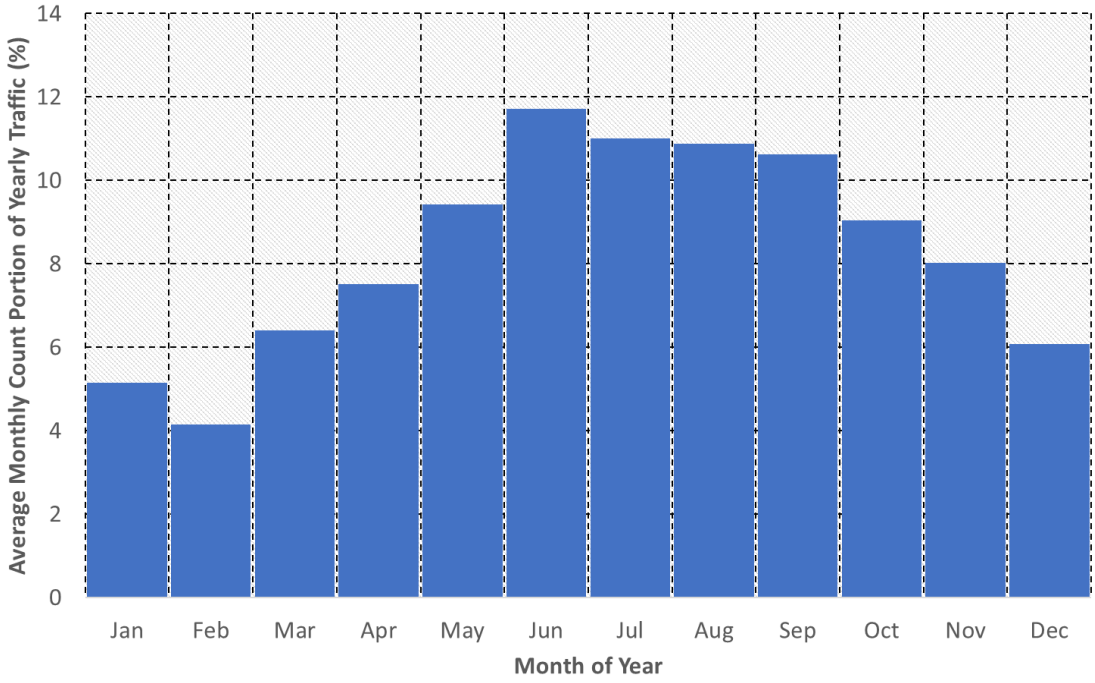

Figure 1.3. Portions of Average Monthly Counts in 2019 for Milton, Ontario

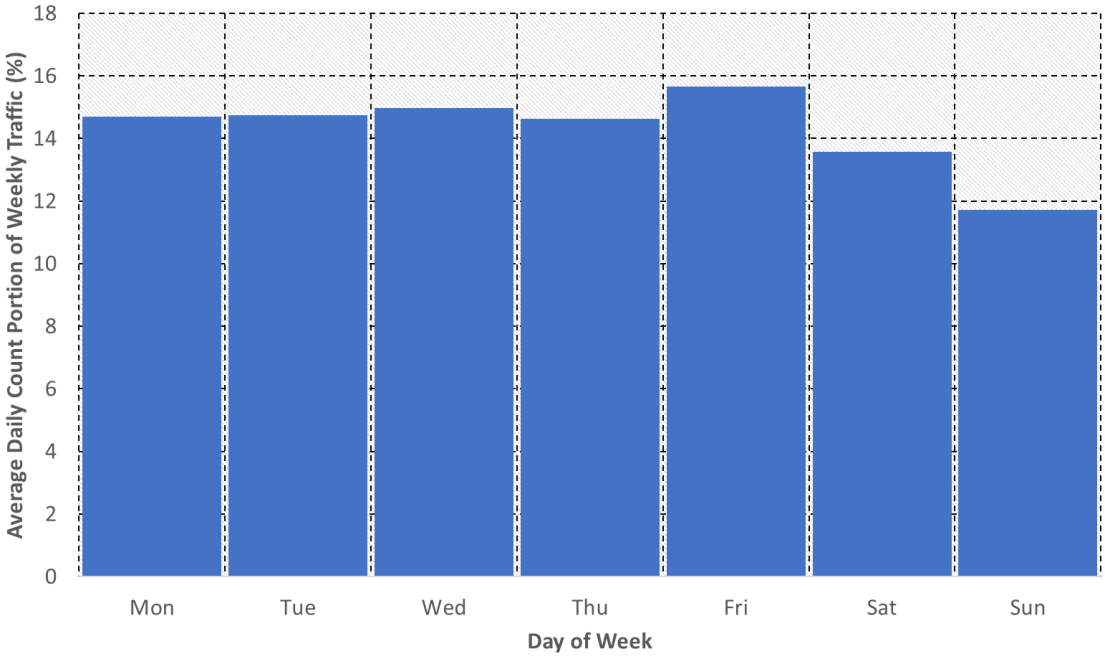

Figure 1.4. Portions of Average Daily Counts in 2019 for Milton, Ontario

1.2.3 Considerations for Obtaining Cleaned Data (P2)

The next issue that needs to be addressed is identifying erroneous or invalid counts in the recorded data to prevent these counts from adversely impacting the subsequent calculation of the AADT. Erroneous counts are those which reflect inaccurate measurements of field conditions. This may occur because of a malfunction of the counting sensor or through an error with the storage of the recorded values within

the data base (U.S. Department of Transportation Federal Highway Administration, 2016). Invalid counts are the counts that are accurate but reflect field conditions that are not representative of "normal" conditions at the site. For example, an emergency road repair necessitates the closure of the intersection for most of the day with the result that the counts for that day are much lower than that would have occurred on that day if the intersection had remained open. Collectively, erroneous counts and invalid counts are referred to as "outliers". The objectives are to identify the outliers using filtering algorithms and then determine methods to minimize their impact on the calculation of AADT.

The filtering algorithms are designed to remove outliers and keep the reliable data. An outlier is a count value that is significantly higher or lower than a series of count values for a given site within a specific window of time. If an individual wants to compute relevant summary values for a set of count data and removes the outliers from the computations, it is likely that the summary values are more representative of the dataset. It is also important to select data filters that are based upon the traffic volumes observed throughout an entire study period. As an example, Table 1.1 highlights the average across the unfiltered bike, pedestrian, and vehicle average daily traffic in 2019 at 7 sites in Milton, Ontario. From Table 1.1, it is clear that the three modes have vastly different magnitudes of daily traffic volumes. Therefore, the filtering algorithms that might be appropriate for one mode may not be directly applicable to data from another mode. Based on the comparison shown in Table 1.1, daily pedestrian counts might be more compatible with cyclists counts. Therefore, it is possible that various aspects of cyclist data filters may be applicable for pedestrians.

Table 1.1. Average Daily Traffic in 2019 by Mode in Milton, Ontario

Mode of Transportation	Average Daily Volume in 2019
Bike Traffic	21
Pedestrian Traffic	198
Vehicle Traffic	27872

1.2.4 Obstacles in Forming Groups of Continuous Count Sites (P3a)

For short-term counts sites, an observed AADT value cannot be directly computed (U.S. Department of Transportation Federal Highway Administration, 2016). However, the question becomes: Are there any other sources of known information that could help produce an estimate of AADT for a short-term count site? To provide some insight into this question, traffic studies could associate short-term count sites with continuous count sites by searching for similar temporal traits. This could imply that the average daily and/or hourly traffic volume for a selection of count sites seem to exhibit similar profiles (patterns) and the practice itself is known as factor grouping (U.S. Department of Transportation Federal Highway Administration, 2016). Figure 1.5 and Figure 1.6 both show a hypothetical example of two distinct day-of-week and hour-of-day patterns, respectively. Note that the data displayed could either come from a CCS or a STCS if the STCS has the minimum quantity of data needed to display the temporal information. For pattern one, the average traffic volumes during Monday to Friday are greater than the mean traffic volumes on Saturday and Sunday. For the average hourly trends displayed by pattern one, it appears that the traffic peaks in the morning and the afternoon. On the other hand, the average weekend traffic volumes for pattern two are larger than the mean weekday traffic volumes. The average hourly traffic is the highest during the middle of the day for pattern two.

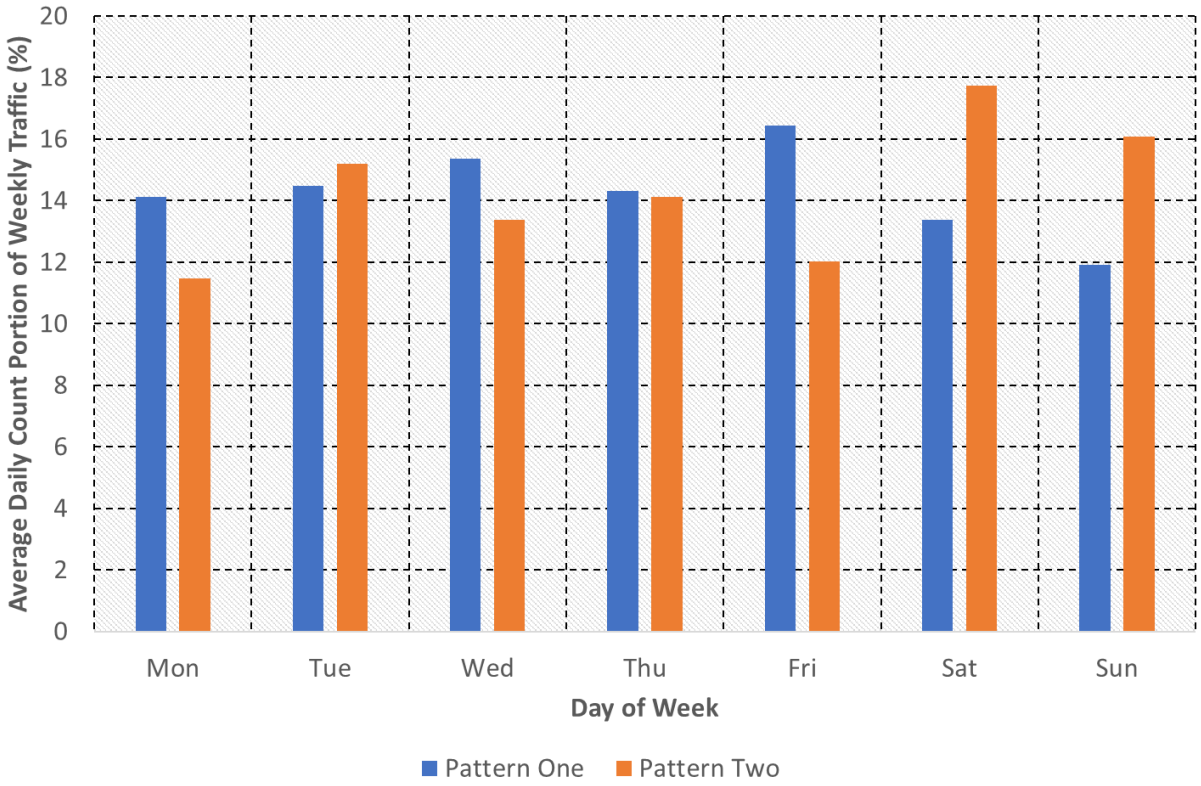

Figure 1.5. Two Hypothetical Sites with Different Day of Week Patterns

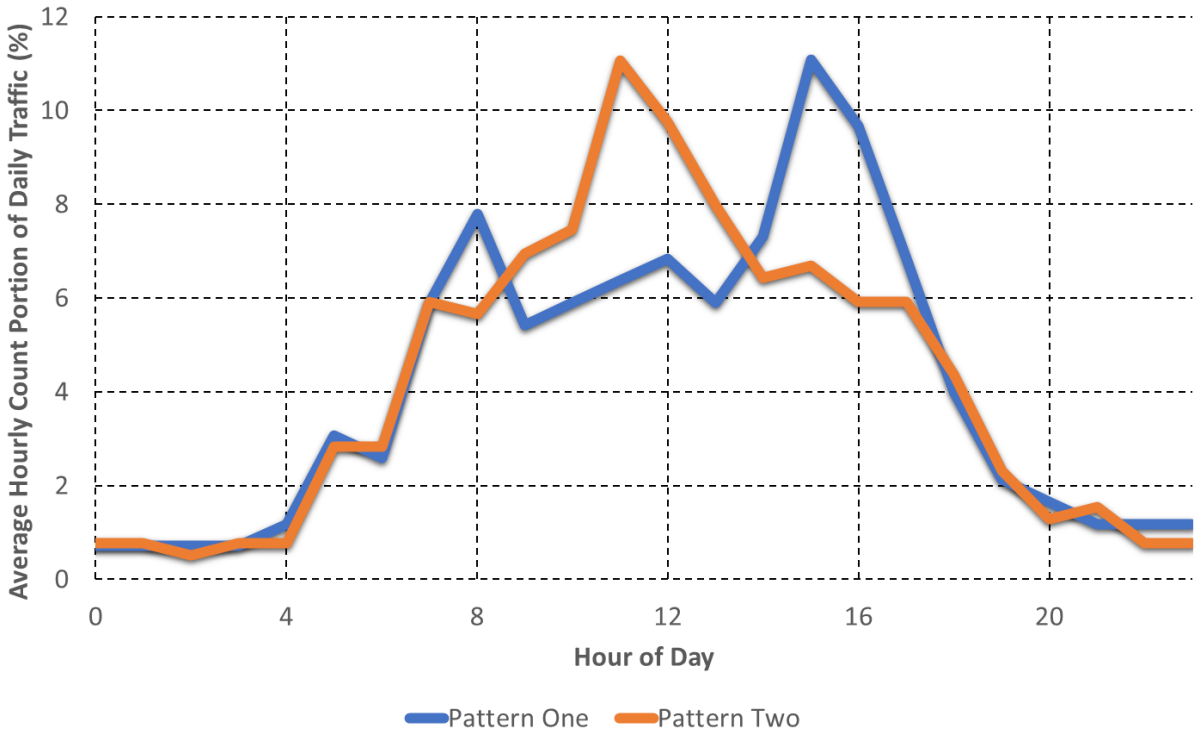

Figure 1.6. Two Hypothetical Sites with Different Hour of Day Patterns

1.2.5 Difficulties in Associating Short-Term and Continuous Data (P3b)

As previously implied, short term counts may be used over a duration that is too short to provide sufficient temporal information to permit factor grouping. The reason why those situations might occur is that the pedestrian count methodology/setup selected. Traditionally, pedestrian counts are collected at some pre-defined points of interest located on a piece of infrastructure exclusively utilized for active transportation modes (U.S. Department of Transportation Federal Highway Administration, 2016). Short-term count durations collected in that manner are usually long enough to obtain day-of-week and hour-of-day traffic volume profiles and it is possible to determine the appropriate factor group based on these temporal patterns.

For the research addressed in this thesis, the pedestrian data are not obtained from dedicated count surveys, but rather from turning movement count (TMC) studies. Those specific studies are already carried out on an ongoing basis (typically once every 2 or 3 years) by the local municipality for the purpose of assessing signalized intersection performance (Mohammed, 2019). A turning movement count (TMC) study consists of observing the traffic on the roadway passing through the intersection and the pedestrians crossing each approach (Mohammed, 2019). Those traffic volumes are recorded by turning movements over each 15-minute period during the data collection period (Mohammed, 2019). The data collection period typically consists of 12 or 8hours, covering the AM peak, mid-day, and PM peak periods (Mohammed, 2019). The data collection day is usually a weekday during the time of year that avoids inclement weather and significant holidays that might impact traffic demand patterns. The collected counts represent a sample collected on only a single day of the year and there is not enough temporal variation information within the count to determine the appropriate factor grouping.

Consequently, some other method is required to determine which factor group the STC should be associated with.

1.2.6 Challenges in Obtaining Estimations for AADT (P4)

After all the count sites are placed within a factor group, the datasets from sites having a complete temporal profile (i.e., continuous counts sites) are utilized to calculate expansion/scaling factors. The scaling factors (*SF*) for a count site typically represents a relationship between the average daily traffic for the entire analysis period such as (*AADT*) and the average daily traffic (*ADT*) associated with a specific aspect of a given temporal profile such as month-of-year or day-of-week as depicted in Eq. 1.1 (U.S. Department of Transportation Federal Highway Administration, 2018).

$$SF = \frac{AADT}{ADT} \quad \text{Eq. 1.1}$$

Once the expansion factors have been computed from the continuous count sites within a factor group, the scaling factors (*SF*) are then applied to the short-term counts (*STC*) collected from the short-term count sites. The application of scaling factors produce an estimated traffic volume (such as *Est AADT*) and is mathematically shown in Eq. 1.2 (U.S. Department of Transportation Federal Highway Administration, 2018). However, it not always clear if the estimates are a good representation of what the observed traffic volume rate could be for any location. This is because pedestrians tend to be more reactive to undesirable weather conditions (Saneinejad, Roorda, & Kennedy, 2012). These weather conditions could include the quantity of precipitation, fluctuations in temperature, and wind speed. Short-term count days should also be collected during times when the traffic is around the average for a given study period (U.S. Department of Transportation Federal Highway Administration, 2016). As an example, Figure 1.7 presents the daily traffic volume for a year and the associated AADT for a hypothetical CCS. The CCS exhibits substantial day-of-week and seasonal variation in the daily traffic volumes. Figure 1.7 also shows two hypothetical STCS with daily counts collected for one week in February at STC1 and one week in October at STC2. From the seasonal variation in the daily traffic volumes and the time of year when the short-term counts are taken, the mean of the daily counts from STC2 provide a much more accurate estimate of the AADT than those from STC1. Consequently, it is necessary to account for these influences on the STCs when estimating the AADT.

$$Est\ AADT = STC \times SF \quad \text{Eq. 1.2}$$

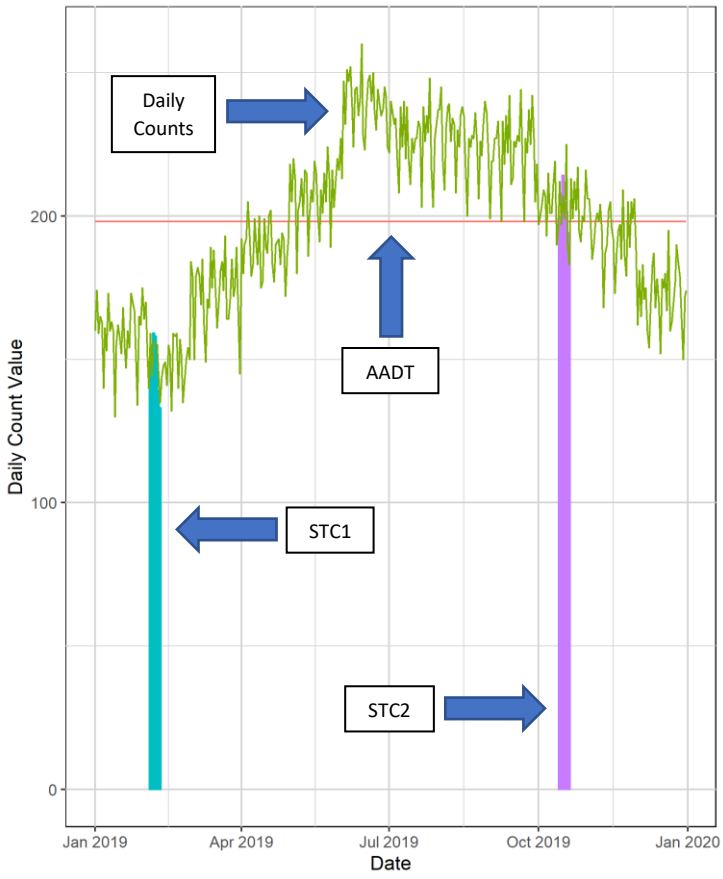

Figure 1.7. Hypothetical Short-Term-Count Selection Example

Another thing to consider is that the process of estimating AADT has mostly been developed for vehicular traffic but the temporal variations (time-of-day, day-of-week, month-of-year, holidays, etc.) and sensitivity to weather, may be quite different between vehicle traffic and pedestrians (U.S. Department of Transportation Federal Highway Administration, 2016). Therefore, it would be ideal to validate the applicability of traditional vehicular scaling factors to pedestrian data. When viewing the average hourly vehicle percentage and 95% confidence interval plot for a collection of sites in Pima County, Arizona for select short-term count days during the first quarter of 2020 in Figure 1.8, it seems that the 95% confidence ranges at each hour are relatively small which implies that there is a lot of consistency between sites. From Figure 1.8, it also appears that the highest hourly portions are present in the morning and afternoon periods. For the average hourly pedestrian percentage and 95% confidence plot in Figure 1.9 (also collected at the same sites and the same days as the data in Figure 1.8), the 95% confidence ranges for each hour are larger than the ranges for vehicles. Therefore, it is not always clear what type of traffic pattern the pedestrian data are exhibiting.

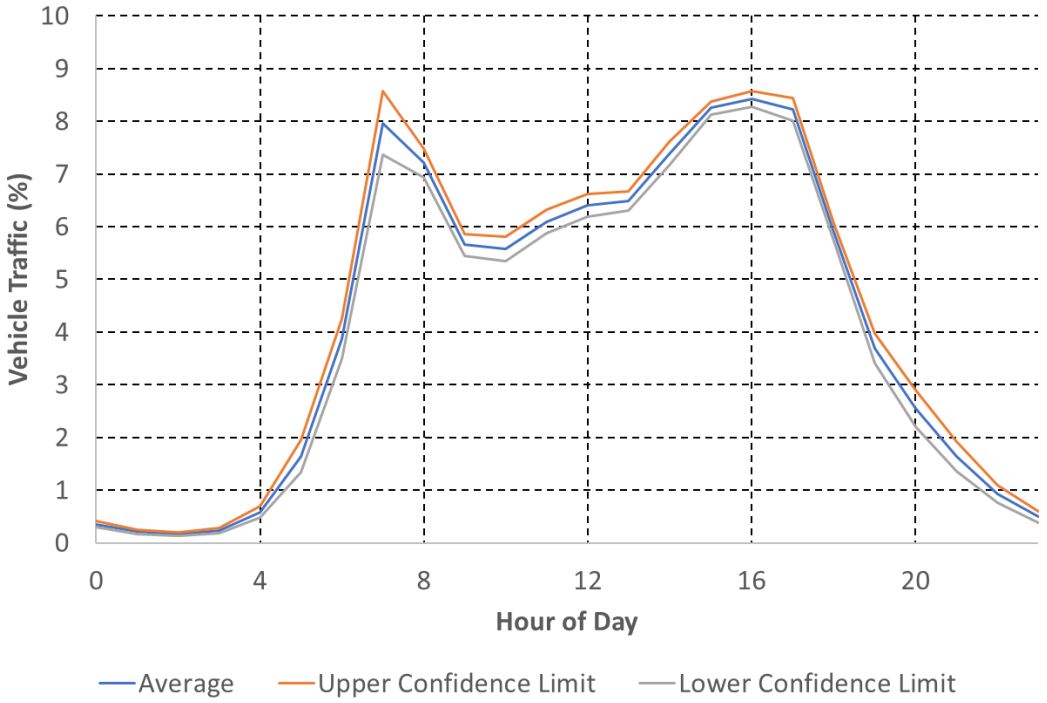

Figure 1.8. Hourly Vehicle Traffic Portions in Pima County, Arizona from January to March 2020

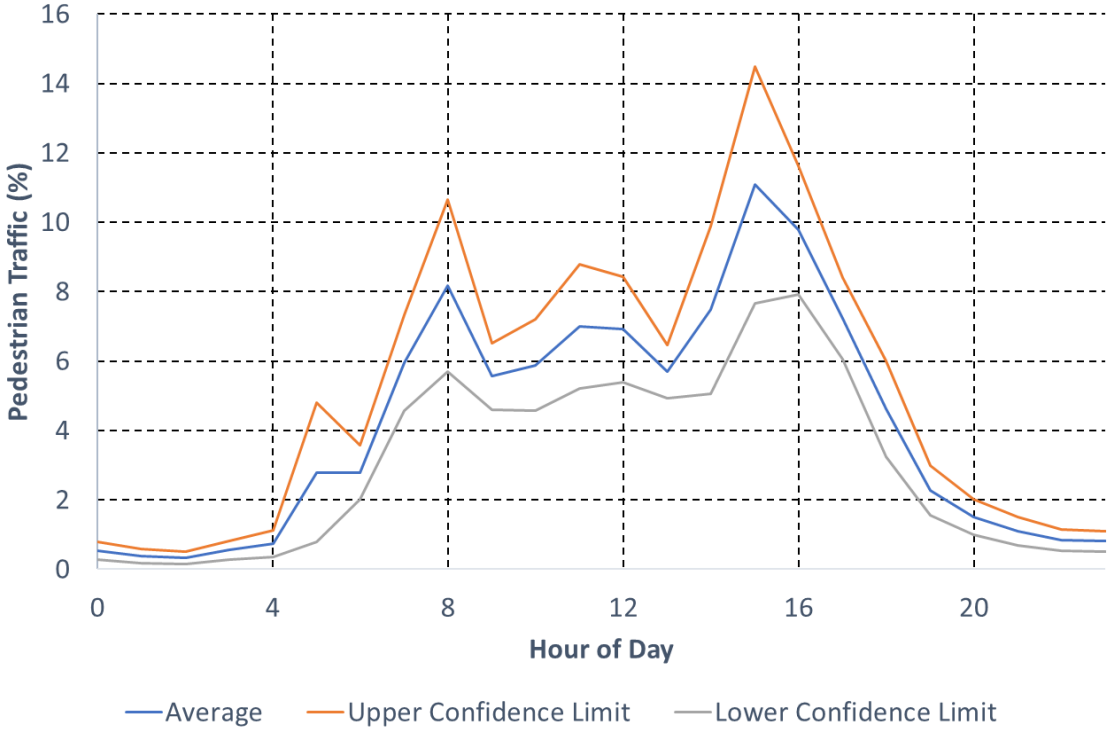

Figure 1.9. Hourly Pedestrian Traffic Portions in Pima County, Arizona from January to March 2020

1.3 Organization of Thesis

The main objective of this research is to develop a pedestrian traffic volume estimation methodology that is applicable to pedestrian counts collected at intersections. The remainder of this thesis is organized as follows:

Chapter 2 provides a review of the methodologies described in the literature for computing AADT from continuous count site data and for estimating AADT from short-term counts. The chapter also describes the existing factor grouping methods for non-motorized traffic . The data requirements for all the AADT calculation methods are highlighted and the advantages and disadvantages of each calculation are identified. The more traditionally used and recently introduced scaling factors and their respective calculation procedures are explained. For the review of pedestrian factor grouping methodologies, the utilization of temporal indices is examined along with the factor grouping based on land uses and the utilization of temporal metrics within a clustering analysis are also reviewed.

Chapter 3 describes the field data used in this thesis including the technology that is used to collect the data, the geographic characteristics in the areas of study and descriptive statistics associated with the collected pedestrian data.

Chapter 4 presents the proposed modifications to a data filtering methodology initially developed for cyclists and the results of applying this method to the field data used in this thesis. This chapter also describes the criteria used for modelling short term counts from continuous count site data. The criterion are weather thresholds, STC hours, days and months, and the elimination of holidays and count days potentially impacted by COVID-19.

Chapter 5 describes the proposed methods by which AADPT/SADPT can be estimated based on turning movement counts and Chapter 6 describes the proposed method for factor grouping all types of count sites.

Finally, conclusions and recommendations are provided in Chapter 7.

Chapter 2 Literature Review

2.1 Introduction

The purpose of this chapter is to provide some insight from the literature on how to address the main technical challenges associated with this research as presented in Figure 1.2. Section 2.2 provides some suggestions to deal with research problem **1**, which focuses on AADT computations from continuous count data. Section 2.3 provides some answers for research problem **4**, which highlights the concept of producing estimated AADT values from short-term data. Section 2.4 of this chapter is assembled to highlight some of the methodologies that can be used to form factor groups from continuous count sites which is specifically related to research problem **3a**. For Section 2.5, the computation of performance metrics is discussed along with the implications of a performance metric value with respect to placing a short-term count site in a factor group, which happens to correspond to research problem **3b**. In Section 2.6, recommendations for developing a research methodology are provided and the recommendations are based on the information presented in Sections 2.2 to 2.5. For research problem **2**, which is concerned with count data filtration, the main aspects of the utilized data filtration methodology were developed by Allen (2021) and is discussed in Chapter 4.

2.2 Computing AADT from Continuous Count Sites

With the research objectives for this thesis have been clearly defined, it is critical to review and evaluate the specific techniques used in past work that could help solve the outlined objectives. The first set of past techniques that are evaluated in this research includes the procedures to compute traffic counts estimates for a given study period from continuous count site data. More specifically, the Traffic Monitoring Guide (TMG) in Eq. 2.1 describes the following methodology to obtain an AADT value for a specific continuous count site (U.S. Department of Transportation Federal Highway Administration, 2016):

$$AADT = \frac{(\sum_{i=1}^{n} VOL_i)}{n} \quad \text{Eq. 2.1}$$

where:

i = An index representing a day of the year

n = Total quantity of days within a year (365 days; 366 in leap years)

VOL_i = The daily traffic volume on day *i*

When data are not available for all the days in a year, Eq. 2.1 cannot be used to compute AADT. It is possible to use a slight variation of Eq. 2.1 by simply re-defining *n* as the quantity of days within the year for which the total daily traffic volume counts are available in Eq. 2.2.

$$AADT = \frac{(\sum_{i=1}^{n} VOL_i)}{n} \quad \text{Eq. 2.2}$$

When the daily traffic volume counts are only available for a subset of days in the year, the use of Eq. 2.2 can lead to a biased estimate of the AADT. For example, consider the cases in which daily traffic volume counts are only available for weekdays but not weekends, or available for all the days in a year

except for the month of December. Therefore, using Eq. 2.2 for these two hypothetical cases leads to biased estimates of the AADT because it is not expected that weekends and weekdays have similar traffic volumes and the same is true for all months of the year as discussed in Chapter 1. Of course, it is more common that the missing data occur randomly throughout the year and then the impact of the missing data are negligible. However, the magnitude of estimation error increases as the number of days out of the year for which daily traffic volume counts are available decreases (U.S. Department of Transportation Federal Highway Administration, 2016). To account for this specific problem, AADT can also be calculated using the AASHTO method as presented in Eq. 2.3 (U.S. Department of Transportation Federal Highway Administration, 2016):

$$AADT = \frac{1}{7}\sum_{i=1}^{7}[\frac{1}{12}\sum_{j=1}^{12}\left(\frac{1}{n}\sum_{k=1}^{n} VOL_{i,j,k}\right)] \quad \text{Eq. 2.3}$$

where:

n = The number of times a specific day of week occurs in a specified month (the maximum value is typically 4 or 5)

k = index for a specific day of week *i* in month *j*

i = An index for day of week ($1 \le i \le 7$)

j = An index for month of year ($1 \le j \le 12$)

$VOL_{i,j,k}$ = The daily traffic volume on day *k*, day of week *i*, and month *j*

For the AASHTO method to work, the variable "*k*" must occur at least once in every month for all days of the week and therefore *n* must be ≥ 1. For example, if there was no data collected on Mondays in January for a specific year, this method would not be valid. This approach removes some of the bias caused from missing days by averaging months of the year and days of the week. However, this formulation requires that 24-hours of data be collected on all days considered – a condition that is not always met. As a result, the authors of the TMG decided to correct this problem by following the proposed FHWA methodology to obtain AADT by using Eq. 2.4 and Eq. 2.5 (U.S. Department of Transportation Federal Highway Administration, 2016):

$$MADT_m = \frac{\sum_{j=1}^{7} w_{j,m} \sum_{h=1}^{24}\left(\frac{1}{n_{h,j,m}}\sum_{i=1}^{n_{h,j,m}} VOL_{i,h,j,m}\right)}{\sum_{j=1}^{7} w_{j,m}} \quad \text{Eq. 2.4}$$

$$AADT = \frac{\sum_{m=1}^{12} d_m \times MADT_{HP_m}}{\sum_{m=1}^{12} d_m} \quad \text{Eq. 2.5}$$

where:

m = An index for month of year ($1 \le m \le 12$)

i = The occurrence number of a specific hour on a specific day of week and month

j = An index for day of week ($1 \leq j \leq 7$)

h = An index for hour of day ($1 \leq h \leq 24$)

$n_{h,j,m}$ = The total number of 1-hour traffic volume counts available for hour h, day of week j, and month m (the maximum value is typically 4 or 5)

d_m = Number of days in a specific month (28, 29, 30, 31)

$w_{j,m}$ = The number of times that day of week j occurs in month m (this value is equal to 4 or 5)

$VOL_{i,h,j,m}$= The measured hourly traffic on occurrence i, hour of day h, day of week j, and month m

$MADT_m$ = Monthly Average Daily Traffic in month m

For the FHWA method to be properly executed, the following assumptions are made. The variable "i" must occur at least once for every day of the week in every month for all hours of the day. Therefore, $n_{h,j,m}$ must be ≥ 1. If no data were collected from 5am to 6am on Wednesdays in October for a given year, the FHWA method would no longer be applicable.

It is also important to note that the variables "w_{jm}" and "d_m" are not dependent on the availability of the data but rather solely on the calendar. For example, there are 31 days in January and 4 Saturdays in January in 2020 and therefore $d_1 = 31$ and $w_{6,1} = 4$, even if no traffic volume count data had been collected for one or more of these Saturdays (U.S. Department of Transportation Federal Highway Administration, 2016).

2.3 Estimating AADT from Short-Term Counts

From reviewing the previous section, an understanding of calculating AADT from continuous count sites should be gained. However, as mentioned in Chapter 1, count data from continuous count sites are not always going to be available and short-term count sites could be used to obtain AADT estimates with specific obstacles to overcome. The TMG recommends the use of scaling factors for month-of-year, day-of-week, and (if necessary) hour-of-day (U.S. Department of Transportation Federal Highway Administration, 2016). These scaling factors are computed from the data obtained from an appropriate continuous count site. In Eq. 2.6, the day-of-week scaling factor can be computed as follows (U.S. Department of Transportation Federal Highway Administration, 2018):

$$D_i = \frac{AADT}{ADT_i} \qquad \text{Eq. 2.6}$$

where:

i = A specific day of week ($1 \leq i \leq 7$)

$AADT$ = Annual Average Daily Traffic

ADT_i = Average Daily Traffic for a collection of a specific day of week i

D_i = Day of week scaling factor corresponding to its respective day of week i

Average Daily Traffic (ADT) for a given day-of-week is computed by utilizing Eq. 2.7 (U.S. Department of Transportation Federal Highway Administration, 2018):

$$ADT_i = \frac{(\sum_{j=1}^{n} VOL_j \times \alpha_j)}{k} \quad \text{Eq. 2.7}$$

where:

j = An index representing a day of the year

n = Total number of days in a year (typically 365 days)

i = A specific day of the week ($1 \leq i \leq 7$)

VOL_i = The daily traffic volume on day *i*

α_j = Binary variable: $\alpha_j = 1$ if day *j* is of day of week type *i*; otherwise $\alpha_j = 0$.

k = The total number of days with a recorded value for VOL_j included in the summation

In Eq. 2.8, the month-of-year scaling factor computation is presented below (U.S. Department of Transportation Federal Highway Administration, 2018):

$$M_j = \frac{AADT}{MADT_j} \quad \text{Eq. 2.8}$$

where:

j = A specific month of year

AADT = Annual Average Daily Traffic

$MADT_j$ = Monthly Average Daily Traffic for a specific month *j*

M_j = month of year scaling factor

Eq. 2.8 requires that the Monthly Average Daily Traffic (MADT) be known. Three different methods have been proposed for computing MADT.

The first method (Eq. 2.9) assumes daily traffic volumes are available for every day in a month (U.S. Department of Transportation Federal Highway Administration, 2018):

$$MADT_j = \frac{(\sum_{i=1}^{n} VOL_{ij})}{n} \quad \text{Eq. 2.9}$$

where:

i = A specific day within a month

j = A specific month

n = The number of days within month *j*

VOL_{ij} = The daily traffic volume on day *i* and in month *j*

The second method (AASHTO method) can be used when total daily traffic volume data are not available for all days within the month as presented in Eq. 2.10 (U.S. Department of Transportation Federal Highway Administration, 2018):

$$MADT_k = \frac{1}{7}\sum_{j=1}^{7}\left(\frac{1}{n_j}\sum_{i=1}^{n_j} VOL_{i,j,k}\right) \quad \text{Eq. 2.10}$$

where:

k = A specific month

j = A specified day of the week

n_j = The frequency of day *j* occurring in month k given that count data are provided

i = An index value for the frequency of day *j* occurring in month *k*

$VOL_{i,j,k}$ = The daily traffic volume for an index value of *i*, day *j* in month *k*

The third method for computing MADT is defined in Eq. 2.4 and can be used when hourly count data are available, but not available for all hours of all days.

When a scaling factor for the month and for the day-of-week have been computed, the AADT can be estimated from the short-term count (assuming the short-term count spans a minimum of one full day) as follows in Eq. 2.11 (U.S. Department of Transportation Federal Highway Administration, 2018):

$$AADT = VOL_i \times M_j \times D_i \times G \quad \text{Eq. 2.11}$$

where:

j = A specific month

i = A specific day

VOL_i = A daily traffic volume on day *i*

M_j = A month of year scaling factor for month *j*

D_i = A day of the week scaling factor for day *i*

G = A growth factor

The growth factor, *G*, in Eq. 2.11 is only used when the year for which the AADT value is being approximated is not the same as the year for which the count data are provided (U.S. Department of Transportation Federal Highway Administration, 2016). When a growth factor is required, it can be calculated using Eq. 2.12 (U.S. Department of Transportation Federal Highway Administration, 2018):

$$G = \frac{AADT_{most\ recent\ year}}{AADT_{preceding\ year}}$$ Eq. 2.12

Nosal et al. (2014) utilized a method for estimating AADT for bicycle counts using a combined month-of-year and day-of-week scaling factor as defined in Eq. 2.13 and Eq. 2.14.

$$DF_{d,m} = \frac{ADB_{d,m}}{AADBT}$$ Eq. 2.13

$$\widehat{AADBT}_{i,y,j} = SDB_{i,y,j,d,m} \times \frac{1}{DF_{d,m}}$$ Eq. 2.14

where:

j = A specific day

i = A specific site

m = Month of year

y = A specific year

d = Day of week

$DF_{d,m}$ = A factor corresponding to day *d* occurring within month *m*

$ADB_{d,m}$ = Average Daily Bicycles specific to day *d* occurring within month *m*

$SDB_{i,y,j,d,m}$ = Short-term Daily Bicycles specific to day *j* which corresponds to the combination of month *m* and day *d* occurring within year *y* at site *i*

AADBT = Annual Average Daily Bicycle Traffic

$\widehat{AADBT}_{i,y,j}$ = Estimated Annual Average Daily Bicycle Traffic for day *j* at site *i* within year *y*

Using Eq. 2.13 and Eq. 2.14 produces an expansion method that is more disaggregated than what is presented in Eq. 2.11 This is because Eq. 2.11 only considers at most 19 scaling factors (12 monthly + 7 daily) and in Eq. 2.13 and Eq. 2.14 there are 84 scaling factors (12 monthly x 7 daily). However, Nosal et al. (2014) also used an even more disaggregated method (creating a scaling factor for each day of the year) for calculating an estimated AADT value as presented in Eq. 2.15 and Eq. 2.16:

$$DF_{y,j} = \frac{DB_{y,j}}{AADBT}$$ Eq. 2.15

$$\widehat{AADBT}_{i,y,j} = SDB_{i,y,j} \times \frac{1}{DF_{y,j}}$$ Eq. 2.16

where:

j = A specific day

i = A specific site

y = A specific year

$DF_{y,j}$ = A factor corresponding to day j within year y

$DB_{y,j}$ = Daily Bicycles specific to day j within year y

$SDB_{i,y,j}$ = Short-term Daily Bicycles specific to day j at site i corresponding to year y

$AADBT$ = Annual Average Daily Bicycle Traffic

$\widehat{AADBT}_{i,y,j}$ = Estimated Annual Average Daily Bicycle Traffic specific to day j at site i within year y

All the AADT estimation methods presented so far require daily count values as an input. When short term counts are less than 24 hours in duration, then an hour-of-day or time-of-day scaling factor is required. This way, the available data can be scaled to represent a full 24-hour count. One of the ways that a time-of-day scaling factor is computed is provided in Eq. 2.17 (U.S. Department of Transportation Federal Highway Administration, 2018):

$$TDF_h = \frac{MHT_h}{\sum_{h=0}^{23} MHT_h} \quad \text{Eq. 2.17}$$

where:

h = A specific hour of day

MHT_h = Mean Hourly Traffic for hour h

TDF_h = A factor for time of day specific to hour h

The result from Eq. 2.17 can be used to convert an hourly count to a daily (24 hour) count and then one of the three previously mentioned AADT estimation methods can be used to estimate AADT.

Another approach is to establish a direct relationship between a specific hourly count and AADT as was done by Nordback et al. (2013b). The hour-of-day factor in that study is equal to the true AADT divided by the specific average hourly count. However, the AADT estimation method used in Nordback et al. (2013b) still considers month-of-year and day-of-week scaling factors. Therefore, the method itself is essentially a traditional expansion method applicable for hourly counts. In El Esawey et al. (2015), two different types of hourly factors are utilized to obtain AADT estimates. However, the hourly factors have a specific focus on the peak hour counts and the factors are more specifically known as a k-factor.

With all the relevant scaling factors outlined, it is also critical to know some of the alternate ways of obtaining pedestrian volume estimates. For a study completed by Griswold et al. (2019) that took place in California and specifically focused on pedestrian crossings at intersections throughout the entire state, a log-linear regression model was utilized to produce an estimated values for annual pedestrian traffic. The log-linear regression model used within Griswold et al. (2019) considered network characteristics and different land use properties as relevant explanatory variables.

Within a different study by Schneider et al. (2013) that took place within San Francisco, California, the researchers also focused on pedestrian traffic at intersections and utilized log-linear regression. The overall purpose of using log-linear regression within Schneider et al. (2013) is to capture any existing connections between various properties surrounding the intersections and estimates of annual pedestrian traffic.

There are also other alternatives that can be implemented to get pedestrian volume estimates such as negative binomial regression which was used by Cao et al. (2006). Within Cao et al. (2006), the neighborhoods of interests are located in Austin, Texas and negative binomial regression is utilized to obtain the quantity of two different trip types pedestrians can make.

2.4 Grouping Continuous Count Sites Based on Temporal Traffic Patterns

With some of the common scaling factors being outlined in the previous section, it is important to realize that scaling factors developed from continuous count sites should ideally be only applicable to short-term count sites that are in the same factor group as the continuous count sites. The TMG identifies the following three methods to create pattern groups all of which do so based on the monthly average daily traffic for motorized vehicles (U.S. Department of Transportation Federal Highway Administration, 2016):

- Traditional Approach: This approach is subjective and consists of visualization of monthly trends and specific characteristics of the road to create monthly pattern groups. The TMG suggests that usually a minimum of 3 to 6 groups are required.
- Cluster Analysis: This method uses the clustering classification technique to create the monthly patterns groups. A feature vector is defined for each continuous count site consisting of the set of ratios of AADT to MADT. A distance metric is chosen such as the Euclidean distance between vectors. Clustering then attempts to classify each site into a group such that the within group variance of the distance metric is minimized and the between group distance metrics are maximized.
- Volume Factor Group Method: This method creates factor group based on the roadway classification system and specifically creates a factor group specific for site located on interstate highways.

The TMG also suggests that at a minimum, the following factor groups for vehicular traffic should be used for a state-wide continuous count program:

- Interstate rural
- Other rural
- Interstate Urban
- Other Urban
- Recreational

Then each short-term count site is associated with one of the factor groups. This is typically done subjectively based on the road classification and geographic location. The mean scaling factors (such as

M_j and D_i) are computed from the set of continuous count stations within each factor group and then the mean scaling factors are used to estimate AADT from the short-term counts.

With the common factor grouping for motorized traffic being outlined, the next issue to address is completing the factor grouping process for pedestrians. In a document prepared by the United States Federal Highway Administration (FHWA) called "Exploring Pedestrian Counting Procedures", they suggest that factor groups for pedestrians should be split up by the specific reason for using a pedestrian facility and by pedestrian facility type. However, the document does not provide any details to quantify the factor group development process (U.S. Department of Transportation Federal Highway Administration, 2016).

In a study by Miranda-Moreno et al. (2013) on estimating AADT for cyclists, the authors propose to allocate sites to one of two factor groups (recreational or utilitarian) based on visual inspection of daily and hourly temporal trends. The initial factor grouping decision was verified through the examination of the 95% confidence intervals of the daily and hourly indices. For example, if a given count site visually appears to follow a recreational traffic pattern, but the confidence intervals do not verify that claim, the site could be classified as mixed recreational. If that is the case, confidence intervals are recalculated for the mixed recreational group and if the site is not inside the confidence interval boundaries, then the site belongs to the unknown group. The same is true, if the site is initially classified as utilitarian, but instead the site could end being classified as mixed utilitarian or again unknown based off the computed confidence intervals. Another important concept that was developed from this study is the idea of providing summary statistics such as the morning to midday (AMI or $I_{AM/mid}$) and weekend to weekday (WWI or $I_{we/wd}$) index and the computation of these two indices are shown in the Eq. 2.18 and Eq. 2.19 (Miranda-Moreno, Nosal, Schneider, & Proulx, 2013).

$$I_{AM/mid} = \frac{\delta_i^{AM}}{\delta_i^{mid}} \qquad \text{Eq. 2.18}$$

$$I_{we/wd} = \frac{\bar{V}_{we}}{\bar{V}_{wd}} \qquad \text{Eq. 2.19}$$

where:

i = A specified day

δ_i^{AM} = Traffic volume between 7am to 9am

δ_i^{mid} = Traffic volume between 11am to 1pm

$\bar{V}_{wd}$ = Average traffic on weekdays

$\bar{V}_{we}$ = Average traffic on weekends

The appearance of AMI and WWI has become relatively frequent in non-motorized traffic studies. For example, Hankey et al. (2014) used AMI and WWI values to classify locations into a specific factor group based off the study from Miranda-Moreno et al. (2013). The criteria for these factor groups are shown

in Table 2.1. The same factor groups as the previously mentioned study (recreational, utilitarian, mixed recreational and mixed utilitarian) are also utilized.

Table 2.1. AMI and WWI Values for Factor Groups (Hankey, Lindsey, & Marshall, 2014)

Location Type	WWI	AMI
Utilitarian	<0.8	>1.5
Mixed Utilitarian	0.8-1.25	0.75-1.5
Mixed Recreational	1-1.8	0.35-1
Recreational	>1.8	<0.35

As indicated in Table 2.1, if a count location has a high AMI value, the count site likely follows a traffic pattern that is predominately utilitarian. Conversely, if a count site has a high WWI value, the count site is likely being used for recreational purposes (Hankey, Lindsey, & Marshall, 2014). It is also worth mentioning that the AMI and WWI classification made by Hankey et al. (2014) was strictly done to see what factor group each of the sites belongs to. Therefore, the actual factor grouping was not caried out for other parts of the research because of the lack of count sites present. Lindsey et al. (2018) also used AMI and WWI indices to classify the given count location into factor groups. In this study they use the label "commuter" instead of "utilitarian", but interestingly none of the count locations used in the study were categorized into the commuter factor group. However, for further analysis in the research sites were split up by their land uses.

In Nordback et al. (2019), the utilization WWI is applied in the factor grouping process and the 3 groups developed from that study are specifically known as: weekday commute, weekend multipurpose, and weekly multipurpose. For the weekday commute factor group, the count sites associated with the group itself experience heavier traffic Monday to Friday when compared with the weekend traffic count totals. The likely cause of experiencing that specific traffic distribution pattern is because the site is located on a route used for travelling to a utilitarian based destination. Therefore, it is typical for most users making those trips to do so Monday to Friday. If a count site belongs to the weekend multipurpose factor group, it is likely that the count site is on a route to a recreational activity such as a sporting event or travelling through an off-road trail. It is also possible for sites to show no difference between weekday and weekend traffic volumes, which means the site has mix of utilitarian and recreational based users travelling through it and that the site belongs to the weekly multipurpose factor group.

Nordback et al. (2019) used the classification criteria as follows:

if (WWI > 1.2)	Weekend multipurpose
if (WWI ≤ 0.8)	Weekday commute
If (0.8 < WWI ≤ 1.2)	Weekly multipurpose

In Johnstone et al. (2018), the authors strictly utilized AMI values and set thresholds based on AMI values to create three factor groups which are also known as the following: hourly commute, hourly noon activity, and hourly multipurpose. If a count site has a larger share of traffic in the morning peak verses the midday peak (i.e. AMI > 1), that is likely a good indication that the count site is along a route that users take when travelling to a destination with an essential purpose (e.g. home-to-work or home-

to-school). In that situation, the sites along that type of trip pathway would be classified as an hourly commute site. If a site consistently experiences heaver traffic later in the morning and/or earlier in the afternoon when compared to the morning rush hour period, that probably implies that the majority of users are not passing through the site for any utilitarian purpose (such as school or work). The reason why that is the case is because the midday hours occur when workplaces and schools are in session. From that setup, it is reasonable to expect that potential count site users within those settings don't travel during the midday hours, which means those count sites belongs to the hourly noon activity factor group. Lastly, it is also possible for some count sites, to have morning and midday peaks that are relatively similar in terms of traffic counts. Those specific sites could have a mix of users making utilitarian and non-utilitarian trips which implies that the site should be placed in the hourly multipurpose group.

Johnstone et al. (2018) used the following classification criteria:

if (AMI > 1.4)	Hourly commute group
if (AMI ≤ 0.7)	Hourly noon activity group
If (0.7 < AMI ≤ 1.4)	Hourly multipurpose group

As mentioned before in Hankey et al. (2014), AMI and WWI are both used to split up sites into factor groups. The advantage of using AMI and WWI together within the factor grouping process is that it provides the ability to classify utilitarian and reactional based traffic patterns on two different levels of data aggregation which includes daily totals for observing day-of-week trends and hourly totals for determining hour-of-day trends. However, the problem with this method is that there is some overlap between the AMI and WWI thresholds for the mixed recreational and mixed utilitarian factor groups. Depending on a given site's AMI and WWI values, it is possible that those two values suggest that the site of interest belongs in two different groups. That specific situation means that it is not really clear what factor group the site belongs to and it is also more than likely that potential solutions for classifying sites with conflicting AMI and WWI values vary from person to person.

In another study by Nordback et al. (2013a) in Colorado the researchers perform factor grouping on the basis of temporal trends and location information such as the count sites being located in the mountains or the Front Range. More specifically, the Front Range is a specific collection of mountains within Colorado (McGuire, 2021). Nordback et al. (2013a) also determine if the count site is in a rural setting or not. As an example, if a count site has more traffic on weekends versus weekdays. If the traffic volume seems to be greater on weekdays, the count site is classified as "high weekdays, low monthly variation" because a count site in that group resembles a commuter-based traffic pattern. The locations in the "high weekdays, low monthly variation" tend to be in more urbanized environments and in the Front Range although, sites can be in rural settings or in the mountains. However, if the traffic volume tends to be greater on weekends, the count site is either classified as "high weekends, high monthly variation" or "high weekends, low monthly variation". If the count site is both located in the mountains and is in a rural setting, the site is classified as "high weekends, high monthly variation" group and any other location type corresponds to the "high weekends, low monthly variation" group. In the "high weekends, low monthly variation" group, count locations are typically in the Front Range and setting for these count sites is mostly urbanized, but some sites can be mountain locations or in rural areas. In general, the factor grouping process in this study involved some visual inspection, location information to place

sites in their respective factor groups and the utilization of AMI and WWI values. This study also examined the effectiveness of AMI and WWI and it was determined that using an AMI value might not indicate if a count site follows a commute pattern when the site AMI value is large enough. This is because the AMI value ignores the afternoon peak. However, count sites with a commuter like pattern generally had a high AMI value and an afternoon peak, which was confirmed by using visual inspection. This study also used cluster analysis and it was determined that using WWI was a good quantitative parameter to utilize when classifying count sites. This is because two distinct clusters were formed using the WWI index.

Classifying locations in a way that considers the spatial variation is likely useful to incorporate because it could help generate a more enhanced understanding of pedestrian behavior as researchers are still in the process of doing so. For example, Jackson et al. (2015) suggested to classify sites based on the site location (i.e., in a rural, urban or university environment). The sites were differentiated by their temporal patterns and these temporal patterns were recreation, commute or mixed. Therefore, a total of 9 different factor groups were proposed.

Olfert et al. (2018) identified that large social events such as concerts, sports events, etc. often occur in the evening and can draw a relatively large crowd of pedestrians during times that are different from traditional spikes in pedestrian demands. As such they used the AMI and WWI indices to create factor groups but also introduced a new parameter called an Evening Portion Ratio (EPR) to incorporate pedestrian traffic building up during after work hours which is depicted in Eq. 2.20 and Eq. 2.21.

$$\overline{EPI} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=18}^{24} \%Daily\ Volume_{ij} \quad \text{Eq. 2.20}$$

$$EPR = \frac{\overline{EPI}_{event}}{\overline{EPI}_{non-event}} \quad \text{Eq. 2.21}$$

where:

i = A specific day

n = The quantity of days

j = Hour of day (24-hour)

$\%Daily\ Volume_{ij}$ = Percentage of daily traffic volume at hour *j* on day *i*

$\overline{EPI}_{event}$ = EPI (Evening Portion Index) value on days with events

$\overline{EPI}_{non-event}$ = EPI (Evening Portion Index) value on days with no events

The factor grouping methodology utilized in the study is a two-level system. The first part is to determine if the count site follows a recreational or utilitarian pattern using the WWI and AMI indices. It should be noted that in their study the count sites all seemed to follow a pattern known as “urban utilitarian”. Once the initial classification is determined, EPR was used to determine if the count site is

highly affected by social gathering in the evening. The criteria for a given count site EPR value to have or not influence special evening events is shown in Table 2.2.

Table 2.2. Range of EPR Values for Influence of Events (Olfert, Poapst, & Montufar, 2018)

Site Type	EPR Range
Not Influenced by Events	0-1.5
Influenced by Events	1.5-4

From this, the urban utilitarian and urban utilitarian event traffic pattern groups were created. However, Olfert et al. (2018) considers only evening events and therefore does not consider large pedestrian crowds that occur during the morning or afternoon (which may occur because of special events or even align with the arrival of transport modes such as passenger ferries or passenger trains). Incorporating events would be an interesting aspect to consider but, doing so might be difficult to incorporate if the events of interest do not generate a relatively large group of people.

In a pedestrian study by Griswold et al. (2018) , two factor group methodologies were compared. The first methodology used land use properties to establish factor groups which consisted of the following: (Central Business District (CBD), Trail, Commercial, School and Other). The second methodology consisted of k-means clustering to define the factor groups. Griswold et al. (2018) showed that both methods provided approximately the same results and therefore the authors recommended to utilize the land use method with some incorporation of clustering. However, it appears that some connections between land uses and temporal trends were established which is important note moving forward with this research but, not all land uses could be affiliated with specific temporal trend.

In a follow up study by Medury et al. (2019), the authors proposed a factor grouping method that uses land use information obtained from Google Places™ as input to a multinomial logistic model. A k-means algorithm was applied in the clustering analysis and 4 different clusters were created (Clusters A, B, C and D). The researchers then used the API for Google Places™ and were able to obtain descriptive land use information of any location of interest. The land use information obtained from Google Places™ was then grouped together in more generalized land use categories shown in Table 2.3.

Table 2.3. Land Use Categories (Medury, Griswold, Huang, & Grembek, 2019)

Land use	API Places (not exhaustive)	Search distance (m)
Professional office/service	Accounting, bank, local government office, doctor, courthouse	500
Transportation/travel	Airport, bus station, campground, car rental, lodging, RV park, train station	800
Entertainment	Bar, movie theater, amusement park, bowling alley, stadium	500
Emergency services	Fire station, hospital, police, ATM	800
Retail	Bookstore, convenience store, pharmacy, shopping mall, laundry, cafe	500
School	School	1200
Religious	Church, mosque, cemetery, synagogue	800
College	University	1200
Park	Park	500

The land use information was used as input variables and the four clusters created were used as specific categories in a multinomial logit model. The model coefficients and the elasticity of the variables were determined and led to the creation of the following factor groups:

- Central Business Districts (Cluster B)
- Isolated Recreational Trails (Cluster C)
- Urban Commuter Trails (Cluster A)
- Summer Vacation Destinations (Cluster D)

Based on the facts presented in the literature review, it appears that there are three main categories of factor grouping methods:

1. Temporal Indices, namely AMI, WWI, and, where appropriate, EPR.
2. Clustering Analysis
3. Land Use characteristics

The advantage of Temporal Indices is that they are easy to apply and only require the count data itself. EPR can only be applied when the short-term counts span a minimum of 24 hours. As for AMI, counts for the morning and midday portions of the day need to be present. For computing WWI, it can only be applied when the short-term counts span a minimum of 24 hours on both a weekday and a weekend day. Though criteria have been published to identify different factor group categories based on the values of these indices, it is not clear that these criteria are transferable to all locations. More importantly, they have only been used to classify the counts sites to a small number of factor groups.

Clustering methods have the advantage that they have a sound statistical foundation and can be used to identify any number of factor groups. However, they also suffer from several limitations. First, it can be difficult to interpret the resulting clusters in terms of describing the main characteristics of the factor groups. Second, conducting the clustering is more complex than applying the Temporal Indices methods and often requires software tools and expertise that practitioners do not possess.

Land Use methods have the advantage that they do not rely on the temporal pattern exhibited by the counts themselves and can be applied to count sites even when the short-term counts are available for only a truly short duration. The logit model proposed by Medury et al. (2019) can be used to define a larger number of factor groups than the temporal indices and the inputs required to apply the model are generally available as they are obtained from Google Places™. Lastly, the model has been shown to perform as well as the k-means clustering method. However, this model has only been recently proposed and it is not clear if the model is transferable to other locations. The most significant challenge in evaluating the factor group methods is the lack of literature in which the Land Use methods have been compared with the Temporal Indices methods.

2.5 Computing Performance Metrics from AADT Estimations

Several different methods exist for estimating AADT from short term counts. It is intended to maximize the estimation accuracy (or conversely minimize the estimation error) to be able to compare different AADT estimation methods. Though different measures of error exist, they all require that the true AADT be known. When only short-term counts are available, the true AADT is not known, a measure of error cannot be computed. Consequently, comparing the accuracy of different AADT estimation methods is done by using data from continuous count stations. The full year of counts at a given station is used to

compute the "true" AADT for that station as stated in Chapter 1. A set of *n* samples of count data are taken from the full set of data to represent a set of *n* short term counts. Each short-term count sample is used as input within the AADT estimating methodology to compute an estimated AADT ($\widehat{AADT_i}$) where i = 1, *n* and represents the sample index # (Wright, Hu, Young, & Lu, 1997). Then a measure of error between the true and estimated AADT can be computed. A common error metric utilized in pedestrian studies is the Mean Absolute Percent Error (MAPE) which is calculated in Eq. 2.22 below:

$$MAPE = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\left| AADT - \widehat{AADT_i} \right|}{AADT} \right) * 100\% \qquad \text{Eq. 2.22}$$

where:

n = The number of short-term count samples

$AADT$ = True AADT computed from full data set from continuous count station

$\widehat{AADT_i}$ = Estimate of AADT from short term count sample *i*

It is also essential to be aware of the limitations that exist when quantifying the AADT error obtained from short-term counts. As mentioned before, scaling factors that are applied to a specific short-term count site are derived from a continuous count site in the same factor group. However, the issue then becomes the choice of scaling factors used. As an example, a continuous count site has a full set of count data obtained (365 days of data collected for 24 hours each day for a given year) and every day is treated as an individual short-term count. Day-of-year scaling factors are developed from the full set of data and applied to each short-term count. Therefore, when the AADT error for the short-term site is computed, it is equal to zero.

This result is obtained because the geographic location and the year of study did not change, and each day-of-year scaling factor applied is calculated to represent a relationship between the true AADT and a specific daily count. It might appear as if this result is favorable, but one should be aware that this result is not realistic. This is because a specific factor group also include short-term counts that were obtained from locations different from the continuous count sites. Although, certain geographic locations might have similar traffic characteristics, it is almost next to impossible for two specific locations to have the exact same traffic data, which implies AADT errors are larger than zero.

This specific setup was used in Figliozzi et at. (2014) for cyclists in Portland, Oregon. However, day-of-year scaling factors were not utilized and ultimately the errors obtained for each day were not zero. Although, the authors of the study do mention that if the developed scaling factors are utilized for some other count location, they would anticipate that the computed errors would be larger. This implies that the errors cited by the study under-estimate the errors that would be experienced in practice.

In a completely different study by Nosal et al. (2014), count data for cyclists was obtained from multiple locations in Montreal, Quebec and Ottawa, Ontario. Some of the sites were used to create scaling factors for various AADT methods and the other sites were used to provide short term counts and to compute the errors for each AADT estimation method. Although the scaling factor and error computation sites were in the same municipality, the results obtained from the study has a range of success that varies.

Nosal et al. (2014) determined that one specific error computation site in Montreal did not appear to match temporally with Montreal's only count site used for scaling factor development. In another finding by Nosal et al. (2014), one specific error computation site in Ottawa produced a relatively low error when the scaling factors came from one specific scaling factor site. Nosal et al. (2014) suggest that the relatively low error value was likely obtained because both count sites followed the same traffic pattern. However, when the scaling factors were computed from Ottawa's other scaling factor site, the computed error was much higher. Nosal et al. (2014) concluded that this was likely because the scaling factor site and error computation sites experienced different traffic patterns and theoretically would have belonged to different factor groups. This result confirms that when temporal patterns between sites do not match, it is difficult to judge the relative performance of different AADT estimation methods because the error introduced by the factor grouping is so large. Therefore, it is necessary to ensure that the error computation (short-term count) and scaling factor (continuous count) sites are assumed to have similar traffic characteristics (same factor group) which could be difficult to accomplish if the short-term count's duration is only 8 hours.

2.6 Summary and Recommendations

Based on the literature review there are several recommendations that can be made for methods to adopt as the benchmark approach:

1. Calculating the actual "true" AADT/SADT:
   It is proposed to use the AASHTO method (Eq. 2.3) because it can effectively handle datasets with missing counts and the data requirements for the AASHTO method as described in section 2.2 are not overly extensive.
2. Factor grouping methods:
   The performance of specific factor grouping methods has not been extensively evaluated. However, from the research that did evaluate the performance of the land use and clustering analysis methods for factor grouping, the results indicated that the performance was not significantly different. Therefore, both the land use and clustering techniques are considered in this research. With no recommendation that was presented based off performance; it would be ideal to use a method that is relatively popular in the literature. Therefore, it is also recommended to utilize AMI and WWI as way to established factor groups for the given count sites (Eq. 2.18 and Eq. 2.19).
3. Computing Expansion Factors:
   For the calculations of AADT expansion factors , the more Traditional or Standard method for expanding short-term counts to an AADT estimate is to utilize a day-of-week and month-of-year scaling factor as previously mentioned (Eq. 2.6, Eq. 2.8 and Eq. 2.11). However, in various studies such as Hankey et al. (2014), Nosal et al. (2014), and El Esawey et al. (2016), the Traditional method seems to be outperformed by the Disaggregate method (Eq. 2.15 and Eq. 2.16) which ultimately uses a day of year scaling factor. Therefore, the Disaggregate factor method is selected as a benchmark expansion factor method selected for this research. Although, the Traditional method scaling factors and a set of other factors known as the day by month scaling factors (Eq. 2.13 and Eq. 2.14) are also computed. This is because it is expected that the 8hr counts could add another degree of variability in the expansion process which means it is unclear what expansion method is truly the best.

Chapter 3 Description of Empirical Data

3.1 Introduction

This chapter describes the empirical (field) data used in this thesis. First, the technology used to collect the data is described. Then, general geographical information about the counts sites is provided in Section 3.2.1. Section 3.2.2 presents an overview of the unfiltered data at each of the count locations. Lastly, Section 3.2.3 describes the relevant weather patterns that pedestrians could encounter at the selected count locations.

For this research, count data were obtained from a vendor having system deployments at intersections in both Milton, Ontario and Pima County, Arizona. The system collects data by processing images collected using a video camera system installed at the intersection. The system characterizes by user type (i.e., motorized vehicle, cyclist, or pedestrian) and motorized vehicles are classified by type of vehicle such as a passenger car, semi-truck, or bus. The counting system collects data in 1-minute intervals if intersection users are present for the given 1-minute interval. Road users such as (motorized vehicles and bicycles) are collected as traditional turning movement counts. Pedestrians are collected based off their motion around the center of the intersection (clockwise (CW) or counter-clockwise (CCW)) from a specific crosswalk side. Figure 3.1 shows a 4-legged intersection highlighting the center of the intersection and all the possible movements a pedestrian can make (8 total).

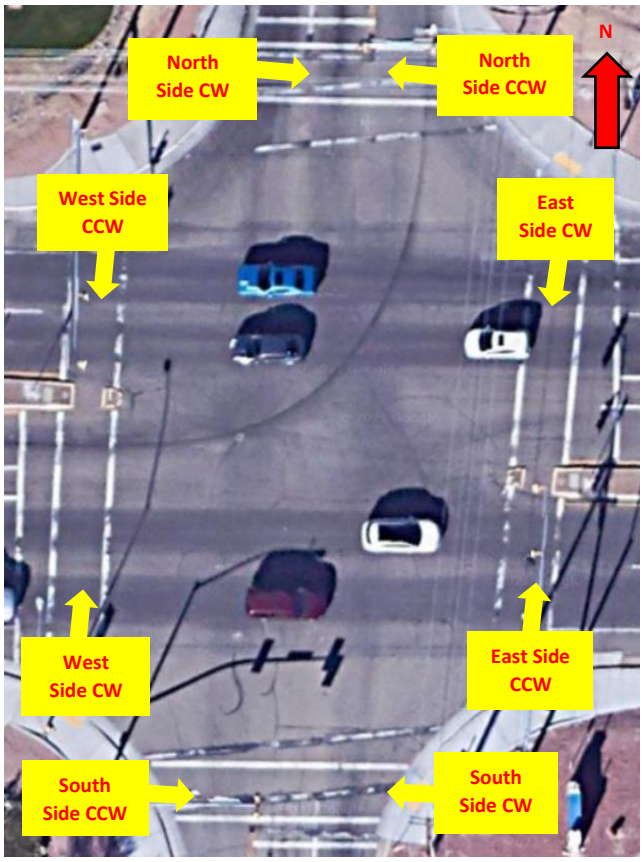

Figure 3.1 Pedestrian Crossing Diagram at the Ina Rd / Mona Lisa Rd Intersection in Pima County, Arizona (Source: Modified from Google Maps™, 2021)

3.2 Description of Observed Land Uses

The type of land use of the area surrounding each count site was determined using available sources, including Google Maps™, and categorized as one of: commercial, industrial, recreational, residential, and undeveloped. A commercial land use would be any section of land with office spaces, restaurants, and other retail establishments. An Industrial land use includes areas with factories, refineries, and warehouses. Recreational uses include sections of land with indoor/outdoor sporting/entertainment facilities and recreational trails that could be used for walking and/or biking. Residential areas include sections of a municipality that mainly consist of single unit dwellings, row houses and apartment buildings. As for undeveloped land, it can be within a rural area or a specific section of an urbanized area however, the section of land is not being used for any specific purpose. Examples of undeveloped land could include grasslands or wooded areas.

3.3 Description of Milton, Ontario

3.3.1 Geographic Information

Milton is a municipality in southern Ontario that is close to the city of Toronto, Ontario which is the most heavily populated urban center in the province and in Canada and its borders are outlined in red as depicted in Figure 3.2. The population of Milton was 110,128 in 2016 and has been growing for the last two census reports (Statistics Canada, 2017). With Milton's proximity to the city of Toronto, Milton is a community that is generally suburban. Milton covers a surface area equal to 363.22 square kilometers (Statistics Canada, 2017) however, the municipality itself has a rural and urban section. All of the count sites used in this study are located in the urbanized portion as presented in Figure 3.3.

Figure 3.2. Location of Milton, Ontario (Source: Google Maps™, 2021)

Figure 3.3. Count Sites within Milton, Ontario (Source: Google Maps™, 2021)

3.3.2 Count Data

There was a total of 30 count sites. However, there are 4 sites with no pedestrian data, so these sites have been excluded from further consideration. The collection periods for each site along with the intersection location and land uses are presented in the Appendix A. Figure 3.4 shows the average daily pedestrian volume from the 26 sites. The majority of sites experience more than 100 pedestrians/day, with the exception of 4 sites which have just below 100 and 5 sites in Milton are well below an average daily count value of 100. For the distribution of daily count values in Milton, Ontario all 26 sites had interquartile ranges between 0 to 1250. However, most of the sites have a Q3 value that is significantly less than 1250 and most of the sites have outlier daily count values as presented in Figure 3.5.

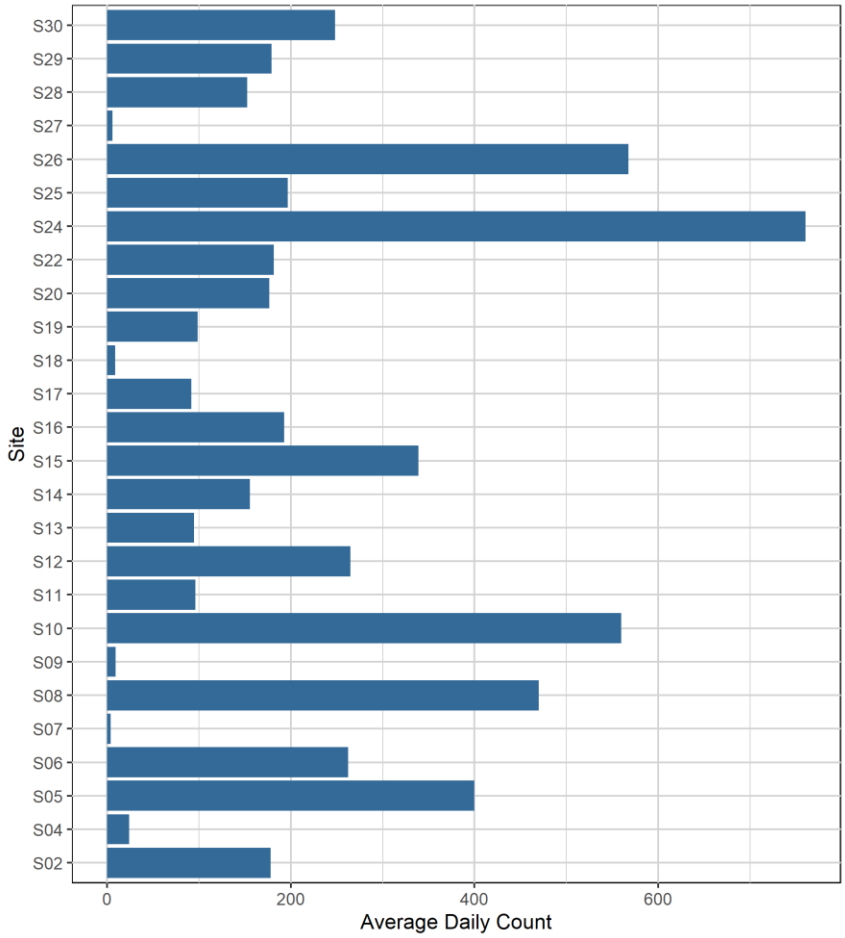

Figure 3.4. Average Daily Counts for Sites in Milton, Ontario

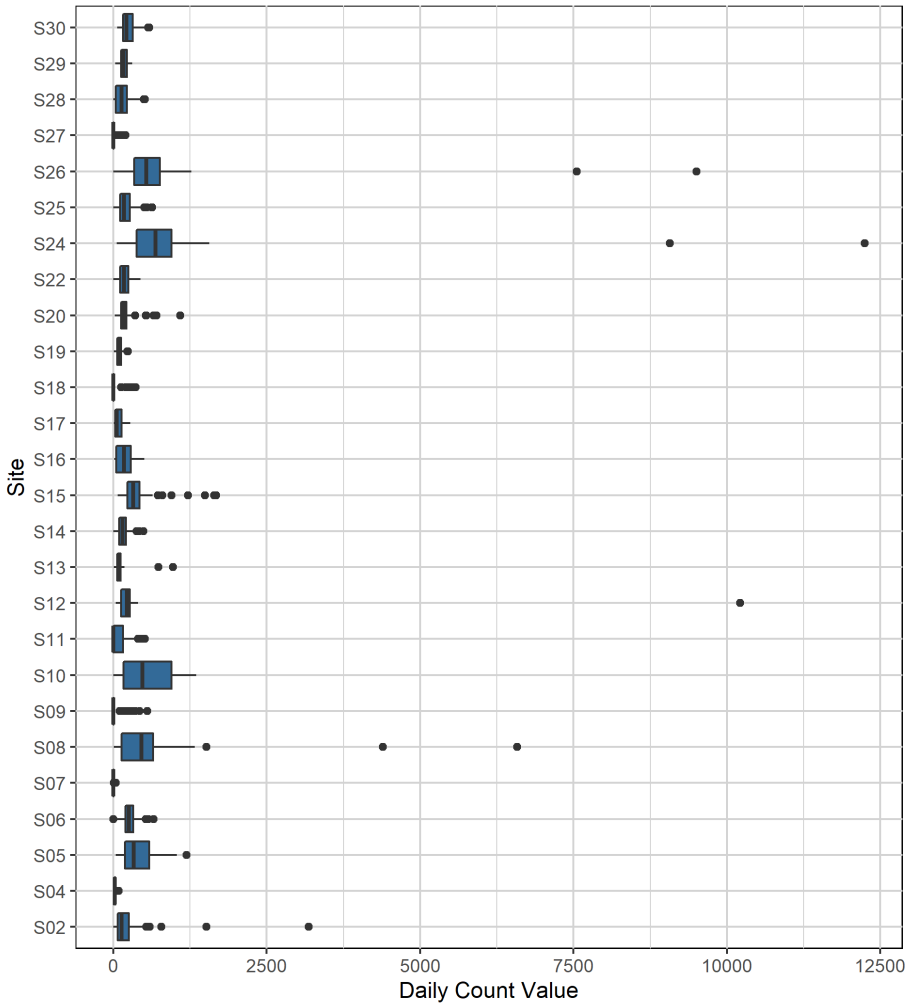

Figure 3.5. Daily Count Value Distribution for Sites in Milton, Ontario

3.3.3 Weather Data

For Milton average daily precipitation and average daily temperature were obtained from Environment Canada. The average daily precipitation in Milton, Ontario seems to show quite a bit of variability on a month-by-month basis. The minimum average daily precipitation occurred in the month of September and has a value just above 1 mm. The maximum average daily precipitation took place in October and the specific value is above 4 mm as shown in Figure 3.6. Average daily temperature is greater than 0°C for all months of the year except for January, February, March, and December. The warmest month is July with an average daily temperature above 20 (°C) and the coldest month is January having an average daily temperature just below -5 (°C) as presented in Figure 3.7.

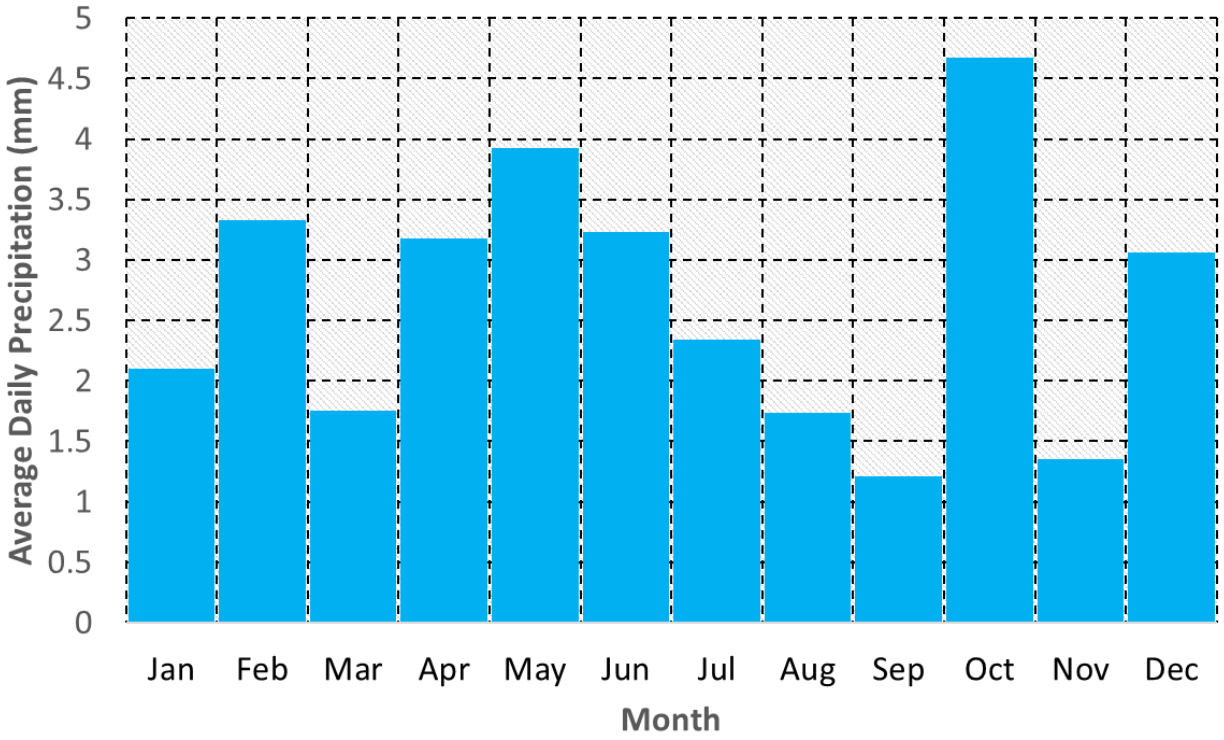

Figure 3.6. Average Daily Precipitation in 2019 for Milton, Ontario

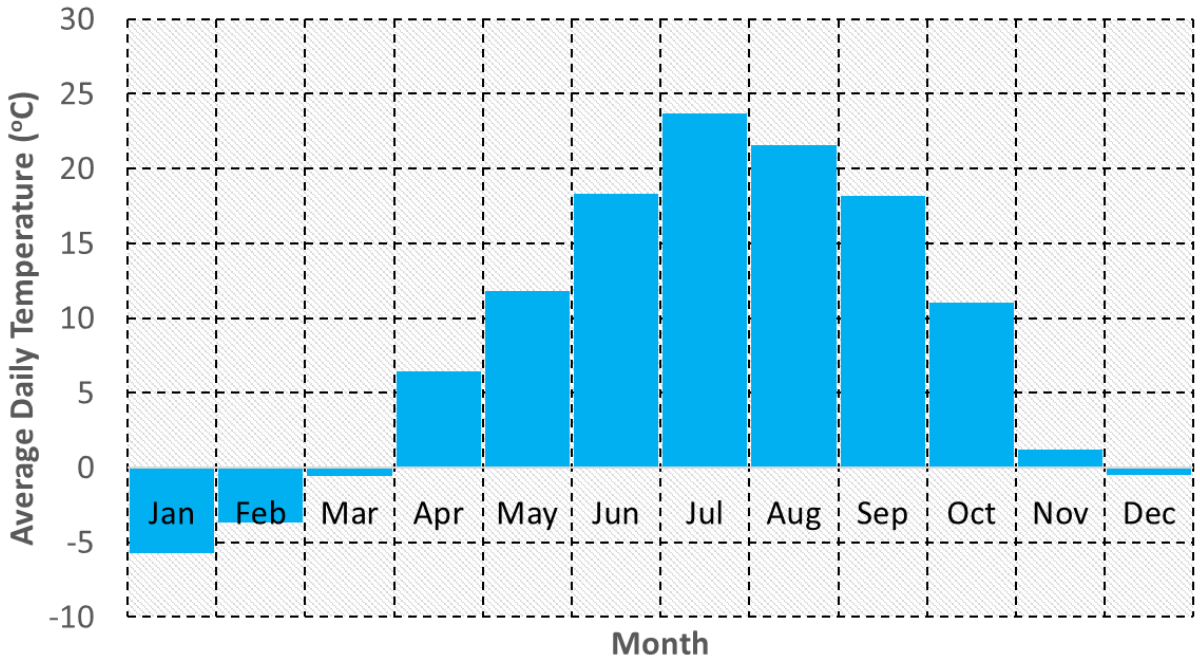

Figure 3.7. Average Daily Temperature in 2019 for Milton, Ontario

3.4 Description of Pima County, Arizona

3.4.1 Geographic Information

Pima County is in the state of Arizona's southern section and is adjacent to the United States and Mexico boundary and the county's borders are highlighted in red as presented in Figure 3.8. The population of Pima County, Arizona was 980,263 in 2010 and based off the last population estimate of 1,047,279 in 2019, it appears that the county itself is growing (U.S. Department of Commerce, n.d.). Pima County has a collection of different communities within its defined boundaries. The City of Tucson is the main community located within Pima County and the other population centers located in the county appear to be suburbs. The count sites used in this thesis are located as shown in Figure 3.9. Pima County is large geographically and has an overall surface area equal to 23,794 square kilometers (U.S. Department of Commerce, n.d.). A large section of Pima County's land mass is associated with the Tohono O'odham Nation Reservation and everything east of that land use is where Tucson and most of the suburban communities are located.

Figure 3.8. Location of Pima County, Arizona (Source: Google Maps™, 2021)

Figure 3.9. Count Sites within Pima County, Arizona (Source: Google Maps™, 2021)

3.4.2 Count Data

There are110 counts sites used for this study. A total of 15 sites in the dataset do not have any count data at all and these were eliminated from further consideration. The site name and location, surrounding land uses, and the data collection period for each site is provided in Appendix B. Figure 3.10 and Figure 3.11 show the average daily pedestrian volume and distribution of daily traffic volumes across for a sample of 34 sites. Most sites have an average daily traffic volume less than 75. The interquartile ranges of daily pedestrian volumes at these sites are mostly between 0 to 100.

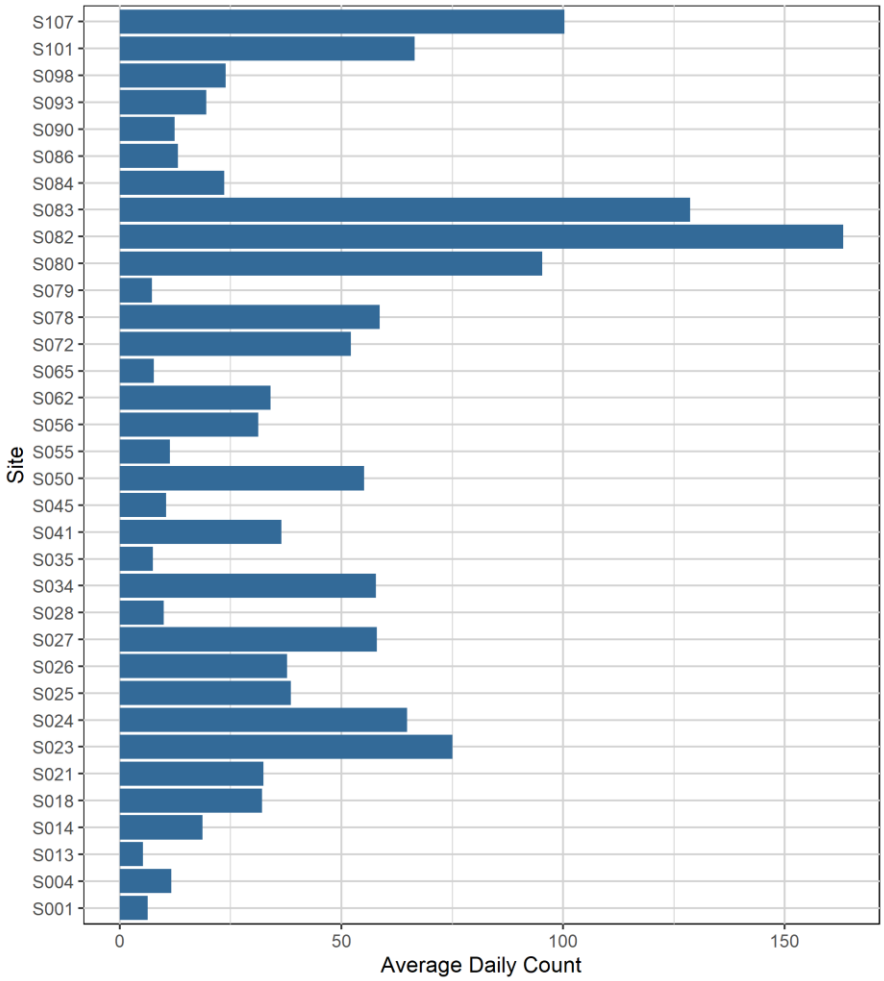

Figure 3.10. Average Daily Counts for Sites in Pima County, Arizona

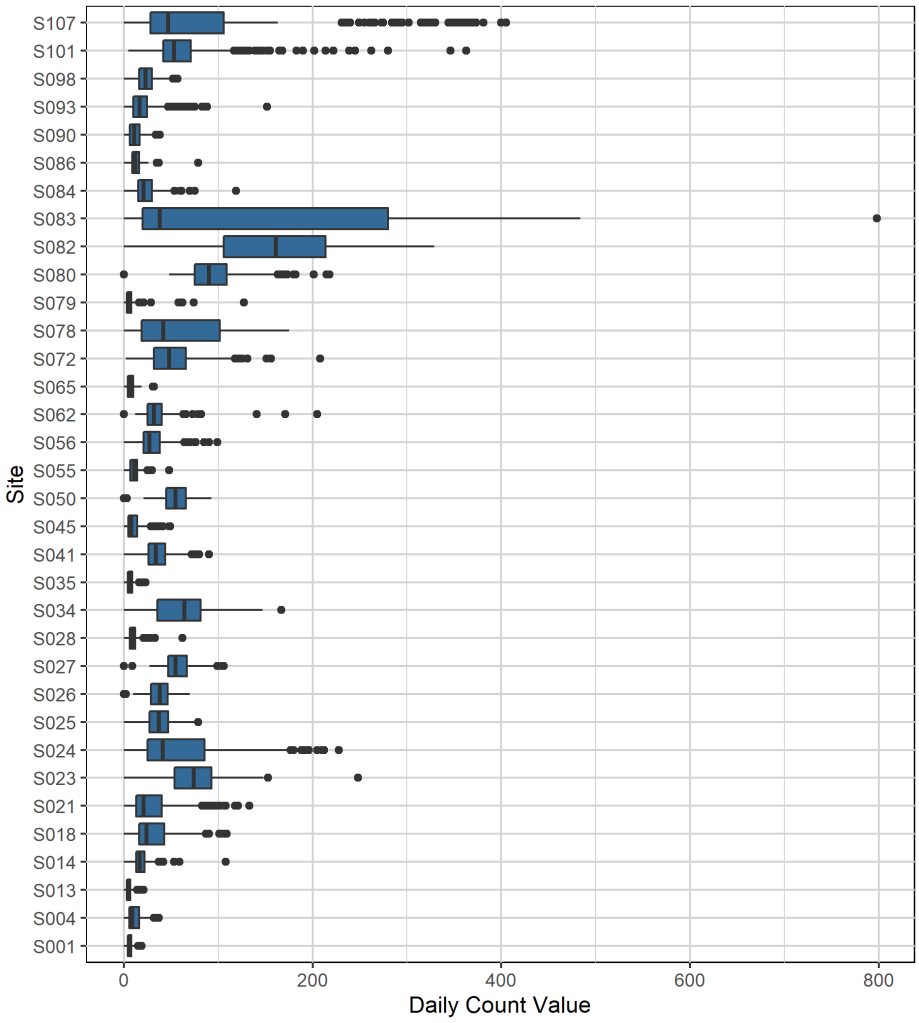

Figure 3.11. Daily Count Value Distribution for Sites in Pima County, Arizona

3.4.3 Weather Data

The average daily precipitation, average daily temperature, and average wind speed were obtained from the National Oceanic and Atmospheric Administration (NOAA). The average daily precipitation varies by month-of-year as shown in Figure 3.12. The month of June has the lowest average daily precipitation value (close to 0 mm) and the highest average daily precipitation value occurs in September where the value is slightly greater than 2 mm. For the average daily temperature in Pima County, Arizona, no month has an average daily temperature below 0°C and colder months occur during the winter season. February has the lowest average daily temperature and July has the highest average daily temperature which hover around values of 10°C and 30°C, respectively, as shown in Figure 3.13. For the average wind speed profile in Pima County, Arizona, the values for each month are similar throughout the year and hover around a speed of 3 km/h as shown in Figure 3.14.

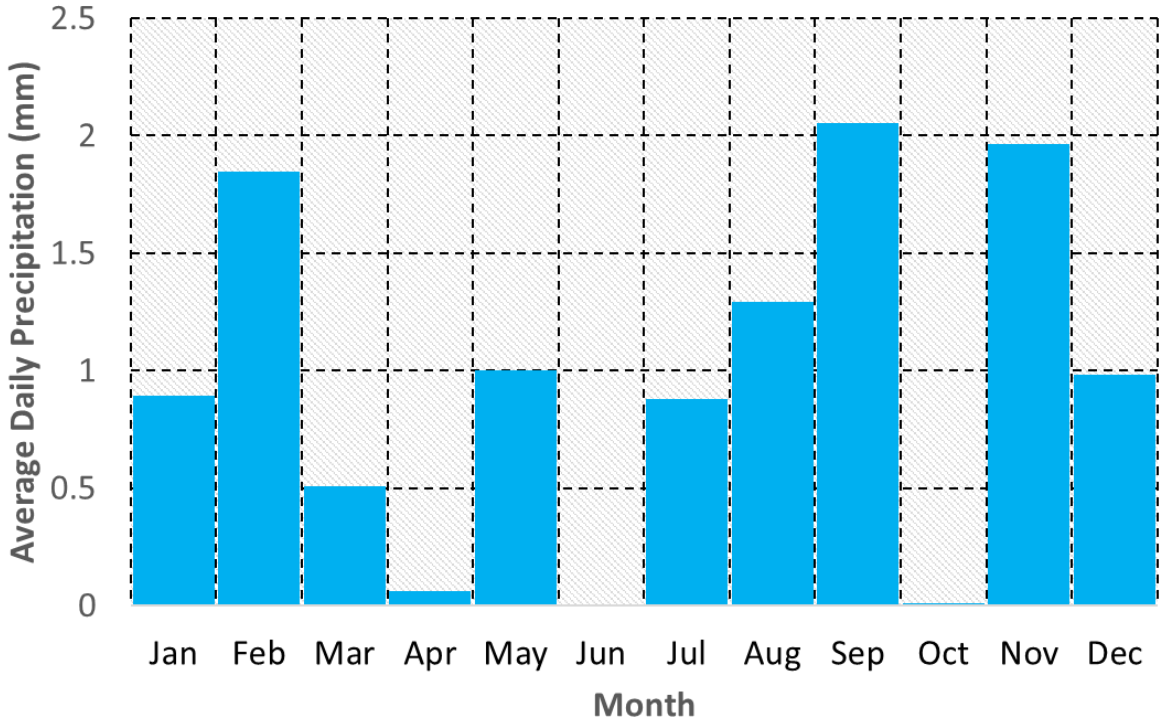

Figure 3.12. Average Daily Precipitation in 2019 for Pima County, Arizona

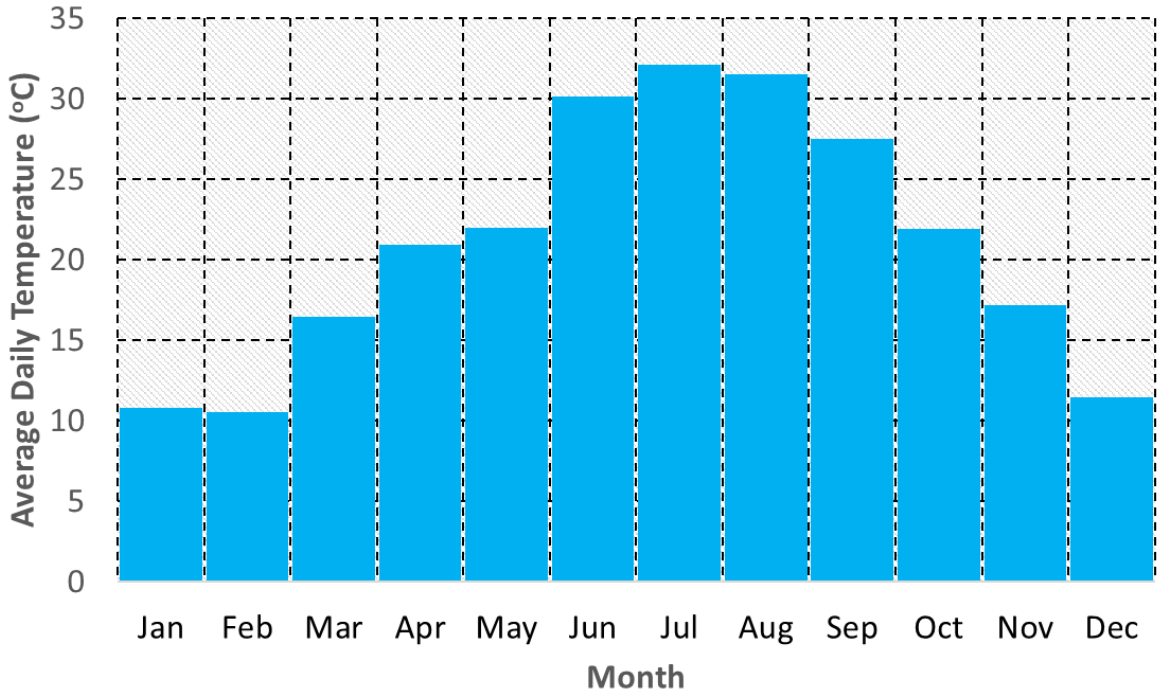

Figure 3.13. Average Daily Temperature in 2019 for Pima County, Arizona

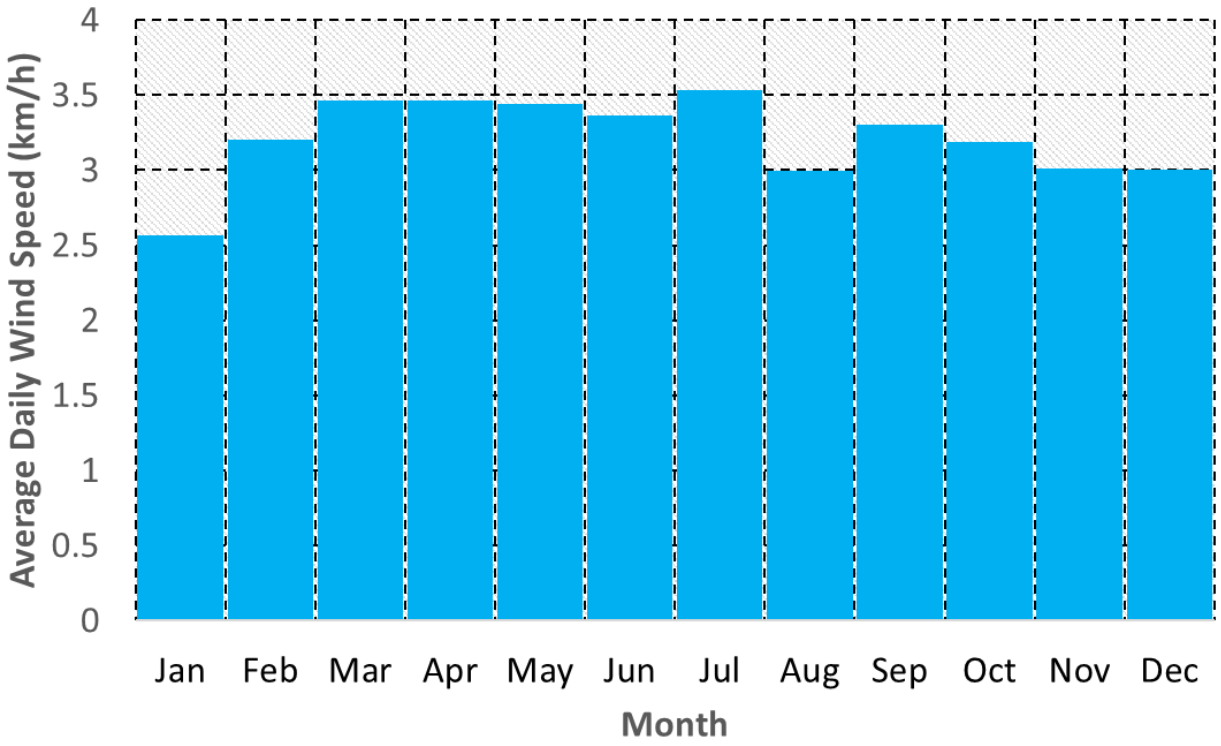

Figure 3.14. Average Daily Wind Speed in 2019 for Pima County, Arizona

Chapter 4 Data Filtering and Site Selection Criteria

4.1 Introduction

As mentioned in Chapter 1, it is generally expected that when a count dataset is received, it is more than likely to observe problematic entries. As an example, a null value is recorded to the database to indicate that a system error has occurred, and no count data has been received (e.g., a sensor error or communication error). It is also possible to have a null entry when the system is operating correctly, but no users pass the sensor and therefore the reported count is truly zero. There may also be specific duration of counts where all the counts for each time interval are equal to one specific non-zero count value, this might imply that there is something wrong with the counter because the counter itself could be stuck on the repeating value. Another example of a problematic entry within a dataset is having a relatively large count value corresponding to a relatively small time interval which could imply that some over counting by the system is occurring.

For some cases, there could be a relatively long duration of zero counts within a dataset, which could imply that the counter at the intersection is powered off and ultimately not collecting the data for the identified duration. It is also important to focus on eliminating exceptionally large count entries within specific timeframe of a given dataset because those entries do not reflect the temporal trend and could impact any further analysis if they are not removed.

This chapter describes various types of data filtering methods applied to the pedestrian count data. The filtering methods have been adapted from those suggested by Allen (2021) for application to cyclist count data and assuming count data were available for 15-minute count periods.

Allen (2021) proposed five separate sequentially applied filters as follows:

1. Null count filter
2. Non-consecutive zero filter
3. Hard cap filter
4. Daily and 8hr zero filter
5. IQR filter

The first three were developed for data aggregated into 15-minute counts which includes a filter for null counts, repeating non-zero values, and a hard cap limit. The other two filters deal with data both on the 8-hour and daily aggregate levels and more specifically, one of these filters removes data outside of a given interquartile range and other filter removes 8-hour and daily zero values. The set of filters, and how they have been modified for this work, is described in more detail in the next part of this thesis and then the filters are applied to the Milton and Pima County pedestrian count data. The purpose of this chapter is to address research problem **2**, which deals with the retention of cleaned count data.

4.2 Proposed Filtering Method

4.2.1 Null Count Filter

The system from which the data used in this thesis were obtained does not make a distinction between a null entry occurring caused by an issue with the counting device and a null entry that is the result of no users present for a given time interval therefore, it is not clear if a problem is existent. The null count filter proposed in this research also acts as a replacement for both the traditional null count filter, where

a given entry has its timestamp recorded but the count value is blank and as well as the consecutive zero filter which was also known as filter f2.

The likelihood of observing zero counts decreases as the average traffic volume increases and as the count period duration increases. As stated before, pedestrian (and cyclists) volumes at intersections tend to be much smaller than the motorized traffic volumes at the intersection. The system from which the count data were obtained also provided motorized traffic counts. It was hypothesized that a system failure or error would impact all counts (i.e., motorized vehicles and pedestrians) and therefore even a relatively short period of zero counts (combination of vehicles and pedestrian counts from all turning movements) would be indicative of a system error. It was proposed that if the total traffic volume from all turning movements combined was equal to zero for three or more consecutive 15-minute intervals, then it was concluded a system error had occurred and those count intervals were flagged as missing data.

It is important to realize that the term "missing counts" is referring to a collective representation of all modes of transportation that are not recorded for a given 15-minute time interval. If any roadway traffic is recorded at a given 15-minute interval (any count > 0) and pedestrians are not recorded at the same 15-minute interval, the timestamp itself, is not considered a missing count and the count value for pedestrians is truly zero.

4.2.2 Non-Consecutive Zero Filter

To deal with the issue of non-consecutive zero counts within a dataset, the non-consecutive zero filter uses a Poisson distribution to validate the likelihood of a specific non-zero count being repeated for x number of succeeding intervals. If a given series of the same non-zero value is deemed to be unlikely, the filter flags all those 15-minute entries. It is also possible to have valid duration of non-zero counts repeating if the recording time intervals are relatively short. In that specific case, it is also more than likely that the deviation of count values are smaller, which could imply that certain count values have a higher frequency for certain hours of the day. Although, if the duration of repeated non-zero counts increases, the likelihood of each recorded count in that specific period being valid decreases.

4.2.3 Hard Cap Filter

The purpose of the hard cap limit is to mark 15-minute data entries that appear to be unreasonably large. Therefore, it is important to select an appropriate hard cap for a set of 15-minute counts and the hard cap value itself is dependent on the range of typical count values for daily totals. The filter itself is not intended to eliminate too many count entries because the targeted entries for the filter itself are just strictly upper outliers. If there is ever a case where too many 15-minute entries are greater than the hard cap, that implies that the flagged values are not all outliers, and the hard cap limit should be increased. Hard cap limits of 500 pedestrians and 250 pedestrians were selected for the Milton and Pima County data, respectively and were chosen based of the observed daily totals. If the daily totals are generally less than 100, the hard cap should be set to 250 and If the daily totals are mostly between 100 to 500, the corresponding hard cap limit would be 500 (Allen, 2021). Any 15-minute counts exceeding those pre-defined limits are flagged.

4.2.4 Daily and 8hr Zero Filter

For eliminating long durations of consecutive zero counts, an 8-hour and 24-hour threshold was utilized. However, a long duration of zero counts could still be equal to zero for every time interval and could

happen if there is something such as construction work present at the intersection and pedestrians cannot use the intersection. Although, the construction work situation does represent a valid case of having a long duration of zero counts, the data itself is not useful since the intersection is out of service. It is also possible to observe a long series of zero counts at an intersection that is located in an area that typically does not attract pedestrians and/or does not have suitable pedestrian infrastructure in place. An example of that is an intersection in a rural area with no sidewalks and even though a long duration of zero pedestrians is likely to be observed, it is still important to remember that long durations of zero are not useful. This is because the data only produces expansion factors equal to zero. If an expansion factor equal to zero is used to expand a non-zero STC from a different site, the estimated daily value for the other site is equal to zero when in reality the true daily value is a non-zero count.

4.2.5 IQR Filter

For removing large data entries that are clearly different from typical counts, the IQR filter makes use of a moving 27-day window and the interquartile range times two plus the third quartile as its boundaries to flag count values that are potentially problematic. The moving 27-day window was utilized in Allen (2021) algorithms and it also assumed that the 27-day window has been applied elsewhere. As an example, if the day of interest is January 15th, the mobile 27-day window would consider the succeeding and previous 13 days which is (January 16th to January 28th) and (January 2nd to January 14th) respectively. Therefore, it is important to realize that the IQR filter is somewhat like the hard cap filter in the sense that the IQR also focuses on upper outliers and should not be eliminating too many values. However, depending on the typical range of 8-hour and daily values, the IQR could either be relatively large or fairly small. If the typical 8-hour and daily values fit within a relatively small range, the IQR too is going to be small and if the IQR is small enough, it is possible to observe flagged entries that do not visually appear to be outliers and/or an excessive quantity of entries being caught by the IQR filter. That limitation could imply that the IQR filter is not effective in dealing with datasets with relatively low count values and that some sort of IQR limit should be implemented. For the Milton data, the issues of small count values were a relatively nonexistent problem. However, for the Pima County data, the problem was much more prevalent, and it was decided to implement a lower IQR cut off value that is equal to 50.

4.3 Application of Filtering

4.3.1 Retention of Cleaned Data

Although the daily and 8 hour data have their own specific filters, it is important to realize that the daily and 8 hour data are also impacted by the filters specifically designed to flag problematic entries at the 15-minute level. Furthermore, problematic data with a duration of 15 minutes impacts all calculated values even if the data aggregation level is not equal to a day or 8 hours. This is because all data entries are composed of at least one or more specific 15-minute count values and as an example, it is quite possible for specific daily or 8 hour counts to not be deemed problematic but, a 15-minute count within that daily or 8 hour period is flagged as problematic, or the opposite case could be true. Therefore, filtering thresholds need to be established and more specifically, problematic entries are not permitted to be used for any further analysis. The only exception to that requirement is that a daily total can have less than 20 15-minute null counts present and the reason for it is because it is quite possible that the flagged null counts could actually be equal to zero therefore, the daily total is not necessarily an uncounted value. There was also an investigation in the count data for Pima County to see if it was worthwhile to increase some of the thresholds for the utilized filters. However, after a closer

examination of the data, it was hypothesized that adjusting thresholds do not appear to significantly increase the quantity of count sites with a reasonable quantity of cleaned count data.

4.3.2 Milton, Ontario Filtering Results

Table 4.1 shows the average percentage of daily, 8 hour and 15-minute entries flagged as problematic for all sites that have data between in July 2019 to February 2020 in Milton, Ontario. From that specific table, very few entries are impacted by the initial filtering process. The average percentage of flagged entries is almost close to 0% with no % value exceeding 5%. For the average percentage values closest to 5%, those values specifically corresponded to daily and 8 hour entries flagged by the daily and 8 hour zero filter. However, for the percentage of daily, 8 hour and 15-minute entries flagged by site as presented in Appendix C, it is clear to see that one specific site has more than 30% of its daily entries flagged and more than 50% of its 8 hour entries flagged as problematic by the daily and 8 hour zero filter. With that specific site corresponding to those percentage values, it is expected that the site cannot be used for further work in this research. However, 12 remaining sites generally have high percentages of cleaned data so, those specific sites should be considered for further analysis in this study.

Table 4.1. Aggregated Filtering Results for Milton, Ontario (July 2019 to February 2020)

Filter	% Flagged (Avg across all sites)
Null Counts (Daily Totals f1)	0.7
Non-Consecutive Zeros (Daily Totals f3)	0.1
Hard Cap (Daily Totals f4)	0.3
Daily Zeros (Daily Totals f5)	2.8
IQR (Daily Totals f6)	0.9
Null Counts (8hr Totals f1)	0.9
Non-Consecutive Zeros (8hr Totals f3)	0.1
Hard Cap (8hr Totals f4)	0.1
8hr Zeros (8hr Totals f5)	4.7
IQR (8hr Totals f6)	1.1
Null Counts (15min Totals f1)	1.3
Non-Consecutive Zeros (15min Totals f3)	0.01
Hard Cap (15min Totals f4)	0.02

After reviewing the general filtering results, it is important to validate them. Figure 4.1 illustrates an example in which six consecutive 15-minute intervals (highlighted in blue) have the same count and are flagged by the filter (f3) as suspect. In the context of the time series of reported count values both before and after these points, it is not clear that these values are erroneous and that this filter is correctly identifying suspect data. However, given how few data are marked as suspect by the filter, no attempt was made to revise or validate this filter as this is likely have almost no impact on future research steps.

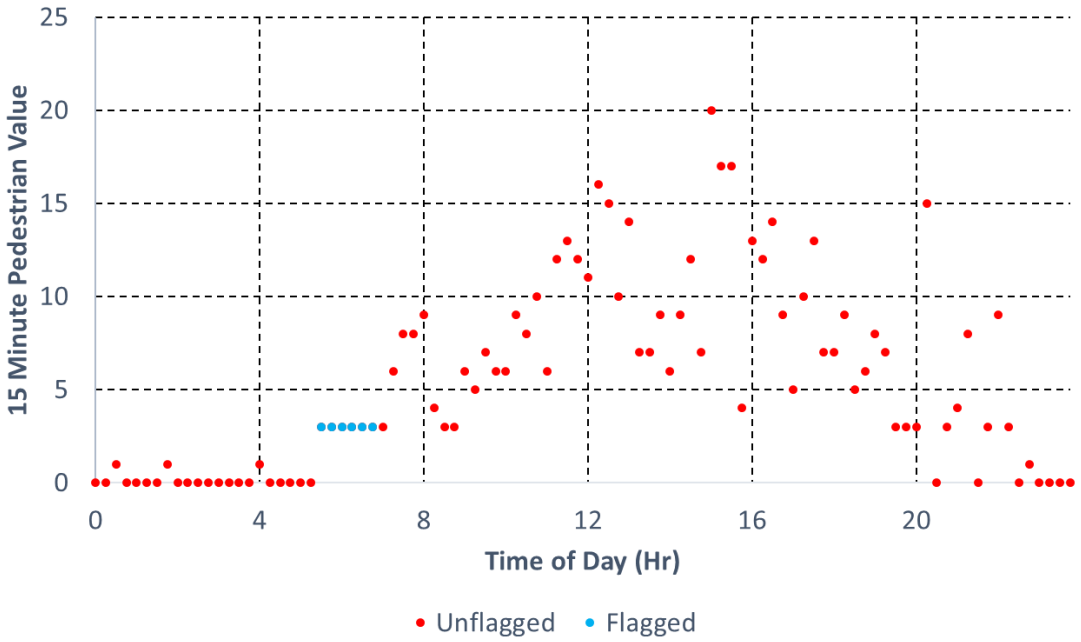

Figure 4.1. Example of Data Entries Flagged by Filter f3 (Site Code: S15)

With the limited quantity of entries being flagged with the Milton dataset, it is probably and good indication that the hard cap (f4) value is acceptable. This is because the filter seems to be eliminating 15-minute counts that are above the hard cap limit (500) and much larger than the typical 15-minute counts being collected as highlighted by the blue markers in Figure 4.2.

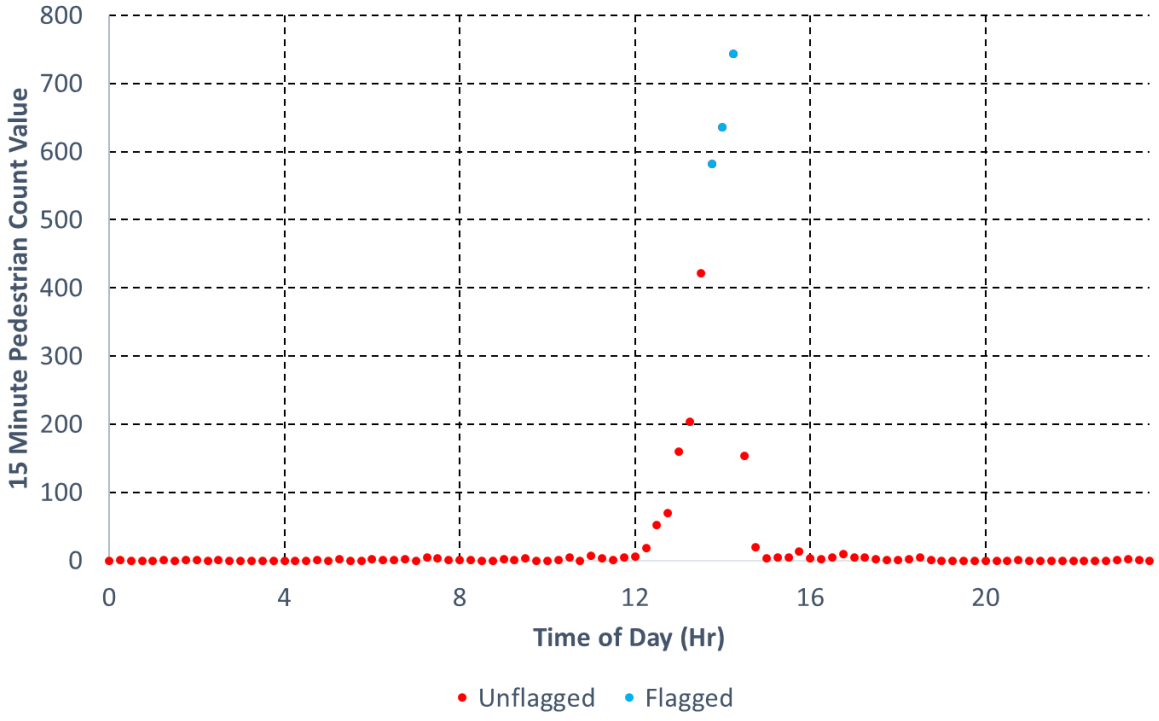

Figure 4.2. Example of Data Entries Flagged by Filter f4 (Site Code: S02)

For the IQR (f6) filtering example (Figure 4.3) shown below, there are two specific daily counts that visually appear to be an extreme outlier within its specific 27-day window and is ultimately flagged by the IQR filter (the two highest daily counts highlighted in blue). From this, there is reason to believe that one of the daily entries is flagged by the IQR filter because that daily entry corresponds to the same date as the 2019 Milton, Ontario Santa Clause parade. That type of event is known to significantly increase pedestrian traffic, which ultimately generates a recorded count value much larger than normal. There is also a couple more daily entries caught by the IQR filter (remaining daily counts highlighted in blue) that do not appear to be much different from its surrounding daily totals. However, when viewing the entire dataset for the site, it is not clear if those values are extreme outliers but removing those entries should not make much of a difference.

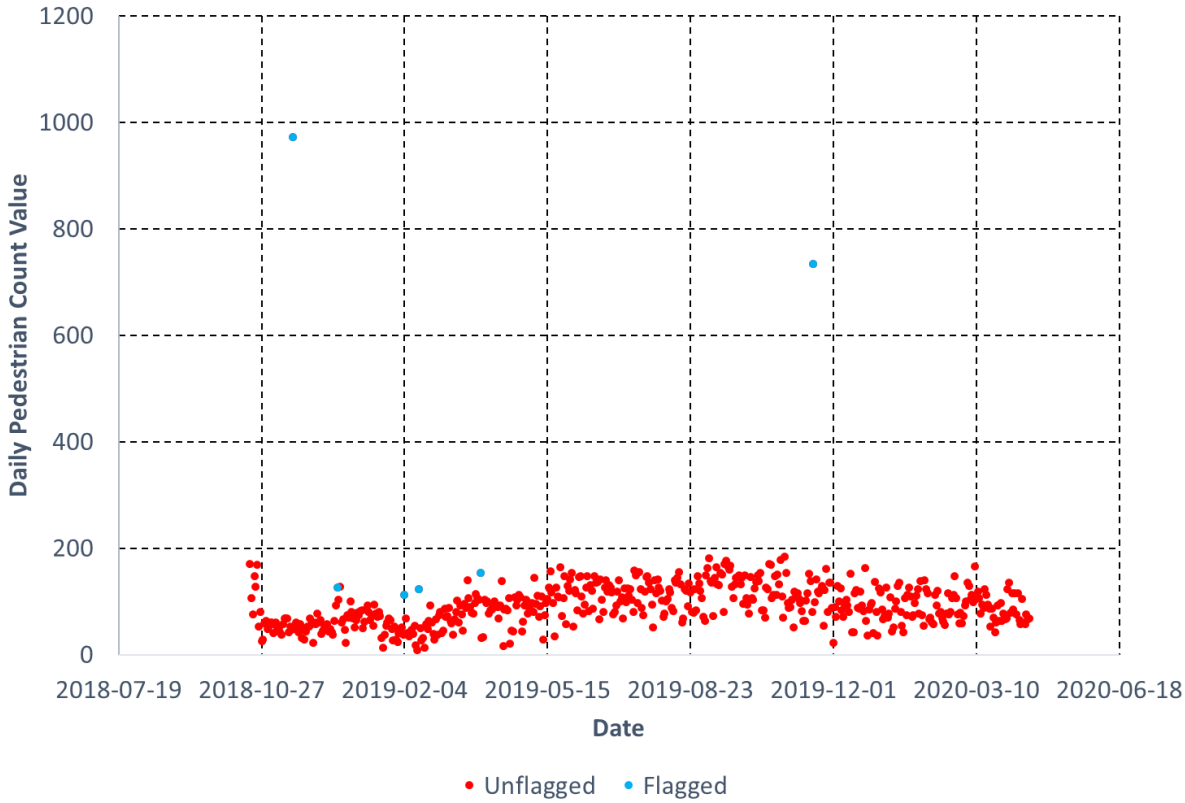

Figure 4.3. Example of Data Entries Flagged by Filter f6 (Site Code: S13)

4.3.3 Pima County, Arizona Filtering Results

The results in Pima County, Arizona showing the average percentage of daily, 8 hour and 15-minute entries captured within the filtering process for sites with data present between January 2020 to March 2020 are presented in Table 4.2. Within that table, no data entries were flagged by the non-consecutive zero and hard cap limit. Furthermore, an exceedingly small percentage (almost equal to 0%) of daily and 8 hour entries were flagged as problematic by the IQR filter. For the null count filter, the percentage of daily, 8 hour and 15-minute entries flagged hovered around 10% and for the daily and 8 hour zero filter, the percentage of data entries captured was close to 20%. The filtering results in Pima County, are not overly surprising given the fact that quite a few sites have exceptionally low pedestrian traffic volumes. In Appendix D where the same results are shown by site, a few sites have almost all of their data entries

flagged by the null count filter and the daily and 8 hour filter. Therefore, those specific results imply that those sites are not useful in this research.

Table 4.2. Aggregated Filtering Results for Pima County, Arizona (January 2020 to March 2020)

Filter	% Flagged (Avg across all sites)
Null Counts (Daily Totals f1)	11.8
Non-Consecutive Zeros (Daily Totals f3)	0
Hard Cap (Daily Totals f4)	0
Daily Zeros (Daily Totals f5)	22.4
IQR (Daily Totals f6)	0.3
Null Counts (8hr Totals f1)	10.4
Non-Consecutive Zeros (8hr Totals f3)	0
Hard Cap (8hr Totals f4)	0
8hr Zeros (8hr Totals f5)	23.4
IQR (8hr Totals f6)	0.1
Null Counts (15min Totals f1)	8.7
Non-Consecutive Zeros (15min Totals f3)	0
Hard Cap (15min Totals f4)	0

4.4 Pedestrian Data Study Period

After the effectiveness of the filters have been examined and determined to be appropriate for the given dataset, the next specific focus of this research is to select a study period. More specifically, it is ideal to select a study period that is equal to an entire year and include a relatively large number of count sites. Doing so is important because the study period captures the different variations of temporal trends that occur within a year for a given location. Having many count sites provides more options for considering numerous factor grouping setups and provides more results that could further validate this research. The one problem present in all datasets is that considering a cleaned full year of data leaves a limited quantity of sites to work with. Therefore, the study period then must represent a specific season instead of a full year to have more optimal quantity of sites to work with.

Another issue that must be considered when selecting a study period is that a relatively large portion of count data obtained from both Milton, Ontario and Pima County, Arizona was collected just prior to and at the start of the COVID-19 global pandemic. Therefore, it is expected that the pandemic significantly reduces traffic volumes for all modes of transportation. An example of this is shown in Figure 4.4 where the figure itself presents daily pedestrian data from March 2020 at site in Pima County located near a school. More specifically, the daily pedestrian totals for the first half of March are typically much larger than the counts from the last half of the month. With all that considered, the study period for Milton, Ontario starts in July 2019 and goes to February 2020 (8 months) and as for Pima County, Arizona, the study period begins in January 2020 and ends in March 2020 (3 months).

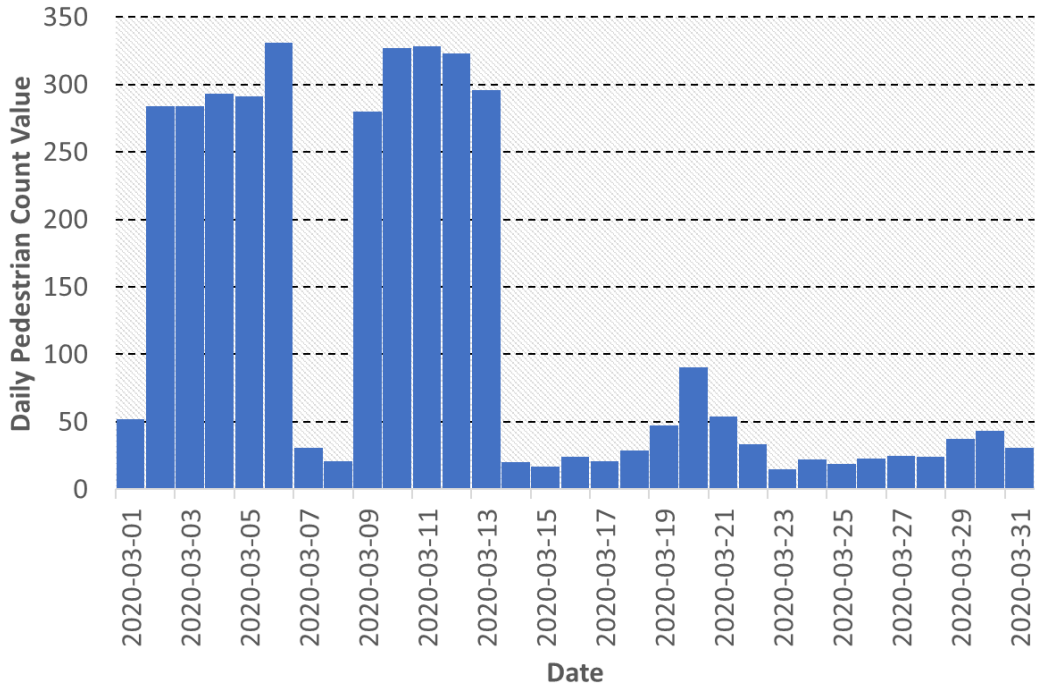

Figure 4.4. An Example of the Impact of COVID-19 on Daily Count Totals (Site Code: S083)

4.5 Site Selection Requirements and Short Term Count Criteria

After selecting a study period for both geographic locations, it is now important to determine if the sites that have data collected within the specific study periods meet the requirements to calculate AMI, WWI, and the observed/true SADPT and expansion factor values. For a given site to have sufficient data for calculating AMI, all the components of Eq. 2.18 need to be available. For WWI, Eq. 2.19 is utilized however, the requirements for calculating WWI in this work are more extensive. Therefore, to calculate WWI in this research, all sites need daily count values that cover all days of the week. Since it was suggested to utilize the AASHTO method for calculating a value such as AADT/SADT, sites in Milton need at least one daily entry for 56 types of daily entries and sites in Pima County need a minimum of one daily entry for 21 unique daily entries. The required daily entries for both datasets are representative of a specific day-of-week and month-of-year combination hence (8 months x 7 days = 56) for Milton and (3 months x 7 days = 21) for Pima County. Once the determination is made that a given site passes the AASHTO requirements for AADT/SADT, most of the expansion factors can be calculated. The only exception to that statement is for the calculation of the k-factor. To determine if a site has adequate data for the k-factor computation, the AASHTO requirements should be applied to the 24hr to 8hr ratios as well. After applying all site selection criteria, 12 sites are available for Milton, Ontario and 25 sites for Pima County, Arizona.

For selecting short-term counts from specific count sites, it is important to remember that turning movement counts are collected during specific months of the year, days of the week and times of the day because certain time periods better reflect the peak usage of hourly vehicular traffic at a specific intersection. Therefore, pedestrian counts at intersections are also representative of the same timeframes since the pedestrian data are collected alongside the vehicular data. For both Milton, Ontario and Pima County, Arizona the times of day to collect count data includes 7am to 9am, 11am to

2pm and 3pm to 6pm occurring on Tuesdays, Wednesdays, and Thursdays. From the constraints determined by the study periods in Milton and Pima County, short-term counts in Milton are collected from September to November and in Pima County, the extraction period is from January to March. It is also important to avoid selecting short-term counts on holidays and on days impacted by COVID-19. This is because it is quite possible that the count values corresponding to those specific days are not at all like the other count values for any given site which was ultimately demonstrated in Figure 4.4. If any input data used in the SADPT calculations considers holidays and/or COVID-19 days, it is possible that the SDAPT estimates are unrepresentative of the true traffic conditions. The list of holidays and days impacted by COVID-19 are presented in Table 4.3 below.

Table 4.3. List of Holidays Impacting Pima County, Arizona

Holiday	Date
New Year's Day	Jan 01, 2020
Winter Break for Schools	Jan 02, 2020 to Jan 03, 2020
Civil Rights Day	Jan 20, 2020
Presidents Day	Feb 17, 2020
Spring Break for Schools	Mar 16, 2020 to Mar 20, 2020
COVID-19	Mar 21, 2020 to Mar 31, 2020

Another aspect to consider when selecting short-term counts is the actual weather conditions which is especially important for pedestrians because of its sensitivity to it. In Table 4.4 and Table 4.5, the weather conditions for both Milton and Pima County are listed and if a specific short-term count is to be utilized for further analysis, the short-term count must pass the listed weather conditions for its geographic location.

Table 4.4. Weather Conditions for Milton, Ontario

Weather Variable	Condition to Pass
Total Daily Rainfall	<= 10mm
Average Daily Temperature	>= -5°C
Total Daily Snow on the Ground	<= 1cm

Table 4.5. Weather Conditions for Pima County, Arizona

Weather Variable	Condition to Pass
Average Daily Windspeed	<= 30km/h
Average Daily Temperature	<= 35°C
Minimum Daily Temperature	>= -5°C
Maximum Daily Temperature	<= 40°C
Total Daily Precipitation	<= 10mm
Total Daily Snow on the Ground	<= 1mm

Chapter 5 Methodologies for Estimating SADPT/AADPT

5.1 Introduction

The main objective of this chapter is to address research problem 4, which is associated with the conversion of short-terms counts to estimate an average daily pedestrian count value for a given timeframe. To fulfill research problem **4**, it is important to understand that expansion methodologies for short-term pedestrian counts have already been developed as stated in Chapter 2. More specifically, those expansion methodologies have been recommended and are known as the Traditional (Eq. 2.6, Eq. 2.8, and Eq. 2.11), AASHTO (Eq. 2.13 and Eq. 2.14), and Disaggregate (Eq. 2.15 and Eq. 2.16) methods. However, if those methods are left as is (no modifications made), they can only be applied to short-term pedestrian counts equal to 24 hours. With the short-term pedestrian counts being equal to a duration of 8 hours in this research, a modified version of those expansion methods is needed if there is any attempt to use them.

As previously mentioned, the literature does provide suggestions for dealing with situations where the short-term counts are less than 24 and those suggestions ultimately imply that modifying daily expansion methods is acceptable. Chapter 2 also suggests that the Disaggregate method has the best performance out of the recommended methods however, those comparisons were completed with daily totals. As previously implied, it is not clear what expansion methodology is best for 8hr pedestrian counts. The mechanics of each expansion method are explained throughout the chapter which could provide some insights for determining the best expansion methodology for 8hr pedestrian counts. Although, the performance of each method is not examined until the next chapter. It should also be noted that the remaining analysis in this thesis does not include any pedestrian count data from Milton, Ontario. It was determined that there were not enough sites remaining after going through the site selection criteria. Therefore, only sites in Pima County, Arizona are being utilized moving forward with this work.

5.2 Computation of Seasonal Values

As previously stated in this thesis, a seasonal average daily pedestrian traffic value is computed from all remaining sites in Pima County, Arizona that includes dates ranging from January 1st, 2020 to March 31st, 2020. It was also mentioned that a modified version of the AASHTO method is used to compute the seasonal average daily pedestrian traffic value and is presented in Eq. 5.1 shown below.

$$SADPT = \frac{1}{7}\sum_{d=1}^{7}[\frac{1}{3}\sum_{m=1}^{3}(\frac{1}{t}\sum_{c=1}^{t}V_{dmc})] \quad \text{Eq. 5.1}$$

where:

m = A specified month of year

d = A specified day of week

c = An appearance value for a specified day *d* and month *m*

t = Total quantity of a specified day *d* and month *m*

V_{dmc} = 24 hour pedestrian volume corresponding to appearance *c* on day *d* within month *m*

$SADPT$ = Seasonal Average Daily Pedestrian Traffic

From Eq. 5.1, it is clear that the only difference between it and the AASHTO method equation presented in the TMG is that the number of months considered is 3 within Eq. 5.1 and is 12 months (a full calendar year) within the TMG (Eq. 2.3). However, both versions of the AASHTO method have the same mathematical structure and consider all 7 days of the week. One of the main concerns with calculating a seasonal average verses an annual average is that a season average does not capture as much seasonal temporal variation as an annual average . As listed in Table 4.3, data collected between March 16th, 2020 to March 31st, 2020 is not considered for SADPT estimations however, that specific data restriction does not apply to the computation of the true SADPT value. Although, it was mentioned that including count data collected from March 16th, 2020 to March 31st, 2020 is concerning because of the COVID-19 pandemic and March school break however, that data are not necessarily invalid. It is just expected that the count values at some locations are lower than normal. As an example, a pedestrian count site could be located near a school and by default, the pedestrian volumes are influenced by the presence of the school. If that is the case, it is reasonable to assume that the pedestrian volumes during the summer months are much lower than months during the school year which ultimately influences the site AADPT value. The COVID-19 pandemic and March break could also have a similar effect on SADPT values for sites in Pima County and could ultimately capture some temporal variation.

Another factor, the k-factor, is needed because the utilized short-term counts consider less than a day's worth of data. The calculated k-factor for each site is incorporated within all the SADPT expansion methods discussed in Section 5.3. However, to calculate the k-factor for a given site, Eq. 5.2 and Eq. 5.3 outline the overall calculation process.

$$R_{dmc} = \frac{V_{dmc}}{V_{dmc(8hr)}} \quad \text{Eq. 5.2}$$

$$k = \frac{1}{7}\sum_{d=1}^{7}[\frac{1}{3}\sum_{m=1}^{3}(\frac{1}{t}\sum_{c=1}^{t}R_{dmc})] \quad \text{Eq. 5.3}$$

where:

m = A specified month of year

d = A specified day of week

c = An appearance value for a specified day *d* and month *m*

t = Total quantity of a specified day *d* and month *m*

V_{dmc} = 24 hour pedestrian volume corresponding to appearance *c* on day *d* within month *m*

$V_{dmc(8hr)}$ = 8 hour pedestrian volume corresponding to appearance *c* on day *d* within month *m*

$SADPT$ = Seasonal Average Daily Pedestrian Traffic

R_{dmc} = Ratio 8 hour to 24 hour pedestrian volumes for appearance *c* on day *d* within month *m*

k = K-factor value

From Eq. 5.3, it is again obvious to see that it is almost identical to the SADPT computation (Eq. 5.1) used in this research. This is because both formulations consider a modified version of the AASHTO method and both formulations do not consider STC requirements however, the ratio of 24hr to 8hr counts replaces the 24hr count term in the k-factor equation. It is also probably reasonable to expect that the ratio of 24hr to 8hr counts might be different during various months of the year and days of the week. As stated previously, the AASHTO method is mathematically set up to handle monthly and daily variations which ultimately implies that using the AASHTO method to compute the k-factor is appropriate.

5.3 Expansion Methodologies

5.3.1 The Traditional Method

The first expansion methodology that is considered in this research is the Traditional method. The Traditional method is commonly used in motorized traffic estimation studies but has also been used in non-motorized traffic estimation studies as well. As stated in the literature, the Traditional method itself requires a month-of-year and day-of-week scaling factor and to calculate those specific scaling factors, a monthly average daily traffic and an average day-of-week traffic value are required. Eq. 5.4, Eq. 5.5, Eq. 5.6, Eq. 5.7, and Eq. 5.8 are needed to produce the SADPT estimates for the Traditional method are listed below.

$$EstSADPT_x = k \times STC_x \times DWSF_d \times MYSF_m \quad \text{Eq. 5.4}$$

$$DWSF_d = \frac{SADPT}{ADPT_d} \quad \text{Eq. 5.5}$$

$$ADPT_d = \frac{1}{q}\sum_{a=1}^{q} V_{da} \quad \text{Eq. 5.6}$$

$$MYSF_m = \frac{SADPT}{MADPT_m} \quad \text{Eq. 5.7}$$

$$MADPT_m = \frac{1}{7}\sum_{d=1}^{7}[\frac{1}{t_{dm}}\sum_{c=1}^{t_{dm}} V_{dmc}] \quad \text{Eq. 5.8}$$

where:

m = A specified month of year

d = A specified day of week

x = A specified short-term count day

c = An appearance value for a specified day *d* and month *m*

a = An appearance value for a specified day *d*

t_{dm} = Total number of days of specified day *d* and month *m*

q = Total quantity of a specified day *d*

k = K-factor value

$ADPT_d$ = Average Daily Pedestrian Traffic for day *d*

$MADPT_m$ = Monthly Average Daily Pedestrian Traffic for month *m*

V_{dmc} = 24 hour pedestrian volume corresponding to appearance *c* on day *d* within month *m*

V_{da} = 24 hour pedestrian volume corresponding to appearance *a* on day *d*

$DWSF_d$ = Day of Week Scaling Factor for day *d*

$MYSF_m$ = Month of Year Scaling Factor for month *m*

STC_x = Short-Term Count for day *x*

$SADPT$ = Seasonal Average Daily Pedestrian Traffic

$EstSADPT_x$ = Estimated Seasonal Average Daily Pedestrian Traffic for day *x*

In Eq. 5.8, the monthly average traffic value is computed using an AASHTO based computation. This is because not every day in each month is present. In that situation, the literature suggests that utilizing an AASHTO based computation is necessary when 100% of the data within a month is not present. On the other hand, in Eq. 5.6, does not separate the average day-of-week traffic value by month, it is just a simple average of either Tuesday, Wednesday, or Thursday daily pedestrian traffic totals occurring between January 1st, 2020 to March31st, 2020. Therefore, the Traditional method requires a total of 6 scaling factors from each site (not including the k-factor) however, 1 scaling factor from each type (MOY and DOW) must be applied for Eq. 5.4 to produce the SADPT estimate.

5.3.2 The AASHTO Method

The next expansion method takes a slightly more disaggregated approach and only requires 1 additional type of scaling factor. The method is known as the AASHTO expansion method and its associated equations (Eq. 5.9, Eq. 5.10, and Eq. 5.11) are listed below.

$$EstSADPT_x = k \times STC_x \times DMSF_{dm} \quad \text{Eq. 5.9}$$

$$DMSF_{dm} = \frac{SADPT}{ADPT_{dm}} \quad \text{Eq. 5.10}$$

$$ADPT_{dm} = \frac{1}{t}\sum_{c=1}^{t} V_{dmc} \quad \text{Eq. 5.11}$$

where:

m = A specified month of year

d = A specified day of week

x = A specified short-term count day

c = An appearance value for a specified day *d* and month *m*

t = Total quantity of a specified day *d* and month *m*

k = K-factor value

$ADPT_{dm}$ = Average Daily Pedestrian Traffic for day *d* within month *m*

V_{dmc} = 24 hour pedestrian volume corresponding to appearance *c* on day *d* within month *m*

$DMSF_{dm}$ = Day of Week Scaling Factor for day *d* within month *m*

STC_x = Short-Term Count for day *x*

$SADPT$ = Seasonal Average Daily Pedestrian Traffic

$EstSADPT_x$ = Estimated Seasonal Average Daily Pedestrian Traffic for day *x*

The AASHTO expansion method itself starts off by determining the average daily traffic for every day-of-week and month-of-year combo within the study period. For this research, there are 9 different average daily traffic values to be computed to correspond with specific short-term count months and days which include: Tuesday, Wednesday, Thursday (3 days) x January, February, March (3 months). After the required average daily totals have been determined, those values are converted to its corresponding AASHTO factor has outlined in Eq. 5.10 which means there are 9 different scaling factors produced from each site.

5.3.3 The Disaggregate Method

The last expansion methodology considered within this research is the most disaggregated of all the expansion methodologies and is known as the Disaggregate method. Eq. 5.12 and Eq. 5.13 outline process for utilizing the Disaggregate method.

$$EstSADPT_x = k \times STC_x \times DYSF_x \quad \text{Eq. 5.12}$$

$$DYSF_x = \frac{SADPT}{V_x} \quad \text{Eq. 5.13}$$

where:

x = A specified short-term count day

k = K-factor value

V_x = 24 hour pedestrian volume on day *x*

$DYSF_x$ = Day of Year Scaling Factor for day *x*

STC_x = Short-Term Count for day *x*

$SADPT$ = Seasonal Average Daily Pedestrian Traffic

$EstSADPT_x$ = Estimated Seasonal Average Daily Pedestrian Traffic for day *x*

From viewing the above equations, there are no average daily values calculated. The Disaggregate method itself, just considers every day for a specific site meeting all short-term count criteria (except for weather) presented in Chapter 4 separately. Then the method divides those specific days by the SADPT value to get all the Disaggregate factor for the given site. With using the Disaggregate method, it is worth knowing that all considered sites do not likely have the same quantity of Disaggregate factors. The reason why that happens is because the filtering results are different for every individual site.

Chapter 6 Factor Grouping

6.1 Introduction

In the AADT/SADPT expansion calculation, there are two main components that make up the input which ultimately includes the short-term count of interest and at least one scaling/expansion factor as shown in Chapter 1. The short-term count is collected from a count site with limited quantity of count data and almost no temporal profile (short-term count site). Scaling factors are computed based on data from a collection of continuous count sites (factor group) and since those sites are continuous, these count data contain temporal information. This chapter focuses on two technical challenges: the first is the method for establishing factor groups with continuous count sites (research problem **3a**); and the second is the method to associate a short-term count site with the appropriate factor group (research problem **3b**).

For problem 3b, it is not possible to use AMI or WWI or other temporal indices because the short-term counts are too short to contain sufficient temporal information. Consequently, it is proposed to explore models that make use of land use information.

Each of these sub-problems is addressed in the following sections.

6.2 Grouping Continuous Count Sites

The literature review revealed the use of temporal indices (i.e. AMI, WWI, and AMI+WWI) for factor grouping. Those specific indices were primarily developed from studies involving cyclists and for those studies involving pedestrians, the indices were calculated from pedestrian data collected on off-road trails and not at roadway intersections. Therefore, it is not clear if AMI and/or WWI effectively capture the behavior of pedestrians utilizing any given roadway intersection. Nevertheless, these three temporal indices are used as benchmark factor grouping methods in this thesis.

Another possible approach to finding the true factor groups is to utilize the k-means clustering methodology. However, the problem with k-mean clustering is that it is not clear how many clusters should be considered or what and how many input variables should be included in the analysis.

The application of the benchmark methods and the k-mean clustering methods are described in the following sub-sections.

6.2.1 Benchmark Factor Grouping Methods

As mentioned in Chapter 2, the usage of AMI and WWI in the factor grouping process have been selected as benchmark methods. Overall, AMI expresses a relationship between the traffic in the morning and midday peak periods of any given day. AMI also summarizes traffic patterns for a specific site on an hourly level. On the other hand, WWI highlights the relationship between weekend and weekday traffic. Therefore, WWI categorizes traffic patterns using daily count totals. Within this research there are 3 distinct benchmark methods which include (1) AMI only (Johnstone, Nordback, & Kothuri, 2018), (2) WWI only (Nordback, et al., 2019), and (3) AMI and WWI combined (Hankey, Lindsey, & Marshall, 2014).

However, in Hankey et al. (2014) a mixed factor group is created to deal with the unclarity surrounding the factor grouping process. Although, there is a potential solution presented in Miranda-Moreno et al. (2013) that could possibly provide more clarity in determining if the site is more recreational or

utilitarian based. However, the intention of this research is to not adjust the threshold values listed in Hankey et al. (2014) and if the solution outlined in Miranda-Moreno et al. (2013) is used, it is possible the threshold values for AMI and WWI in Hankey et al. (2014) could change. Another issue with using the solution in Miranda-Moreno et al. (2013) is that it might be difficult to tell what temporal trends the low traffic volume sites in this work are revealing as discussed before. Therefore, adding the mixed group appeared to be the most appropriate solution. Table 6.1 and Table 6.2 shown below highlight the requirements for each group in the AMI and WWI combined case. 19 of 25 sites somewhat exhibit characteristics of both utilitarian and recreational traffic patterns. Those results are not unexcepted based on the fact that hourly and daily trends are not always connected for utilitarian and recreational behavior for the AMI only and WWI only cases.

Table 6.1. Boundaries for Combined AMI and WWI Factor Groups Modified from (Hankey, Lindsey, & Marshall, 2014)

Factor Group	WWI	AMI
Utilitarian	<0.8	>1.5
Mixed Utilitarian	>=0.8 & <1	>1 & <=1.5
Mixed Recreational	>1.25 & <=1.8	>=0.35 & <0.75
Recreational	>1.8	<0.35
Mixed	>=1 & <=1.25	>=0.75 & <=1

Table 6.2. Conditions for the Modified Combined AMI and WWI Factor Groups

Factor Group	Criteria
Utilitarian	1) AMI and WWI = Utilitarian 2) AMI/WWI = Utilitarian and AMI/WWI = Mixed Utilitarian
Mixed Utilitarian	1) AMI and WWI = Mixed Utilitarian 2) AMI/WWI = Mixed Utilitarian and AMI/WWI = Mixed 3) AMI/WWI = Utilitarian and AMI/WWI = Mixed
Mixed Recreational	1) AMI and WWI = Mixed Recreational 2) AMI/WWI = Mixed Recreational and AMI/WWI = Mixed 3) AMI/WWI = Recreational and AMI/WWI = Mixed
Recreational	1) AMI and WWI = Recreational 2) AMI/WWI = Recreational and AMI/WWI = Mixed Recreational
Mixed	Any criteria not previously listed

6.2.2 K-Means Factor Grouping Methods

The k-means clustering method provides a factor grouping method that is both objective and flexible, but it is necessary to define the feature vector and the number of clusters. Given the objective is to distinguish between sites that have different expansion factor values, it is proposed to compose the feature vector of the k-factor, and the DOW and MOY scaling factors. Only the DOW scaling factors

associated with short-term count days (i.e. Tuesdays, Wednesday, and Thursdays) need be considered so for a typical application, the feature vector consists of 1+3+12 = 16 values. For the Pima County application, for which we only have valid data from three months (Jan, Feb, and March), the feature vector consists of just 1+3+3 = 7 values.

Having decided which variables to utilize in the feature vector, it is necessary to select the number of clusters and this requires a means of comparing the two different factor grouping solutions that have different numbers of groups. For this research, two different k-means cases were considered where one case has two clusters and the other case has three clusters. Using a three cluster setup makes sense because from the benchmark cases, there generally seems to be 3 types of factor groups. With using a three cluster setup, there is a possibility that the clustering results could match the results of one of the benchmark methods. If that is the case, it proves that the use of the specified benchmark method in future studies is suitable because of its simplicity and it matches the results of a more complex factor grouping method. Using a two cluster setup is also a good option because if certain sites seem to fit within the "mixed factor group", it is possible to see if those sites are more associated with the utilitarian or recreational factor group.

6.2.3 Comparing Factor Grouping Methods

The results from the three benchmark and proposed factor grouping methods are compared qualitatively and quantitively, as described in the next sections.

6.2.3.1 Qualitative comparison of factor grouping method results

The grouping results for the benchmark methods and the proposed k-means clustering methods are provided in Table 6.3. The results are colour coded to clearly show the groupings. Sites that are assumed to be more utilitarian based are yellow, recreational based are blue, and sites with both patterns are green. The only exception to that rule are sites within the mixed utilitarian group which uses red. For a given grouping method, sites within the same group are indicated in the same colour. Comparisons across grouping methods can also be made on the basis of the colour coding even when the factoring methods may assign different test descriptors to the groups. For example, consider Site S001 in Table 6.3. This site is assigned to Cluster 0 by the K-means (n=2) grouping method and to Cluster 1 by the K-means (n=3) method. These cluster labels are arbitrary and the light blue colours in the cells indicates that both these methods identify this site into the same group. The same can be observed from the group assignment from the three benchmark methods. The group labels vary by method, but the site is effectively assigned to the same group.

A review of the results in Table 6.3 indicates that there is little consistency in the site groupings by the different grouping methods even when the grouping methods have the same number of groups and provide similar number of sites in each group. Consider the results from the AMI Only and WWI Only methods. They both create 3 factor groups and the number of sites assigned to each factor group are almost identical (the AMI Only case has 11, 10, and 4 sites in each of its factor groups and the WWI Only case has 10, 10, and 5 sites in each of its factor groups). However, the allocation of the sites to each group is not very consistent between the two methods. As previously implied, it is not unreasonable for someone to assume that hourly and daily trends could be associated with each other but, the hourly commute group from the AMI Only case has 4 sites and the weekday commute group from the WWI Only case has 10 sites associated with it. That result implies that sites having a heavier portion of traffic occurring on weekdays when compared to Saturday and Sunday do not always display a traditional

hourly peak pattern. It also does not appear as if a midday peak pattern is always associated with heaver weekend traffic. This is because for the AMI only case, the hourly noon activity group has 10 sites and the weekend multipurpose group only has 5 sites fitting within its boundaries.

Table 6.3. Factor Group Classification Table by Factor Grouping Method

Site Code	K-means (n=2)	K-means (n=3)	AMI and WWI	AMI Only	WWI Only
S001	Cluster 0	Cluster 1	Mixed Recreational	Hourly Noon Activity	Weekend Multipurpose
S004	Cluster 0	Cluster 1	Mixed Utilitarian	Hourly Multipurpose	Weekday Commute
S018	Cluster 1	Cluster 2	Mixed	Hourly Multipurpose	Weekly Multipurpose
S021	Cluster 1	Cluster 2	Utilitarian	Hourly Commute	Weekday Commute
S023	Cluster 1	Cluster 2	Mixed	Hourly Noon Activity	Weekday Commute
S024	Cluster 0	Cluster 2	Utilitarian	Hourly Commute	Weekday Commute
S031	Cluster 1	Cluster 0	Mixed	Hourly Multipurpose	Weekend Multipurpose
S041	Cluster 1	Cluster 0	Mixed Recreational	Hourly Noon Activity	Weekly Multipurpose
S045	Cluster 1	Cluster 0	Mixed Recreational	Hourly Noon Activity	Weekend Multipurpose
S050	Cluster 0	Cluster 1	Mixed Utilitarian	Hourly Multipurpose	Weekly Multipurpose
S055	Cluster 0	Cluster 1	Utilitarian	Hourly Commute	Weekly Multipurpose
S056	Cluster 1	Cluster 0	Mixed Utilitarian	Hourly Multipurpose	Weekly Multipurpose
S062	Cluster 1	Cluster 2	Mixed Utilitarian	Hourly Multipurpose	Weekday Commute
S065	Cluster 0	Cluster 1	Mixed	Hourly Multipurpose	Weekly Multipurpose
S072	Cluster 1	Cluster 0	Mixed Recreational	Hourly Multipurpose	Weekly Multipurpose
S078	Cluster 1	Cluster 2	Mixed	Hourly Noon Activity	Weekday Commute
S079	Cluster 0	Cluster 1	Mixed	Hourly Noon Activity	Weekly Multipurpose
S082	Cluster 1	Cluster 2	Mixed	Hourly Noon Activity	Weekly Multipurpose
S083	Cluster 1	Cluster 2	Utilitarian	Hourly Multipurpose	Weekday Commute
S084	Cluster 1	Cluster 0	Mixed Recreational	Hourly Noon Activity	Weekly Multipurpose
S090	Cluster 1	Cluster 2	Utilitarian	Hourly Multipurpose	Weekday Commute
S093	Cluster 1	Cluster 0	Mixed Recreational	Hourly Noon Activity	Weekend Multipurpose
S098	Cluster 1	Cluster 2	Mixed	Hourly Noon Activity	Weekday Commute
S101	Cluster 1	Cluster 0	Mixed Recreational	Hourly Multipurpose	Weekend Multipurpose
S107	Cluster 1	Cluster 2	Utilitarian	Hourly Commute	Weekday Commute

In the combined AMI and WWI case, there are more than 3 factor groups considered. The utilitarian, mixed utilitarian, mixed recreational, recreational, and mixed groups each have 6, 4, 7, 0, and 8 sites respectively. For the K-means case with two clusters, the quantity of sites in each of the two clusters is quite different, cluster 0 only has 7 sites and cluster 1 has the remaining 18 sites. Most of the sites in cluster 0 appear to be utilitarian or a mix of recreational and utilitarian if the sites are to be placed in

one of the literature cases but, 1 of the sites in cluster 0 does display some strict recreational behavior. For the K-means case with three clusters, clusters 0, 1, and 2 each have 8, 6, and 11 sites respectively. The sites within cluster 1 for the three cluster setup fit within cluster 0 of the two cluster setup and there is only a difference in size by one site between those two clusters. It might also be justified to assume that the literature cases for factor grouping are not the best for creating factor groups for pedestrians. This is because if AMI and/or WWI are useful thresholds in the factor grouping process, they should produce factor groups that are almost identical to the clustering method results and that does not appear to happen.

6.2.3.2 Quantitative comparison of factor grouping method results

To complete the evaluation of factor grouping methods, the impact of the factor grouping method on the AADPT estimation accuracy was determined. Accuracy was quantified in terms of the mean absolute percent error (MAPE) and mean absolute error (MAE) of the AADPT estimate assuming the true factor group to which the short term count site belongs is known. As an example, group/cluster 0 in the k-means case with two clusters has 7 sites and if someone wanted to expand a short-term count corresponding to a specific day and the 4th site listed for cluster 0 (Table 6.3), each specific scaling factor from all expansion methods (Traditional, AASHTO, and Disaggregate) are calculated as an average of the true scaling factors from the other sites in the group (sites 1-3, and 5-7) for given type of scaling factor. Once all the scaling factors are calculated for each the expansion methods, the applicable scaling factors associated with the short-term count day for site 4 are then used as multiplying factors for the short-term count value which produces the SADPT estimate for the site and day. After obtaining all SADPT from every eligible short-term count day from a given site, MAPE and MAE values are obtained and using simple average computation of the site error value produces error values for a given factor grouping case.

In Table 6.4, MAPE and MAE values corresponding to the factor grouping methods are shown. The best factor grouping method is the K-means case with three clusters and the worst performing case is the AMI Only case. T-tests were conducted to determine if the differences in AADPT estimation accuracy between the different factor grouping methods were statistically significant. These results for MAPE and MAE are shown in Appendix E, and most of the case by case comparisons show that the differences in MAPE and MAE values are significant. In terms of the expansion methods, the Traditional method had the lowest MAPE values and the Disaggregate method had the highest MAPE values. Another key observation is that the Disaggregate method did not outperform the Traditional method and as stated before, the Disaggregate method has always been the best performing in past studies. One of the main concerns with the MAPE values for the factor grouping cases is that they range between 45% to 70%, which is much higher than those reported in the literature. The range of MAE values for the factor grouping cases is between 30 to 60, and sites with lower pedestrian volumes could have an influence on the larger MAPE values and smaller MAE values.

Table 6.4. AADPT Estimation Error Metrics by Factor Grouping Method

Metric	Expansion Method	Factor Grouping Method				
		K-means (n=2)	K-means (n=3)	AMI and WWI	AMI Only	WWI Only
MAPE (%)	Traditional	51.0	47.5	52.2	55.5	48.3
	Disaggregate	63.8	55.0	63.0	67.8	55.8
	AASHTO	53.0	48.8	54.5	58.0	49.4
MAE	Traditional	39.4	33.0	39.1	44.6	35.8
	Disaggregate	51.6	34.8	45.9	55.6	39.5
	AASHTO	41.6	33.8	40.9	47.2	36.9

Estimation errors were also examined by individual site. MAPE and MAE by site are presented in Appendix F. Note that are 5 factor grouping methods and 3 different expansion method for each site in this analysis, which means that each site has 15 site MAPE and 15 MAE values corresponding to it.

Figure 6.1 shows the MAPE errors for each site considering all expansion methods and the k-means clustering factor grouping method with two clusters. There seems to be collection of sites with respectable MAPE values when compared with the literature. From Appendix F, the site MAE values have a relatively large range and it is likely large because of the collection of low and high pedestrian volume sites in the dataset. However, 3 specific sites, which include Calle del Marques and Sunrise Dr (Site S021), Linda Vista Bl and Thornydale Rd (Site S083), and Sunrise Dr and Swan Rd (Site S107) consistently had MAPE values above 100% for most scenarios considered.

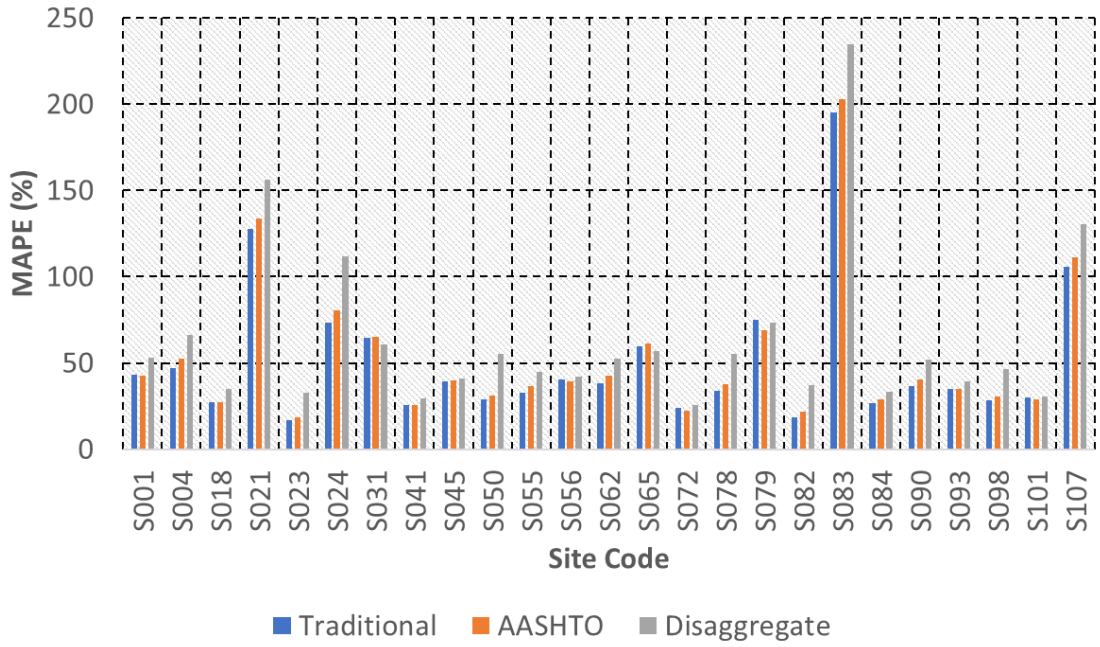

Figure 6.1. MAPE Values for the K-means (n=2) Factor Grouping Method

6.2.4 Modified K-Means Factor Grouping Methods

An investigation was carried out on these three sites and it was observed that they are all located very close to high schools. To explore if being located close to a high school plays a role in increasing MAPE values, the relationship between AADPT estimation accuracy (MAPE) and the closest walking distance to a nearby high school (PIMA COUNTY SCHOOL SUPERINTENDENT, 2021) was examined (Figure 6.2). The results in the figure show (1) that for sites located less than 1km walking distance from a high school, estimation errors are much higher than for sites located farther away from a high school; and (2) that there is no discernable impact of distance to the nearest high school when the distance is equal to or greater than 1km. Therefore, none of the current factor grouping cases were able to capture the connection between close proximity to high schools and site MAPE.

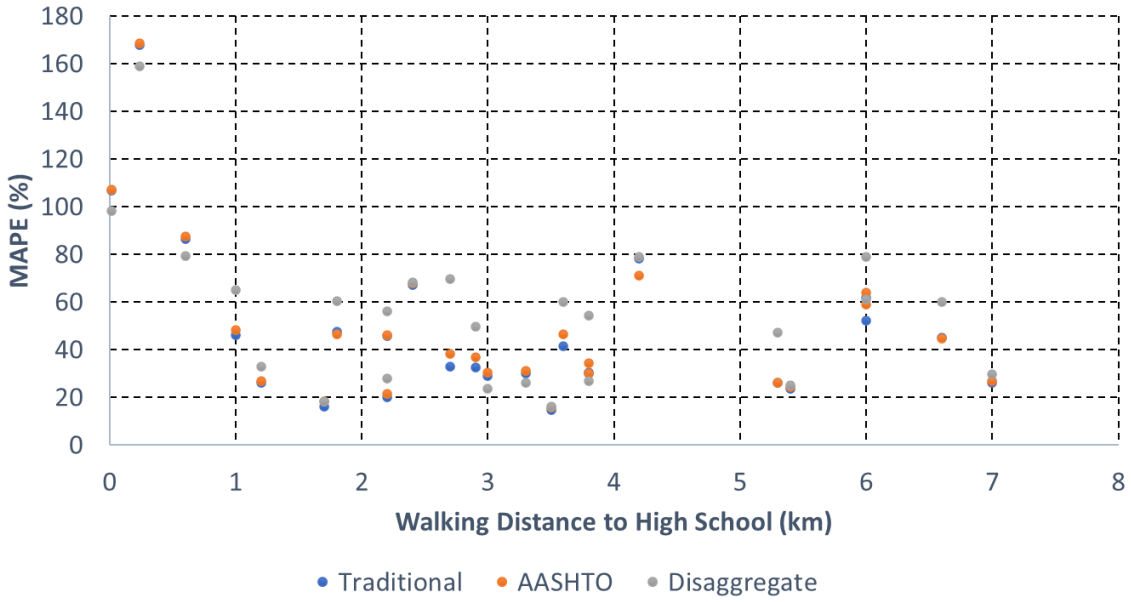

Figure 6.2. Walking Distance to High School vs MAPE for the K-means (n=3) Factor Grouping Method

With the observation made in the above paragraph, both of the k-means clustering factor grouping methods were modified to place the three sites close to high schools (i.e. intersections of Calle del Marques and Sunrise Dr, Linda Vista Bl and Thornydale Rd, and Sunrise Dr and Swan Rd) in their own exclusive high school factor group. The modified K-means clustering method consists of applying the standard K-means clustering to all 25 sites and then extracting the sites located less than 1km from a high school from the clusters they have assigned to and placing them in their own cluster. In this way, the modified K-means clustering method with n clusters results in n+1 groups.

The results (MAPE and MAE) results for all the factor grouping methods, including the modified K-means clustering methods, are provided in Table 6.5. From these results we can observe:

1. The modified k-means clustering methods provide better AADPT estimation accuracy than the standard K-means clustering methods and the three benchmark methods.
2. The modified K-means clustering method performs better with n=3 than with n=2.
3. The Traditional expansion method performs the best across the different factor grouping methods and the Disaggregate expansion method performs the worst.

There are 21 t-test results presented in Appendix G for MAPE and MAE. 17 of the case by case comparisons have significant differences and for only 4 comparisons, the differences in AADPT are not statistically significant.

Table 6.6 provides the site groupings that result from the application of the factor grouping methods. Please note that the colour coding does not change from Table 6.3 however, sites in the high school group use the color pink. It is observed that group/cluster 0 in the two cluster setup remains the same between the previous and modified versions. The same is also true for clusters 0 and 1 within the three cluster setup. Even though the clustering cases now have improved MAPE values, all the MAPE values are still above 40% which again is not ideal. Therefore, it is likely necessary to evaluate the impact of the high school group on the site MAPE values.

Table 6.5. Updated Error Metrics by Factor Grouping Method

Metric	**Expansion Method**	**Factor Grouping Method**						
		K-means (n=2) Mod	K-means (n=3) Mod	K-means (n=2)	K-means (n=3)	AMI and WWI	AMI Only	WWI Only
MAPE	Traditional	44.2	42.6	51.0	47.5	52.2	55.5	48.3
	Disaggregate	52.4	48.0	63.8	55.0	63.0	67.8	55.8
	AASHTO	46.1	43.9	53.0	48.8	54.5	58.0	49.4
MAE	Traditional	27.9	24.9	39.4	33.0	39.1	44.6	35.8
	Disaggregate	31.2	23.6	51.6	34.8	45.9	55.6	39.5
	AASHTO	29.4	25.8	41.6	33.8	40.9	47.2	36.9

Even though the modified k-means clustering cases are the best performing it is important to realize that there are practical issues with implementing that methodology. It is not clear on how many clusters should be considered and what temporal metrics should be inputted into the analysis as well. With the benchmark methods, all the thresholds for allocating sites have already been set and it is just a matter of knowing where a given site's AMI and/or WWI values fall between.

Table 6.6. Updated Factor Group Classification Table by Factor Grouping Method

Site Code	K-means (n=2) Mod	K-means (n=3) Mod	K-means (n=2)	K-means (n=3)	AMI and WWI	AMI Only	WWI Only
S001	C0	C1	C0	C1	MR	HNA	WDM
S004	C0	C1	C0	C1	MU	HM	WC
S018	C1	C2	C1	C2	MX	HM	WYM
S021	HSG	HSG	C1	C2	UT	HC	WC
S023	C1	C2	C1	C2	MX	HNA	WC
S024	C0	C2	C0	C2	UT	HC	WC
S031	C1	C0	C1	C0	MX	HM	WDM
S041	C1	C0	C1	C0	MR	HNA	WYM
S045	C1	C0	C1	C0	MR	HNA	WDM
S050	C0	C1	C0	C1	MU	HM	WYM
S055	C0	C1	C0	C1	UT	HC	WYM
S056	C1	C0	C1	C0	MU	HM	WYM
S062	C1	C2	C1	C2	MU	HM	WC
S065	C0	C1	C0	C1	MX	HM	WYM
S072	C1	C0	C1	C0	MR	HM	WYM
S078	C1	C2	C1	C2	MX	HNA	WC
S079	C0	C1	C0	C1	MX	HNA	WYM
S082	C1	C2	C1	C2	MX	HNA	WYM
S083	HSG	HSG	C1	C2	UT	HM	WC
S084	C1	C0	C1	C0	MR	HNA	WYM
S090	C1	C2	C1	C2	UT	HM	WC
S093	C1	C0	C1	C0	MR	HNA	WDM
S098	C1	C2	C1	C2	MX	HNA	WC
S101	C1	C0	C1	C0	MR	HM	WDM
S107	HSG	HSG	C1	C2	UT	HC	WC

C0 = Cluster 0, C1 = Cluster 1, C2 = Cluster 2, HSG = High School Group,
UT= Utilitarian, MU = Mixed Utilitarian, MX = Mixed, MR = Mixed Recreational,
HC = Hourly Commute, HM = Hourly Multipurpose, HNA = Hourly Noon Activity,
WC = Weekday Commute, WYM = Weekly Multipurpose, WDM = Weekend Multipurpose

Even though it appears that the addition of the high school group has generally improved MAPE values on the factor group level, it is still critical to determine why the high school group still has large errors. After viewing the site MAPE values for all 3 sites in the high school factor group (Appendix H), 2 of the 3 sites have MAPE values that are substantially less than 100% and for some of the MAPE values, they are more in line with the literature results. For the site of Linda Vista and Thornydale (Site S083), the errors have improved but they still hover around 100%. One of the main differences between the site of Linda Vista and Thornydale and the other two sites is that a shopping plaza is located near the remaining two sites. Therefore, the pedestrian demands at these other two sites are influenced by both the nearby high school and the retail plaza and consequently these sites are expected to have a larger portion of

weekend traffic. The WWI values support this observations. Linda Vista and Thornydale has a WWI value of 0.17 whereas, the other two sites have WWI values close to 0.35. Overall, the WWI values of all three high school sites indicate that the sites likely display utilitarian behavior because of their proximity to a high school. However, the difference in WWI values does indicate that Linda Vista and Thorndale's temporal trend is unique within the high school group because weekend traffic portion is much smaller than the weekend portion for the other two sites. For this analysis, the Linda Vista and Thornydale site is not reassigned to another factor group. With the factor groups not changing, site MAPE and MAE values still generally have a wide spectrum of values.

6.3 Estimating AADPT from Short-Term Counts

The previous section quantified AADPT estimation errors when expansion factors are computed from the factor groups and when the correct factor group is known for each short-term count site. However, in practice, the factor group is not known in advance and there must be a way to determine the appropriate factor group. This section describes two approaches taken to determine expansion factors for short-term count sites:

1. Direct estimation of expansion factors from the short-term count site characteristics.
2. Identification of the most appropriate factor group based on short-term counts site characteristics, and then determination of the expansion factors from the factor group.

Each of these approaches is described in the following sections.

6.3.1 Direct Estimation of Temporal Trends using Regression Modelling

It is proposed to estimate the temporal trend factor(s) on the basis of the land use characteristics at the short term site. For the modified k-means clustering methods (n=2 and n=3) the feature vector consists of 7 temporal scaling factors and thus linear regression models are required to estimate each of these seven values. The three benchmark factor grouping methods require the estimation of AMI and/or WWI.

Land use characteristics within 400m of each site are captured via five binary land use variables (1=land use present; 0 otherwise) as defined in Table 6.7. The variables are relatively broad categories except for the convenience store flag. The land use characteristic data were obtained and complied from Google Places™ by another student (A. Jafari). The school flag considers post-secondary schools, high schools, and elementary schools. The retail flag reflects any retail location such as restaurants, big box, grocery, and clothing stores. For the parks and recreation variable, other recreational facilities could include gyms and sporting arenas. Finally, for the residential flag, that variable could include anything from a subdivision of single unit dwellings or a high rise apartment building complex.

Table 6.7. List of Land Use Variables Considered for Regression Modelling

Land Use Variable	**Description**
ConvSFlag	Convenience Stores
Park_Rec	Parks and other Recreational Areas/Facilities
Residential	Residential Areas
RetailFlag	Retail Establishments
SchoolFlag	Educational Institutions

9 different land use regression models were developed where each model corresponded to one of the dependent variables. All the land use variables listed in Table 6.7 were used in a step wise regression in R with 95% confidence intervals to produce the results for each of the models. Detailed regression model fitting results are provided in Appendices I through M. A summary is provided in Table 6.8. All reported model coefficients and constants are statistically significant at the 95% confidence level.

The results indicate that linear models, relating land use characteristics to temporal factors, could be established for only the AMI, WWI, January, March, and Tuesday scaling factors. Therefore, the K-factor, February, Wednesday, and Thursday scaling factors did not produce any significant land use regression models. The predicted values for those dependent variables are equal to the average across all 25 sites for each variable respectively. For the AMI land use model, park/recreational and school land uses are significant and the AMI land use model is the only model with two independent variables. The VIF value indicate that there is no substantial correlation between these two independent variables. For the WWI and Tuesday land use models, only the school land use was significant and for the January and March land use models, the convenience store land use was the only significant independent variable.

In Table 6.8, the prediction error metrics for each of the dependent variables is also displayed. These values were computed by comparing the value of the dependent variable determined from the proposed regression model to the "true" value computed from the site count data. The majority of the MAPE values are below 15% which is quite good. However, for WWI the MAPE is around 35% and for AMI, the MAPE is even larger and hovers above 60%. The error metrics obtained are not overly surprising because the range of values for AMI and WWI appears to be larger than the range of values for the other dependent variables. One of the concerns with the AMI land use model is that it predicts a negative AMI value for a couple of the sites. A negative AMI value would never be realistic because a traffic count can never be less than zero and the input for AMI exclusively considers traffic counts. Another issue is that the adjusted r-squared values for every land use model is less than 0.4. That finding implies that the given land uses are not overly reliable at predicting the selected temporal metrics.

Table 6.8. Summary of Land Use Regression Model Calibration

Model (Dependent variable)	**Independent Variable (coefficient)**	**constant**	**VIF Value**	**Adjusted R^2**	**MAPE (%)**
AMI	SchoolFlag (1.581) Park_Rec (-1.014)	0.954	1.133	0.395	61.8
WWI	SchoolFlag (-0.443)	1.027	NA	0.275	34.5
K-factor	NA	1.903*	NA	NA	13.6
Jan	ConvSFlag (-0.128)	1.019	NA	0.189	7.2
Feb	NA	1.035*	NA	NA	9.6
Mar	ConvSFlag (0.237)	0.982	NA	0.220	11.1
Tue	SchoolFlag (0.200)	1.031	NA	0.275	10.0
Wed	NA	0.995*	NA	NA	12.4
Thu	NA	1.005*	NA	NA	10.4

*One average value for all sites

The values predicted for each site from the regression equations or average values are used as the scaling factor components in the Traditional method (Eq. 5.4) for the purposes of obtaining SADPT estimates for its corresponding site. That type of SADPT estimation is known as the direct estimation method and since Eq. 5.4 is utilized, only predicted values for the k-factor and all 3 MOY and DOW factors are required. The purpose of the direct estimation method is to see if it is worthwhile to attempt placing short-term count sites into a specific factor group. To make that determination, the SADPT errors from the direct estimation method and all the relevant factor grouping methods are compared in Section 6.3.3. If the SADPT errors from the direct estimation method are similar or significantly less than the factor grouping SADPT errors, that is likely an indication that factor grouping isn't necessary for short-term counts. It is also worth noting that there was no attempt in this research to develop regression equations for the AASHTO and Disaggregate scaling factors.

6.3.2 Identification of Factor Group using Regression Modelling

In this method, the factor group to which the short-term count site should be associated is identified by comparing one or more temporal trend measures estimated based on the short-term count site characteristics to the same temporal trend measured used to establish the factor groups of the continuous count site. The regression models described in the previous sections are used to estimate the temporal trend measures for the short-term count sites. For the benchmark factor grouping methods, it is required to compute the estimated AMI/WWI value and then apply the factor grouping criteria accordingly. For the k-means clustering methods, the actual clustering analysis is not carried out again however, all 25 sites are treated as short-term count sites and use the regression models to predict each value in a feature vector of the 7 factors. The short-term count feature vector is then compared with the average feature vector computed for each factor group within both of the clustering cases in Section 6.2.4. When the short-term count feature vectors are being compared with an average factor group feature vector, the mean absolute error (MAE) considering all 7 scaling factors shared between the vectors was computed. The factor group that corresponds to the lowest MAE value is the factor group the given site belongs to. Since 4 of the 7 elements consisting of the short-term count feature vector are based on average values, it was also decided to match short-term count sites with a k-means cluster using a feature vector consisting of 3 components that have a regression equation associated with it. Those specific values include the January, March, and Tuesday scaling factors and again the cluster corresponding to the smallest MAE value based off all 3 components is the known as the appropriate factor group for the respective short-term count site.

It is also important to remember that the input for the factor group scaling factors doesn't change from the original factor groups outlined in Section 6.2. That statement remains true even if specific factor groups loses or gains sites in the land use factor grouping prediction process. The only exception is if a given short-term count location matches the location of a continuous count site in specific factor group from Section 6.2. If that is the case, the input from that site is removed from the factor group scaling factor computations.

The models for determining the factor group for a short-term count sites are assessed in two ways. The first is to qualitatively examine the factor groupings that result from the group prediction models. The second is to quantify the accuracy of the factor group identification methods in terms of factor group prediction accuracy. Each of these is discussed in the following sections.

6.3.2.1 Qualitative assessment of predicted factor groups:

The factor group membership prediction for each of the factor grouping methods is shown in *Table 6.10* (color coding remains the same as Table 6.6) and the factor groups that have known associated land uses are presented in *Table 6.9*. It can be observed that:

1. The hourly commute or utilitarian group are both the same as the weekday commute group.
2. Cluster 1 in the K-means clustering (n=2) method is the same as cluster 2 in the K-means (n=3) method.
3. There is a total of 9 unique factor grouping arrangements (not including the high school group) created from the land use factor group prediction models.
4. The Hourly Commute, Utilitarian or Weekday Commute group must have sites that are within 400 meters of any type of school.
5. The hourly multipurpose or mixed group must have sites that are not located within 400 meters of both a school and recreational area. For the hourly noon activity or mixed recreational group, sites within that group must be located within 400 meters of a recreational area but further than 400 meters from any type of school. The weekly multipurpose group is simply a combination of the hourly multipurpose/mixed and hourly noon activity/mixed recreation groups where the sites cannot be less than 400 meters situated from any type of school. For any of the factor groups formed from the k-means clustering analysis, those sites must not be within 1 kilometer of a high school but other land uses for the k-means factor groups are unknown.

Table 6.9. Land Use Conditions for Factor Grouping Placements

Factor Group	**Conditions**					
	< 400m from a School Area	=> 400m from a School Area	< 400m from a Recreational Area	=> 400m from a Recreational Area	< 1km from a High School	=> 1km from a High School
Hourly Commute, Utilitarian, or Weekday Commute	Required	NA	NA	NA	NA	NA
Hourly Multipurpose or Mixed	NA	Required	NA	Required	NA	NA
Hourly Noon Activity or Mixed Recreational	NA	Required	Required	NA	NA	NA
Weekly Multipurpose	NA	Required	NA	NA	NA	NA
High School Group	NA	NA	NA	NA	Required	NA

With knowing the associated land uses across the different factor groups, STC sites can be clearly placed into a group. The associated land uses can also explain why the Combined AMI and WWI and the AMI Only cases are the same. Both of those factor grouping methods consider the same two land use variables that are utilized to estimate AMI values. As stated before, WWI only considers one land use variable in its regression model and that land use variable is one of the two land uses considered in the AMI regression model. Ultimately, the AMI model is more dominant than the WWI model because the additional land use variable in the AMI model gives it the ability to create more factor groups. Therefore, the site assignment to the given factor groups is identical. The AMI Only and the combined AMI and WWI cases also have a site distribution of 16, 6, and 3 sites and the WWI case has one factor group with 19 sites and the other factor group in that case has only 6 sites. The distribution of the 7V clustering cases includes the high school group of 3 sites and the larger cluster of 22 sites. The 3 variable (3V) clustering cases also consider the high school group and have a distribution 19, 3, and 3 sites for 2 clusters and 18, 4, and 3 sites for 3 clusters.

Table 6.10. Land Use and Original Factor Group Classification Comparison by Factor Grouping Method

Site Code	K-means (n=2) Mod (7V)		K-means (n=2) Mod (3V)		K-means (n=3) Mod (7V)		K-means (n=3) Mod (3V)		AMI and WWI		AMI Only		WWI Only	
	Org	LU	Org	LU	Org	LU	Org	LU	Org	LU	Org	LU	Org	LU
S001	C0	C1	C0	C1	C1	C2	C1	C1	MR	MX	HNA	HM	WDM	WYM
S004	C0	C1	C0	C1	C1	C2	C1	C1	MU	MR	HM	HNA	WC	WYM
S018	C1	C1	C1	C1	C2	C2	C2	C2	MX	MX	HM	HM	WYM	WYM
S021	HSG	HSG	HSG	HSG	HSG	HSG	HSG	HSG	UT	UT	HC	HC	WC	WC
S023	C1	C1	C1	C1	C2	C2	C2	C1	MX	MX	HNA	HM	WC	WYM
S024	C0	C1	C0	C0	C2	C2	C2	C2	UT	UT	HC	HC	WC	WC
S031	C1	C1	C1	C1	C0	C2	C0	C1	MX	MX	HM	HM	WDM	WYM
S041	C1	C1	C1	C1	C0	C2	C0	C1	MR	MX	HNA	HM	WYM	WYM
S045	C1	C1	C1	C1	C0	C2	C0	C1	MR	MX	HNA	HM	WDM	WYM
S050	C0	C1	C0	C1	C1	C2	C1	C1	MU	MX	HM	HM	WYM	WYM
S055	C0	C1	C0	C1	C1	C2	C1	C1	UT	MX	HC	HM	WYM	WYM
S056	C1	C1	C1	C1	C0	C2	C0	C1	MU	MX	HM	HM	WYM	WYM
S062	C1	C1	C1	C1	C2	C2	C2	C1	MU	MX	HM	HM	WC	WYM
S065	C0	C1	C0	C1	C1	C2	C1	C1	MX	MX	HM	HM	WYM	WYM
S072	C1	C1	C1	C1	C0	C2	C0	C1	MR	MX	HM	HM	WYM	WYM
S078	C1	C1	C1	C1	C2	C2	C2	C1	MX	MX	HNA	HM	WC	WYM
S079	C0	C1	C0	C1	C1	C2	C1	C1	MX	MX	HNA	HM	WYM	WYM
S082	C1	C1	C1	C1	C2	C2	C2	C2	MX	MX	HNA	HM	WYM	WYM
S083	HSG	HSG	HSG	HSG	HSG	HSG	HSG	HSG	UT	UT	HM	HC	WC	WC
S084	C1	C1	C1	C1	C0	C2	C0	C1	MR	MR	HNA	HNA	WYM	WYM
S090	C1	C1	C1	C1	C2	C2	C2	C1	UT	MX	HM	HM	WC	WYM
S093	C1	C1	C1	C1	C0	C2	C0	C1	MR	MR	HNA	HNA	WDM	WYM
S098	C1	C1	C1	C0	C2	C2	C2	C2	MX	UT	HNA	HC	WC	WC
S101	C1	C1	C1	C0	C0	C2	C0	C1	MR	UT	HM	HC	WDM	WC
S107	HSG	HSG	HSG	HSG	HSG	HSG	HSG	HSG	UT	UT	HC	HC	WC	WC

C0 = Cluster 0, C1 = Cluster 1, C2 = Cluster 2, HSG = High School Group,
UT= Utilitarian, MU = Mixed Utilitarian, MX = Mixed, MR = Mixed Recreational,
HC = Hourly Commute, HM = Hourly Multipurpose, HNA = Hourly Noon Activity,
WC = Weekday Commute, WYM = Weekly Multipurpose, WDM = Weekend Multipurpose,
Org = Original Factor Group, LU = Predicted Land Use Factor Group

The likely reason why the n=2 and n=3 clustering cases considering 7 variables (7V) were the exact same is because most of the components making up the clustering analysis feature vector did not have a regression model. Since that was the case, there likely was not enough variation between the site feature vectors because all the feature vectors have 4 components equal to one average value for each component. In the 3V setup with n=3, 18 of the 22 sites originally assigned to cluster 2 in the 7V setup with n=3 were moved to cluster 1. That specific observation proves that considering a different quantity of variables between k-means cases can potentially alter the factor grouping placement of short-term count sites.

6.3.2.2 Factor group prediction accuracy:

Factor group prediction accuracy is determined as the number of sites for which the predicted group is correct divided by the total number of group membership predictions (in this application this is 25). Prediction accuracy, provided in Table 6.11, ranges from 72% to 44%. The case that best assigned the sites to their original factor group in Section 6.2 is the modified k-means clustering case with n=2 and 7 variables. The worst case is the modified k-means cases with n=3 and 7 variables. The reason why that likely happened is because the best and worst cases have the same predicted factor grouping arrangement. However, the original n=2 and n=3 cases are different and since the original n=3 case considered one more factor group that the original n=2 case, the prediction score for n=3 is automatically lower. One aspect that the best and worst factor grouping cases have in common is that they contain the high school group. It does not matter what the predicted values are for the feature vector for the sites in the high school group. From Section 6.2, it was determined that the three high school group sites should only be associated with each other. That means the minimum accuracy for the modified clustering cases is 12% (3 sites/25 sites). The Benchmark factor grouping methods do not treat the sites near high schools differently from other sites.

Table 6.11. Factor Group Prediction Accuracy of Land use Regression Models

Factor Grouping Case	K-means (n=2) Mod (7V)	K-means (n=2) Mod (3V)	K-means (n=3) Mod (7V)	K-means (n=3) Mod (3V)	AMI and WWI	AMI Only	WWI Only
Number of Factor Groups	2	3	2	3	3	3	2
Prediction Accuracy	72%	68%	44%	52%	52%	52%	60%

With knowing the assignment accuracy of all factor grouping cases, it is important to realize that factor grouping cases that better predict the allocation of sites does not mean the actual site SADPT estimation errors are the best out of all the cases considered. As an example, case 1 has a better factor group prediction accuracy value than case 2. However, the smaller quantity of sites not predicted correctly in

case 1 have extremely high SADPT errors. Therefore, when the summary error metrics are computed for both case 1 and case 2, the summary errors for case 1 end up exceeding case 2.

6.3.3 AADPT estimation accuracy

The previous sections described two methods for determining the expansion factors for short-term count sites. For the second method (Identifying factor group), a qualitative and quantitative assessment was made on the factor group identification. In this section, the AADPT estimation errors for the two methods and the benchmark methods are computed and compared. The AADPT estimation error metrics are provided inTable 6.12. All the modified k-means clustering methods perform better than the Benchmark methods with the exception k-means case with 3 variables and n=3, and as indicated in Appendix N, most of those improvements are statistically significant. Initially, the direct estimation method obtained SADPT estimation errors that were similar to the benchmark method errors. However, after excluding the high school sites from the direct estimation process and placing those sites again in the high school group, the overall case errors went down to 42.6% from 52%.

In terms of the best performing expansion methods, again the Traditional method is the best with the lowest MAPE values, followed by the AASHTO and Disaggregate expansion methodologies. The MAE and MAPE values for all land use factor grouping cases range from 25 to 50 and from 35% to 70% respectively.

Table 6.12. AADPT Estimation Error Metrics for Short-term count sites

Metric	**MAPE**			**MAE**		
Expansion Method	**T**	**D**	**A**	**T**	**D**	**A**
Direct Est	52.0	NA	NA	40.7	NA	NA
Direct Est HSG Sep	42.6	NA	NA	26.8	NA	NA
K-means (n=2) Mod (7V)	41.5	46.5	42.9	25.9	27.4	27.1
K-means (n=2) Mod (3V)	43.8	50.7	45.8	27.8	30.7	29.4
K-means (n=3) Mod (7V)	38.4	35.4	39.1	24.0	19.5	24.5
K-means (n=3) Mod (3V)	53.2	66.2	56.2	30.6	33.6	32.3
AMI and WWI	52.1	62.2	53.9	39.1	45.6	40.7
AMI Only	52.5	59.5	54.5	40.4	45.0	42.3
WWI Only	52.4	62.7	54.5	38.5	44.7	40.1

T = Traditional; D=Disaggregate; A=AASHTO

With most of the k-means clustering cases performing the best using the predicted temporal metrics, it is again important to know it's limitations. From this analysis, it was decided to compare predicted feature vectors with factor group feature vectors using all 7 and just 3 specific variables. Even though it was explained why those two arrangements were considered, it is also possible to get much better or worse SADT estimation results with different setups of variables to compare. Therefore, it is not clear if the association between the predicted and factor feature vectors truly produced the best SADPT estimation errors. The direct estimation method with the high school group separated also performed well and it might be a more practical choice to use. This is because there is a discrete number of expansion factors that could be used for a specific dataset which means that there is less uncertainty surrounding the direct estimation SADPT error results.

The impact of having to estimate the factor group to which a short-term count site belongs can be determined by comparing SADPT estimation errors when the factor group is known to when the factor group must be estimated. This comparison is provided in *Table 6.13* where the majority cases have better performance metrics using the predicted temporal metrics. However, the WWI only case and the k-means case with 3 variables and n=3 do not have any errors metrics that outperform the land use predictions.

Table 6.13. Differences in Error Metrics when factor group is known and when it must be estimated

Metric	**Expansion Method**	**Factor Grouping Method**						
		K-means (n=2) Mod (7V)	K-means (n=2) Mod (3V)	K-means (n=3) Mod (7V)	K-means (n=3) Mod (3V)	AMI and WWI	AMI Only	WWI Only
MAPE Difference (%)	Traditional	6.1	0.9	9.8	-24.9	0.1	5.4	-8.5
	Disaggregate	11.2	3.3	26.3	-37.8	1.3	12.2	-12.3
	AASHTO	7.0	0.7	11.0	-28.1	1.2	6.0	-10.4
MAE Difference (%)	Traditional	7.1	0.4	3.8	-22.9	-0.1	9.4	-7.5
	Disaggregate	12.3	1.5	17.2	-42.2	0.7	19.1	-13.2
	AASHTO	8.0	0.2	5.0	-25.4	0.6	10.4	-8.7

Overall, the k-means case with 3 variables and clusters considered has much worse SADPT error values than its original group. The likely cause of that specific difference is that the association of just 3 specific variables with actual regression models does not seem to capture the full temporal profile of some of the sites. Therefore, it is important to try and select variables for comparison that truly reflect the temporal trends of a given factor group so that the summary errors for the case are reduced. However, as stated before, the answer for selecting the optimal collection of variables to compare between predicted and factor group metrics it always obvious.

For the individual site error metrics which includes MAPE and MAE as displayed in Appendix O, the values again seem to display a wide range of success which is again likely dependent on the site and the specific factor group the sites are placed in.

Chapter 7 Conclusions and Recommendations

7.1 Conclusions

This research was able to provide solutions to the five main research objectives outlined in the first chapter.

1. The modified AASHTO method for computing SADPT values for continuous count sites was determined to be flexible with respect to missing data entries.
2. This research was able to adapt filtering algorithms developed by Allen (2021) for cyclist count data and successfully apply them to pedestrian count data. In the datasets for both Milton, Ontario and Pima County, Arizona, the daily/8hr zero and null count filters flagged the highest percentage of problematic entries. For the daily/8hr zero filter more than 2.5% of entries were captured within the Milton dataset and for Pima County, around 20% of entries were marked as problematic by that filter. The null count filter was able to flag around 1% of the data collected in Milton and about 10% of entries in Pima County. For filters such as the IQR filter or hard cap limit, it is not intended for those filters to mark an extensive portion of the count data. Therefore, the hard cap filter had its threshold value increased and a lower limit was applied to the IQR filter and ultimately those filters marked less than 1.5% for both datasets.
3. For the SADPT expansion methodologies, the Traditional, AASHTO, and Disaggregate methods were evaluated. It was worthwhile to try all three expansion methodologies because the SADPT error metrics collected in this research yielded different performance results than those in previous studies. The majority of case by case comparisons that were made between the error metrics indicated that the Traditional method is the best and the Disaggregate method is the worst expansion method. Within the literature, the opposite is true and the likely reason why there is a difference in performance between the expansion methods is the low pedestrian volume sites. The problem with low pedestrian volume sites is that they may have perfectly valid larger count values within their dataset from time to time. When the comes time to calculate the expansion factors, the day of year scaling factor for example cannot mitigate the effect of the random value. This is because the Disaggregate expansion methodology treats every day in the study period as unique. A Traditional scaling factor requires a daily average for either a given month or day within a study period. That specific daily average considers an input of multiple data entries that have a greater potential to reduce the impact of noticeably different counts.
4. When the site error metrics are compared with some of the literature values, some of the sites have error values that are generally respectable and similar to error metrics collected from past studies. However, some of the remaining sites had extremely large error values. Therefore, some of the likely causes of those large site errors could include the incorrect factor group placement for the site and/or the variability within the collection of short-term counts associated with the site. It was also determined that both the Benchmark(the utilization of AMI and/or WWI) and the k-means factor grouping methods for continuous count sites were not the best factor grouping solutions for this dataset. However, 3 sites that are closely positioned near high schools were placed within their own unique factor group for all of the k-means clustering cases, and some of the various error metrics did seem to improve and the Disaggregate expansion method also performed the best for the 3 high school sites. However, even with the

high group considered within the clustering analysis, there is still some uncertainty with respect to the quantity of clusters and what variables to consider in that type of analysis.

5. For the land use regression models, the MAPE values were between 5% to 15% with the exception of the models used to predict AMI and WWI which range from 30% to 65%. The likely reason why those MAPE values were obtained is that most of the utilized temporal tends have a relatively small range of values. However, the adjusted r-squared values for all the land use models fell between 0.15 to 0.4. For the factor grouping cases considered for land use prediction, the assignment accuracy was above 50% except for one case that predicted 44% of the sites original factor groups. However, assignment accuracy is not always associated with the actual SADPT errors. For most factor grouping cases, the land use prediction models produced more favorable SADPT errors when compared to the factor grouping arrangements for the continuous count sites. In one specific k-means case, the land use prediction SADPT MAPE value was more than 25% better than the original SADPT MAPE value. Overall, the combination of variables to associate short-term count sites with a k-means cluster is important for produces low SADPT error values.

7.2 Recommendations

The research presented in this thesis gives rise to the following 5 recommendations:

1. The first recommendation that can be made from this work is to utilize a dataset that allows sites to have their true AADPT value computed and that has very large collection of counts sites with mostly reliable data. For this research, most of the sites in the Pima County, Arizona dataset had to be removed from the analysis because of problems identified in Chapter 4. Sites with a large portion of cleaned data had most of their data entries removed because it was collected after the COVID-19 pandemic was declared and it was questionable if that data would represent the typical temporal trends at the site. Using a large collection of count sites that have enough reliable data to compute AADPT instead of SADPT would further validate the estimation accuracy by providing more data within the given sites and across the study region as well. Another benefit to using a larger dataset with reliable annual count data is that there is the possibility of observing more temporal trends for pedestrians. Having more temporal trends could mean that there could be more factor groups to consider and that could possibly reduce AADPT estimation errors because the sites within the groups have more specific temporal trends shared.
2. Another recommendation that can be made for future research is selecting count locations in a densely populated urban center. The reason why it is important to consider sites located in those environments is because those sites likely have high pedestrian volumes and there could be more of a risk of pedestrians getting injured in those settings as well. Therefore, if pedestrian volumes are not calculated or estimated, there is no way of addressing the safety concerns that numerous pedestrians could be dealing with on a consistent basis.
3. If pedestrian count sites are placed within their correct factor group the variation in performance metrics should be relatively low and the majority of obtained performance metrics should be in line with previous studies. It also might be a good idea to compare the Traditional, AASHTO, and Disaggregate methods when utilizing more heavily urbanized pedestrian count locations because there is likely no guarantee that the best performing expansion method is going to be the same as this research or other previous studies. It is also very possible that

temporal trends for pedestrians may appear to be very different depending on the geographic location for the count sites. Therefore, it is probably worthwhile to incorporate as many geographic locations as possible for potential count sites in future studies and continue to compare AADPT expansion methodologies to see what expansion methodology work best in certain situations and to truly get an enhanced understanding pedestrian traffic at intersections.

4. Commonly used literature and k-means factor grouping methods have their limitations. Those methods do not always have the ability to reduce estimation errors for all count sites. From that, it would be good idea to identify problematic sites and once the source of the problem is identified for each site, place each of those sites within a different factor group if possible.
5. Finally, the last aspect of this research that can be improved is the land use regression modelling. It could be useful to model the distance between a specific site and a given land use as a function of a given temporal metric such as the scaling factors used in the clustering analysis or AMI/WWI. For this work, if land uses were equal to or more than 400 meters away from a given site, it was assumed that the land use had no influence on the site. However, in reality, there is no set distance for land uses to be temporally influential. Another benefit of modelling distance as a function of a temporal metric is that the independent variable (the distance value) is no longer binary which means that there could be a greater variation of predicted values for all temporal metrics considered. In general, it is expected that a collection of count sites have slightly different values from each other for all of their associated temporal metrics.

References

Allen, B. (2021). *Estimating Intersection Annual Average Daily Bicycle Traffic from 8-hour Turning Movement Counts.* Waterloo, ON: University of Waterloo.

Cao, X., Handy, S. L., & Mokhtarian, P. L. (2006). The influences of the built environment and residential self-selection on pedestrian behavior: evidence from Austin, TX. *Journal of Transportation, 33*, 1-20.

El Esawey, M. (2016). Toward a Better Estimation of Annual Average Daily Bicycle Traffic Comparison of Methods for Calculating Daily Adjustment Factors. *Journal of the Transportation Research Board, 2593*(1), 28-36.

El Esawey, M., & Mosa, A. I. (2015). Determination and Application of Standard K Factors for Bicycle Traffic. *Journal of the Transportation Research Board, 2527*, 58-68.

Environment Canada. (2021). Daily Data Report: OAKVILLE TWN ONTARIO.

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), 04014014-(1-8).

Griswold, J. B., Medury, A., Schneider, R. J., & Grembek, O. (2018). Comparison of Pedestrian Count Expansion Methods: Land Use Groups versus Empirical Clusters. *Journal of the Transportation Research Board, 2672*(43), 87-97.

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. *Journal of the Transportation Research Board, 2673*(4), 941-950.

Hankey, S., Lindsey, G., & Marshall, J. (2014). Day-of-Year Scaling Factors and Design Considerations for Nonmotorized Traffic Monitoring Programs. *Journal of the Transportation Research Board, 2468*(1), 64-73.

Jackson, K. N., Stolz, E., & Cunningham, C. (2015). Nonmotorized Site Selection Methods for Continuous and Short-Duration Volume Counting. *Journal of the Transportation Research Board, 2527*(1), 49-57.

Johnstone, D., Nordback, K., & Kothuri, S. (2018). Annual Average Nonmotorized Traffic Estimates from Manual Counts: Quantifying Error. *Journal of the Transportation Research board, 2672*(43), 134-144.

Lindsey, G., Singer-Berk, L., Wilson, J. S., Oberg, E., & Hadden-Loh, T. (2018). Challenges in Monitoring Regional Trail Traffic. *Journal of the Transportation Research Board, 2672*(43), 98-109.

McGuire, P. (2021, August 11). *Colorado's Front Range—Much More Than Big Cities*. Retrieved November 24, 2021, from Uncover Colorado: https://www.uncovercolorado.com/front-range-colorado/

Medury, A., Griswold, J. B., Huang, L., & Grembek, O. (2019). Pedestrian Count Expansion Methods: Bridging the Gap between Land Use Groups and Empirical Clusters. *Journal of the Transportation Research Board, 2673*(5).

Miranda-Moreno, L. F., Nosal, T., Schneider, R. J., & Proulx, F. (2013). Classification of Bicycle Traffic Patterns in Five North American Cities. *Journal of the Transportation Research Board, 2339*(1), 68-79.

Mohammed, H. J. (2019). *Kansas LTAP Fact Sheet Using Turning Movement Counts (TMC) at Intersections.* Lawrence, KS: University of Kansas.

National Oceanic and Atmospheric Administration. (2021). Daily Summaries.

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? *Journal of the Transportation Research Board, 2673*(10), 295-310.

Nordback, K., Marshall, W. E., & Janson, B. N. (2013a). *Development of Estimation Methodology for Bicycle and Pedestrian Volumes Based on Existing Counts.* Colorado Department of Transportation.

Nordback, K., Marshall, W. E., Janson, B. N., & Stolz, E. (2013b). Estimating Annual Average Daily Bicyclists Error and Accuracy. *Journal of the Transportation Research Board, 2339*(1), 90-97.

Nosal, T., Miranda-Moreno, L. F., & Krstulic, Z. (2014). Incorporating Weather: Comparative Analysis of Annual Average Daily Bicyclist Traffic Estimation Methods. *Journal of the Transportation Research Board, 2468*(1), 100-110.

Olfert, C., Poapst, R., & Montufar, J. (2018). Incorporating the Effect of Special Events into Continuous Count Site Selection for Pedestrian Traffic. *Journal of the Transportation Research Board, 2672*(43), 65-74.

PIMA COUNTY SCHOOL SUPERINTENDENT. (2021). *District Schools*. Retrieved October 27, 2021, from http://www.schools.pima.gov/schools/public-schools

Saneinejad, S., Roorda, M. J., & Kennedy, C. (2012). Modelling the impact of weather conditions on active transportation travel behaviour. *Journal of Transportation Research Part D, 17*(2), 129-137.

Schneider, R. J., Henry, T., Mitman, M. F., Stonehill, L., & Koehler, J. (2013). *Development and Application of the San Francisco Pedestrian Intersection Volume Model.* Berkeley, CA: University of California.

Statistics Canada. (2017, November 29). *Census Profile, 2016 Census Milton, Town [Census subdivision], Ontario and Ontario [Province]*. Retrieved April 22, 2021, from Statistics Canada: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CSD&Code1=3524009&Geo2=PR&Code2=35&SearchText=Milton&SearchType=Begins&SearchPR=01&B1=All&TABID=1&type=0

U.S. Department of Commerce. (n.d.). *QuickFacts Pima County, Arizona*. Retrieved April 22, 2021, from United States Census Bureau: https://www.census.gov/quickfacts/fact/table/pimacountyarizona/LND110210

U.S. Department of Transportation Federal Highway Administration. (2016). *Exploring Pedestrian Counting Procedures A Review and Compilation of Existing Procedures, Good Practices, and Recommendations.* Washington, D.C.

U.S. Department of Transportation Federal Highway Administration. (2016). *Traffic Monitoring Guide.* Washington, D.C.

U.S. Department of Transportation Federal Highway Administration. (2018). *Traffic Data Computation Method Pocket Guide.* Washington, D.C.

Wright, T., Hu, P. S., Young, J., & Lu, A. (1997). *VARIABILITY IN TRAFFIC MONITORING DATA FINAL SUMMARY REPORT.* Oak Ridge, Tennessee.

Appendices

Appendix A: List of Milton, Ontario Intersections

Table A.1. Milton Ontario Intersections Part 1

Site Name	Site Code	Start Date	End Date	Land Uses
Bronte Street and Derry Road	S01	No Data	No Data	Residential and Commercial
Bronte Street North and Main Street West	S02	Nov 06, 2018	Apr 16, 2020	Commercial
Fifth Line and Yukon Court	S03	No Data	No Data	Undeveloped and Industrial
James Snow Parkway and Derry Road	S04	Jun 06, 2018	Nov 11, 2019	Residential and Undeveloped
Laurier Avenue and Commercial Street	S05	Jul 23, 2019	Apr 16, 2020	Residential and Undeveloped
Laurier Avenue and Ontario Street South	S06	Jun 14, 2019	Apr 16, 2020	Residential and Commercial
Louis St Laurent Avenue and Commercial Plaza	S07	Jul 04, 2019	Apr 16, 2020	Undeveloped
Louis St Laurent Avenue and Farmstead Drive	S08	Aug 25, 2019	Apr 16, 2020	Residential and Undeveloped
Main Street and Sherwood	S09	Apr 06, 2020	Apr 16, 2020	Undeveloped
Main Street E and Harris / Pearson	S10	Oct 16, 2019	Apr 16, 2020	Residential
Main Street East and Commercial Street	S11	Apr 06, 2020	Apr 16, 2020	Commercial
Main Street East and Leisure Centre Driveway	S12	Oct 28, 2019	Apr 16, 2020	Recreational and Undeveloped
Main Street East and Mall Entrance	S13	Oct 19, 2018	Apr 16, 2020	Commercial
Main Street East and Maple / Sinclair	S14	Nov 04, 2019	Apr 16, 2020	Residential
Main Street East and Ontario Street North	S15	Oct 12, 2018	Apr 16, 2020	Commercial

Table A.2. Milton Ontario Intersections Part 2

Site Name	Site Code	Start Date	End Date	Land Uses
Main Street West and Savoline Boulevard	S16	Feb 11, 2020	Apr 16, 2020	Residential, Recreational and Undeveloped
Main Street West and Scott Boulevard	S17	Jan 30, 2020	Apr 16, 2020	Residential and Undeveloped
Martin Street and Main Street East	S18	Apr 06, 2020	Apr 16, 2020	Commercial
Ontario Street South and Childs Drive	S19	Oct 12, 2018	Apr 16, 2020	Residential, Commercial, and Undeveloped
Ontario Street South and Pine Street	S20	Oct 12, 2018	Apr 16, 2020	Residential and Commercial
Steeles Avenue East and Martin Street	S21	No Data	No Data	Commercial
Thompson Road and Childs Drive	S22	Jun 05, 2018	Apr 16, 2020	Residential
Thompson Road and Derry Road	S23	No Data	No Data	Residential and Commercial
Thompson Road and Drew Centre	S24	Jul 19, 2019	Apr 16, 2020	Commercial, Recreational, and Industrial
Thompson Road and Laurier Avenue	S25	Jun 06, 2018	Apr 16, 2020	Residential
Thompson Road and Main Street	S26	Jun 08, 2018	Apr 16, 2020	Residential and Commercial
Thompson Road and McCuaig Drive	S27	Mar 09, 2020	Apr 16, 2020	Residential
Thompson Road and Nipissing Road	S28	Sep 17, 2018	Apr 16, 2020	Residential and Commercial
Thompson Road North and Maple Avenue	S29	Jan 23, 2020	Apr 16, 2020	Residential and Commercial
Thompson Road North and Woodward Avenue	S30	Jan 27, 2020	Apr 16, 2020	Residential and Undeveloped

Appendix B: List of Pima County, Arizona Intersections

Table B.1. Pima County Arizona Intersections Part 1

Site Name	Site Code	Start Date	End Date	Land Uses
1st Av / Christie Dr / Ina Rd	S001	May 12, 2019	Sep 10, 2020	Undeveloped and Residential
1st Av / Orange Grove Rd	S002	No Data	No Data	Residential and Undeveloped
36th St / Palo Verde Rd	S003	Jan 02, 2020	Sep 10, 2020	Undeveloped, Industrial and Commercial
37th St / Golf Links Rd / Palo Verde Rd	S004	Jan 02, 2020	Sep 10, 2020	Industrial, Recreational and Commercial
Abrego Dr / Continental Rd	S005	Mar 16, 2020	Sep 10, 2020	Residential
Abrego Dr / Esperanza Bl	S006	Jun 10, 2020	Sep 10, 2020	Residential and Recreational
Aero Park Bl / Nogales Hy / Vamori St	S007	Mar 15, 2020	Sep 10, 2020	Undeveloped, Industrial and Commercial
Aerospace Pw / Nogales Hy	S008	Mar 15, 2020	Sep 10, 2020	Industrial and Undeveloped
Aerospace Pw / Raytheon Pw	S009	Mar 15, 2020	Sep 10, 2020	Undeveloped and Industrial
Ajo Wy / Alvernon Wy	S010	Mar 15, 2020	Sep 10, 2020	Industrial and Undeveloped
Ajo Wy / Dodge Bl	S011	Mar 15, 2020	Sep 10, 2020	Commercial and Undeveloped
Ajo Wy / Palo Verde Rd	S012	Mar 15, 2020	Sep 10, 2020	Commercial and Undeveloped
Alvernon Wy / Brandi Fenton Dw / River Rd	S013	Mar 01, 2020	Sep 10, 2020	Recreational and Undeveloped
Alvernon Wy / Dodge Bl	S014	Mar 03, 2020	Sep 10, 2020	Recreational
Alvernon Wy / Irvington Rd	S015	No Data	No Data	Industrial and Undeveloped
Alvernon Wy / Valencia Rd	S016	Jan 02, 2020	Sep 10, 2020	Residential and Undeveloped
Benson Hy / Palo Verde Rd	S017	No Data	No Data	Commercial and Residential
Benson Hy / Swan Rd / Valencia Rd	S018	Jan 02, 2020	Sep 10, 2020	Commercial, Residential and Undeveloped

Table B.2. Pima County Arizona Intersections Part 2

Site Name	Site Code	Start Date	End Date	Land Uses
Broadmont Dr / Palo Verde Rd	S019	No Data	No Data	Commercial and Undeveloped
Calle Bosque / Territory Dr / Craycroft Rd	S020	Jun 29, 2020	Sep 10, 2020	Residential and Undeveloped
Calle del Marques / Sunrise Dr	S021	Dec 05, 2019	Sep 10, 2020	Residential
Camino Casa Verde / La Canada Dr / Paseo del Chino	S022	Mar 15, 2020	Sep 10, 2020	Commercial and Residential
Camino de la Tierra / Ina Rd	S023	Mar 31, 2019	Sep 10, 2020	Commercial and Residential
Camino De La Tierra / Orange Grove Rd	S024	Nov 17, 2019	Sep 10, 2020	Commercial and Residential
Camino de la Tierra / Valencia Rd	S025	Feb 10, 2020	Sep 10, 2020	Commercial, Residential and Undeveloped
Camino de Oeste / Valencia Rd	S026	Feb 09, 2020	Sep 10, 2020	Commercial, Residential and Undeveloped
Camino Del Sol and Camino Encanto	S027	Feb 19, 2020	Sep 10, 2020	Recreational , Residential, and Undeveloped
Camino Verde / Valencia Rd	S028	Feb 09, 2020	Sep 10, 2020	Undeveloped, Commercial and Residential
Campbell Av / Skyline Dr	S029	Mar 30, 2020	Sep 10, 2020	Commercial
Campbell Av / Speedway Bl	S030	No Data	No Data	Commercial
Campo Abierto / Sunrise Dr	S031	Dec 05, 2019	Sep 10, 2020	Commercial and Undeveloped
Campus Park Wy / Shannon Rd	S032	Jul 06, 2020	Sep 10, 2020	Residential
Cardinal Av / Drexel Rd	S033	Jun 30, 2020	Sep 10, 2020	Commercial, Residential, and Undeveloped
Cardinal Av / Valencia Rd	S034	Mar 08, 2020	Sep 10, 2020	Commercial
Casino del Sol Dr / Valencia Rd	S035	Feb 09, 2020	Sep 10, 2020	Undeveloped and Commercial
Cloud Rd / Sabino Canyon Rd	S036	Mar 15, 2020	Sep 10, 2020	Residential

Table B.3. Pima County Arizona Intersections Part 3

Site Name	Site Code	Start Date	End Date	Land Uses
Club Dr / Shannon Rd	S037	Jun 28, 2020	Sep 10, 2020	Residential
Colossal Cave Rd / Mary Ann Cleveland Wy	S038	Nov 12, 2019	Sep 10, 2020	Residential, Commercial and Undeveloped
Continental Rd / Continental Plaza / I19 Frontage Rd	S039	No Data	No Data	Commercial
Continental Rd / La Canada Dr	S040	Mar 11, 2020	Sep 10, 2020	Commercial and Recreational
Cortaro Farms Rd / Magee Rd / Shannon Rd	S041	Jan 05, 2020	Sep 10, 2020	Residential and Undeveloped
Cortaro Farms Rd / Oldfather Dr	S042	Jul 14, 2020	Sep 10, 2020	Residential
Cortaro Farms Rd / Thornydale Rd	S043	Jan 05, 2020	Sep 10, 2020	Residential, Commercial and Undeveloped
Craycroft Rd / River Rd	S044	Mar 11, 2020	Sep 10, 2020	Residential, Recreational, Commercial and Undeveloped
Craycroft Rd / Sunrise Dr	S045	Nov 20, 2019	Sep 10, 2020	Residential and Undeveloped
Curtis Rd / La Cholla Bl	S046	Mar 11, 2020	Sep 10, 2020	Commercial, Recreational, and Undeveloped
Desert Bell Dr / La Canada Dr / La Canoa	S047	Mar 16, 2020	Sep 10, 2020	Commercial, Residential and Undeveloped
Desert View HS Dw / Valencia Rd	S048	No Data	No Data	Residential and Undeveloped
Drexel Rd / Mission Rd	S049	No Data	No Data	Commercial and Undeveloped
Drexel Rd / Palo Verde Rd	S050	Jan 02, 2020	Sep 10, 2020	Residential and Commercial
East Catalina Highway and North Mount Lemmon Short Road	S051	Jul 14, 2020	Sep 10, 2020	Undeveloped and Residential
Esperanza Bl / La Canada Dr	S052	Mar 11, 2020	Sep 10, 2020	Commercial
Flowing Wells Rd / Wetmore Rd	S053	Mar 11, 2020	Sep 10, 2020	Commercial
Foothills Mall Dr / La Cholla Bl	S054	Mar 11, 2020	Sep 10, 2020	Commercial

Table B.4. Pima County Arizona Intersections Part 4

Site Name	Site Code	Start Date	End Date	Land Uses
Hardy Rd / La Canada Dr / Overton Rd	S055	Jan 05, 2020	Sep 08, 2020	Residential
Hardy Rd / Thornydale Rd	S056	Jan 05, 2020	Sep 10, 2020	Residential, and Undeveloped
Hermans Rd / Nogales Hy	S057	Mar 11, 2020	Sep 10, 2020	Undeveloped, Industrial and Commercial
Hospital Dr / La Cholla Bl	S058	Mar 11, 2020	Sep 10, 2020	Commercial and Industrial
Houghton Rd / Sahuarita Rd	S059	Mar 11, 2020	Sep 10, 2020	Commercial, Residential and Undeveloped
Ina Rd / La Canada Dr	S060	Jun 10, 2019	Sep 10, 2020	and Residential
Ina Rd / La Cholla Bl	S061	Apr 23, 2020	Sep 10, 2020	Commercial
Ina Rd / Mona Lisa Rd	S062	May 07, 2019	Sep 10, 2020	Commercial, Residential, and Undeveloped
Ina Rd / Pima Canyon Dr / Skyline Dr	S063	Nov 20, 2019	Sep 10, 2020	Commercial
Ina Rd / Shannon Rd	S064	Apr 27, 2020	Sep 10, 2020	Commercial and Recreational
Ina Rd / Westward Look Dr	S065	Nov 19, 2019	Sep 10, 2020	Residential and Undeveloped
ITD VLAN Test	S066	No Data	No Data	Commercial
Kinney Rd / Western Wy	S067	Aug 25, 2020	Sep 10, 2020	Commercial and Residential
Knollwood Dr / River Rd / Sabino Canyon Rd	S068	Apr 13, 2020	Sep 10, 2020	Residential
Kolb Rd / Mountain Shadows Pl / Ventana Canyon Dr	S069	Mar 11, 2020	Sep 10, 2020	Residential and Recreational
Kolb Rd / Sabino Canyon Rd	S070	Apr 21, 2020	Sep 10, 2020	Residential and Undeveloped
Kolb Rd / Snyder Rd	S071	Mar 10, 2020	Sep 10, 2020	Residential and Undeveloped
Kolb Rd / Sunrise Dr	S072	Oct 09, 2019	Sep 10, 2020	Commercial

Table B.5. Pima County Arizona Intersections Part 5

Site Name	Site Code	Start Date	End Date	Land Uses
La Canada Dr / Magee Rd	S073	Mar 11, 2020	Sep 10, 2020	Residential
La Canada Dr / Orange Grove Rd	S074	Apr 23, 2020	Sep 10, 2020	Commercial, Residential, and Undeveloped
La Canada Dr / River Rd	S075	Dec 10, 2019	Sep 10, 2020	Commercial
La Cholla Bl / Magee Rd	S076	Jan 05, 2020	Sep 10, 2020	Residential, Commercial and Undeveloped
La Cholla Bl / Omar Dr	S077	Mar 10, 2020	Sep 10, 2020	Residential and Recreational
La Cholla Bl / Orange Grove Rd	S078	Nov 18, 2019	Sep 10, 2020	Commercial
La Cholla Bl / Overton Rd	S079	Jan 05, 2020	Sep 10, 2020	Residential and Undeveloped
La Cholla Bl / River Rd	S080	Feb 24, 2020	Sep 10, 2020	Commercial
La Cholla Bl / Rudasill Rd	S081	Mar 16, 2020	Sep 10, 2020	Commercial and Undeveloped
La Cholla Bl / Ruthrauff Rd	S082	May 07, 2019	Sep 10, 2020	Commercial
Linda Vista Bl / Thornydale Rd	S083	Oct 10, 2018	Sep 10, 2020	Recreational and Undeveloped
Magee Rd / Shannon Rd / Tuscany Dr	S084	Jan 05, 2020	Sep 10, 2020	Recreational
Magee Rd / Thornydale Rd	S085	Oct 30, 2019	Sep 10, 2020	Residential and Commercial
Mark Rd / Valencia Rd	S086	Feb 10, 2020	Sep 10, 2020	Undeveloped and Residential
Maryvale Av / Ruthrauff Rd	S087	No Data	No Data	Undeveloped
Mission Rd / Valencia Rd	S088	Apr 23, 2020	Sep 10, 2020	Undeveloped and Residential
Nogales Hy / Old Nogales Hy	S089	Mar 10, 2020	Sep 10, 2020	Commercial, Residential and Undeveloped
Orange Grove Rd / Shannon Rd	S090	Nov 18, 2019	Sep 10, 2020	Residential
Orange Grove Rd / Skyline Dr	S091	Nov 20, 2019	Sep 10, 2020	Residential

Table B.6. Pima County Arizona Intersections Part 6

Site Name	Site Code	Start Date	End Date	Land Uses
Overton Rd / Shannon Rd	S092	Jan 05, 2020	Sep 10, 2020	Residential and Undeveloped
Overton Rd / Thornydale Rd	S093	Oct 09, 2018	Sep 10, 2020	Commercial, Recreational and Undeveloped
Pontatoc Rd / River Rd	S094	Aug 17, 2020	Sep 10, 2020	Residential and Recreational
Pontatoc Rd / Sunrise Dr	S095	Dec 05, 2019	Sep 10, 2020	Residential and Undeveloped
River Rd / Swan Rd	S096	Mar 10, 2020	Sep 10, 2020	Residential and Undeveloped
River Rd / Via Entrada	S097	Mar 10, 2020	Sep 10, 2020	Commercial and Residential
Romero Rd / Ruthrauff Rd	S098	Sep 29, 2019	Sep 10, 2020	Commercial, Residential, and Recreational
Romero Rd / Wetmore Rd	S099	Mar 16, 2020	Sep 10, 2020	Residential, Commercial and Undeveloped
Sabino Canyon Rd / Snyder Rd	S100	Aug 18, 2020	Sep 10, 2020	Residential, Commercial and Undeveloped
Sabino Canyon Rd / Sunrise Dr	S101	Oct 03, 2019	Sep 10, 2020	Commercial, Residential, Recreational and Undeveloped
San Marcos / Mission Road (Shop Test Cabinet)	S102	No Data	No Data	Industrial
Silverbell Rd / Sunset Rd	S103	May 27, 2020	Sep 10, 2020	Residential and Undeveloped
Speedway Bl and Alvernon Wy	S104	No Data	No Data	Commercial
Speedway Bl and Tucson Bl	S105	No Data	No Data	Commercial
Suncrest Pl / Sunrise Dr	S106	No Data	No Data	Residential
Sunrise Dr / Swan Rd	S107	Dec 05, 2019	Sep 10, 2020	Commercial
Sunrise Dr / Via Palomita	S108	No Data	No Data	Undeveloped
Tanque Verde Loop Road and Tanque Verde Road	S109	Jun 16, 2020	Sep 10, 2020	Undeveloped
Valencia Rd / Wade Rd	S110	May 13, 2020	Sep 10, 2020	Undeveloped

Appendix C: Filtering Results for Milton, Ontario by Site (Jul 2019 to Feb 2020)

Table C.1. Daily Filtering Results for Milton Ontario

Site Code	Null Counts (%)	Non-Consecutive Zeros (%)	Hard Cap (%)	Daily Zeros (%)	IQR (%)
S02	0	0	0	0	0
S05	2.2	0	0	0	0
S06	6.6	0	0	4.4	0
S07	0	0	0	31.9	0
S13	0	1.1	0	0	0
S15	0	0	0	0	3.3
S19	0	0	0	0	0
S20	0	0	0	0	3.3
S22	0	0	0	0	0
S24	0	0	2.2	0	2.2
S25	0	0	0	0	0
S26	0	0	1.1	0	2.2
S28	0	0	0	0	1.1

Table C.2. 8 Hour Filtering Results for Milton Ontario

Site Code	Null Counts (%)	Non-Consecutive Zeros (%)	Hard Cap (%)	8hr Zeros (%)	IQR (%)
S02	2.2	0	0	0	0
S05	2.2	0	0	0	0
S06	6.6	0	0	4.4	1.1
S07	0	0	0	56	0
S13	0	1.1	0	0	0
S15	0	0	0	0	3.3
S19	0	0	0	0	0
S20	0	0	0	0	3.3
S22	0	0	0	0	0
S24	0	0	1.1	0	2.2
S25	0	0	0	0	1.1
S26	0	0	0	0	2.2
S28	1.1	0	0	0	1.1

Table C.3. 15 Minute Filtering Results for Milton Ontario

Site Code	Null Counts (%)	Non-Consecutive Zeros (%)	Hard Cap (%)
S02	3	0	0
S05	0.8	0	0
S06	7.4	0	0
S07	0.1	0	0
S13	0.05	0.1	0
S15	3.3	0	0
S19	0.05	0	0
S20	0.05	0	0
S22	0.05	0	0
S24	0.05	0	0.2
S25	0.05	0	0
S26	1.2	0	0.01
S28	0.2	0	0

Appendix D: Filtering Results for Pima County, Arizona by Site (Jan 2020 to Mar 2020)

Table D.1. Daily Filtering Results for Pima County Arizona

Site Code	Null Counts (%)	Non-Consecutive Zeros (%)	Hard Cap (%)	Daily Zeros (%)	IQR (%)
S001	0	0	0	0	0
S003	100	0	0	1.1	0
S004	2.2	0	0	1.1	0
S016	57.8	0	0	100	0
S018	3.3	0	0	1.1	0
S021	0	0	0	0	0
S023	0	0	0	0	1.1
S024	0	0	0	0	0
S031	4.4	0	0	2.2	0
S038	0	0	0	100	0
S041	2.3	0	0	1.1	1.1
S043	2.3	0	0	98.9	0
S045	0	0	0	2.2	0
S050	2.2	0	0	1.1	0
S055	2.3	0	0	1.1	0
S056	2.3	0	0	1.1	0
S060	47.3	0	0	0	0
S062	0	0	0	0	1.1
S063	100	0	0	13.2	0
S065	0	0	0	0	0
S072	0	0	0	0	0
S075	8.8	0	0	98.9	0
S076	2.3	0	0	100	0
S078	0	0	0	0	0
S079	2.3	0	0	3.4	3.4
S082	0	0	0	0	0
S083	1.1	0	0	0	0
S084	4.6	0	0	1.1	2.3
S085	78	0	0	0	0
S090	0	0	0	0	0
S091	3.3	0	0	100	0
S092	2.3	0	0	100	0
S093	2.2	0	0	1.1	0
S095	0	0	0	98.9	0
S098	1.1	0	0	0	1.1
S101	2.2	0	0	0	1.1
S107	1.1	0	0	0	0

Table D.2. 8 Hour Filtering Results for Pima County Arizona

Site Code	Null Counts (%)	Non-Consecutive Zeros (%)	Hard Cap (%)	8hr Zeros (%)	IQR (%)
S001	0	0	0	4.4	0
S003	100	0	0	1.1	0
S004	2.2	0	0	1.1	0
S016	33.3	0	0	100	0
S018	2.2	0	0	1.1	0
S021	0	0	0	0	0
S023	0	0	0	0	1.1
S024	1.1	0	0	0	0
S031	5.5	0	0	2.2	0
S038	0	0	0	100	0
S041	2.3	0	0	1.1	0
S043	2.3	0	0	98.9	0
S045	0	0	0	7.7	0
S050	3.3	0	0	1.1	0
S055	3.4	0	0	2.3	0
S056	2.3	0	0	1.1	0
S060	17.6	0	0	0	0
S062	0	0	0	0	0
S063	100	0	0	31.9	0
S065	0	0	0	2.2	0
S072	0	0	0	0	0
S075	4.4	0	0	98.9	0
S076	2.3	0	0	100	0
S078	0	0	0	0	0
S079	2.3	0	0	9.2	0
S082	0	0	0	0	0
S083	2.2	0	0	0	0
S084	5.7	0	0	1.1	2.3
S085	78	0	0	0	0
S090	0	0	0	0	0
S091	3.3	0	0	100	0
S092	2.3	0	0	100	0
S093	2.2	0	0	1.1	0
S095	2.2	0	0	98.9	0
S098	1.1	0	0	0	0
S101	1.1	0	0	0	0
S107	1.1	0	0	0	0

Table D.3. 15 Minute Filtering Results for Pima County Arizona

Site Code	Null Counts (%)	Non-Consecutive Zeros (%)	Hard Cap (%)
S001	0	0	0
S003	70.6	0	0
S004	1	0	0
S016	18.1	0	0
S018	9.1	0	0
S021	0	0	0
S023	0.03	0	0
S024	0.1	0	0
S031	3.4	0	0
S038	0	0	0
S041	1	0	0
S043	4	0	0
S045	1.4	0	0
S050	1.1	0	0
S055	1.5	0	0
S056	1	0	0
S060	17.8	0	0
S062	0	0	0
S063	97.4	0	0
S065	0	0	0
S072	0	0	0
S075	17	0	0
S076	1	0	0
S078	6.8	0	0
S079	0.9	0	0
S082	1.2	0	0
S083	0.5	0	0
S084	1.9	0	0
S085	52.6	0	0
S090	0	0	0
S091	1.8	0	0
S092	0.9	0	0
S093	1.6	0	0
S095	0.2	0	0
S098	0.3	0	0
S101	0.8	0	0
S107	7.2	0	0

Appendix E: T-Test Results by Factor Grouping Methods

Table E.1. T-Test Results for MAPE Values

Case	WWI Only	AMI Only	AMI and WWI	K-means (n=3)
K-means (n=2)	0.101	**0.004**	0.490	0.081
K-means (n=3)	**0.020**	**0.021**	**0.024**	NA
AMI and WWI	**0.030**	**0.016**	NA	NA
AMI Only	**0.024**	NA	NA	NA

Values in **bold** are statistically significant at 95% confidence level

Table E.2. T-Test Results for MAE Values

Case	WWI Only	AMI Only	AMI and WWI	K-means (n=3)
K-means (n=2)	0.125	**0.010**	0.322	0.087
K-means (n=3)	**0.027**	**0.032**	**0.035**	NA
AMI and WWI	**0.041**	**0.029**	NA	NA
AMI Only	**0.034**	NA	NA	NA

Values in **bold** are statistically significant at 95% confidence level

Appendix F: SADPT Estimation Metrics for Sites in Pima County, Arizona

Table F.1. MAPE Traditional (Original Unmodified)

Site Code	K-means (n=2)	K-means (n=3)	AMI and WWI	AMI Only	WWI Only
S001	43.4	45.1	40.3	39.8	43.0
S004	46.8	52.4	32.7	32.4	28.3
S018	27.5	26.2	28.5	30.7	33.8
S021	127.9	106.9	113.4	140.9	108.3
S023	16.8	16.2	31.2	24.3	16.1
S024	73.1	29.2	28.9	33.3	28.6
S031	64.8	67.3	64.5	66.0	67.6
S041	25.4	30.0	30.9	27.1	31.6
S045	39.5	46.0	46.6	42.3	54.4
S050	29.1	33.1	19.8	20.1	22.0
S055	32.7	32.7	52.1	51.4	36.0
S056	40.4	47.6	45.7	46.4	48.1
S062	38.3	30.0	57.9	54.4	30.2
S065	59.7	61.9	55.7	56.1	57.2
S072	24.1	26.1	27.5	27.1	28.4
S078	34.0	23.7	60.3	50.4	24.3
S079	74.9	78.3	65.0	66.1	67.3
S082	18.6	14.7	40.8	31.7	44.7
S083	195.3	167.8	180.9	242.0	170.8
S084	26.9	41.6	43.9	33.1	41.4
S090	36.5	30.4	31.2	51.1	30.4
S093	34.9	46.3	48.4	40.1	61.2
S098	28.2	20.2	41.3	36.9	21.0
S101	30.0	26.2	25.5	26.9	26.1
S107	105.7	86.6	92.7	116.8	88.0

Table F.2. MAPE AASHTO (Original Unmodified)

Site Code	K-means (n=2)	K-means (n=3)	AMI and WWI	AMI Only	WWI Only
S001	42.4	44.9	40.6	38.9	44.8
S004	52.3	59.0	39.7	35.1	28.5
S018	27.4	27.1	27.4	32.1	35.8
S021	133.8	107.2	117.1	147.7	109.2
S023	18.6	18.4	33.8	27.8	18.4
S024	80.5	30.6	31.1	36.3	30.9
S031	65.2	68.0	62.8	67.4	68.8
S041	25.9	34.3	37.1	30.9	34.2
S045	39.7	46.1	45.6	41.5	52.2
S050	31.3	38.3	23.1	20.4	21.8
S055	36.4	36.8	51.7	51.0	37.7
S056	39.5	46.7	47.3	45.3	44.7
S062	42.8	31.3	67.9	59.3	30.7
S065	61.4	64.0	54.2	57.1	57.8
S072	22.3	26.2	28.2	25.9	29.2
S078	37.7	24.6	64.2	57.2	25.2
S079	68.7	71.3	63.6	61.7	63.7
S082	22.0	16.0	44.2	38.4	50.5
S083	202.8	168.7	184.7	249.6	172.2
S084	28.9	46.6	48.2	37.1	45.5
S090	40.3	30.2	30.1	54.7	30.2
S093	35.1	48.5	51.1	42.1	63.8
S098	30.7	21.6	46.5	41.8	22.8
S101	29.2	27.0	25.3	27.3	26.2
S107	111.0	87.7	96.5	122.9	89.4

Table F.3. MAPE Disaggregate (Original Unmodified)

Site Code	K-means (n=2)	K-means (n=3)	AMI and WWI	AMI Only	WWI Only
S001	52.9	60.0	48.2	44.4	61.8
S004	66.2	78.9	48.8	44.1	32.0
S018	34.7	33.1	41.5	42.5	48.0
S021	156.1	98.2	110.3	144.6	98.2
S023	32.9	18.5	57.2	56.0	19.5
S024	111.6	23.7	24.6	29.4	21.6
S031	60.7	68.3	62.5	62.5	82.2
S041	29.2	54.5	55.1	42.4	45.7
S045	41.2	56.3	52.8	45.6	69.9
S050	55.4	69.7	35.0	29.3	37.4
S055	44.9	49.7	59.0	59.9	44.1
S056	41.9	60.3	49.3	45.5	52.1
S062	52.7	26.2	81.9	72.8	25.6
S065	56.9	61.2	55.3	57.4	56.4
S072	25.6	47.4	46.1	33.0	48.2
S078	55.1	25.3	91.8	87.8	26.3
S079	73.3	79.0	67.9	63.7	63.6
S082	37.3	16.2	68.6	67.4	78.4
S083	234.7	159.1	178.9	284.9	161.3
S084	33.3	60.1	61.1	48.8	52.9
S090	52.1	26.8	29.1	64.7	26.3
S093	39.4	65.1	62.3	48.6	95.2
S098	46.3	27.9	70.1	70.8	31.8
S101	30.4	29.8	26.6	30.0	37.1
S107	130.4	79.5	90.2	119.6	79.6

Table F.4. MAE Traditional (Original Unmodified)

Site Code	K-means (n=2)	K-means (n=3)	AMI and WWI	AMI Only	WWI Only
S001	3.3	3.4	3.0	3.0	3.2
S004	9.0	10.1	6.3	6.2	5.5
S018	15.4	14.7	15.9	17.2	18.9
S021	72.7	60.8	64.5	80.1	61.6
S023	15.1	14.5	28.0	21.9	14.5
S024	84.9	33.9	33.5	38.7	33.2
S031	9.3	9.7	9.3	9.5	9.7
S041	8.9	10.5	10.8	9.5	11.0
S045	2.5	2.9	3.0	2.7	3.5
S050	19.8	22.5	13.5	13.7	14.9
S055	4.8	4.8	7.7	7.6	5.3
S056	14.5	17.1	16.4	16.7	17.2
S062	13.4	10.5	20.2	19.0	10.6
S065	5.5	5.8	5.2	5.2	5.3
S072	18.8	20.4	21.4	21.2	22.1
S078	35.3	24.6	62.5	52.2	25.2
S079	4.8	5.0	4.1	4.2	4.3
S082	39.1	31.1	86.0	66.7	94.2
S083	338.6	290.9	313.6	419.5	296.2
S084	8.0	12.4	13.1	9.9	12.4
S090	6.1	5.1	5.2	8.5	5.1
S093	8.0	10.7	11.1	9.2	14.1
S098	8.8	6.3	12.9	11.5	6.6
S101	16.4	14.3	13.9	14.7	14.2
S107	222.7	182.6	195.4	246.2	185.4

Table F.5. MAE AASHTO (Original Unmodified)

Site Code	K-means (n=2)	K-means (n=3)	AMI and WWI	AMI Only	WWI Only
S001	3.2	3.4	3.1	2.9	3.4
S004	10.1	11.4	7.6	6.8	5.5
S018	15.3	15.2	15.3	18.0	20.0
S021	76.1	61.0	66.6	84.0	62.1
S023	16.7	16.5	30.4	24.9	16.5
S024	93.4	35.5	36.1	42.2	35.9
S031	9.4	9.8	9.0	9.7	9.9
S041	9.0	12.0	12.9	10.8	11.9
S045	2.5	2.9	2.9	2.7	3.3
S050	21.3	26.0	15.7	13.8	14.8
S055	5.4	5.4	7.6	7.5	5.6
S056	14.2	16.7	17.0	16.2	16.0
S062	15.0	10.9	23.8	20.7	10.7
S065	5.7	5.9	5.0	5.3	5.4
S072	17.4	20.5	22.0	20.3	22.8
S078	39.0	25.5	66.6	59.3	26.2
S079	4.4	4.5	4.0	3.9	4.0
S082	46.2	33.7	93.0	80.8	106.4
S083	351.6	292.5	320.3	432.7	298.6
S084	8.6	13.9	14.4	11.1	13.6
S090	6.7	5.1	5.0	9.2	5.0
S093	8.1	11.2	11.8	9.7	14.7
S098	9.6	6.8	14.5	13.1	7.1
S101	15.9	14.7	13.8	14.9	14.3
S107	234.0	184.8	203.3	259.2	188.5

Table F.6. MAE Disaggregate (Original Unmodified)

Site Code	K-means (n=2)	K-means (n=3)	AMI and WWI	AMI Only	WWI Only
S001	4.0	4.5	3.6	3.3	4.6
S004	12.7	15.2	9.4	8.5	6.2
S018	19.5	18.5	23.2	23.8	26.9
S021	88.8	55.8	62.7	82.2	55.9
S023	29.5	16.6	51.3	50.3	17.5
S024	129.5	27.5	28.6	34.1	25.0
S031	8.7	9.8	9.0	9.0	11.8
S041	10.2	19.0	19.2	14.8	15.9
S045	2.6	3.6	3.4	2.9	4.5
S050	37.7	47.4	23.8	19.9	25.4
S055	6.6	7.4	8.7	8.9	6.5
S056	15.0	21.6	17.7	16.3	18.7
S062	18.4	9.2	28.7	25.5	9.0
S065	5.3	5.7	5.1	5.3	5.2
S072	20.0	37.0	36.0	25.8	37.7
S078	57.2	26.2	95.1	91.0	27.2
S079	4.7	5.0	4.3	4.0	4.0
S082	78.5	34.0	144.4	142.0	165.1
S083	406.9	275.9	310.2	494.0	279.6
S084	10.0	17.9	18.2	14.6	15.8
S090	8.7	4.5	4.9	10.8	4.4
S093	9.1	15.0	14.3	11.2	21.9
S098	14.5	8.7	21.9	22.1	9.9
S101	16.6	16.3	14.5	16.4	20.3
S107	274.8	167.7	190.1	252.1	167.9

Appendix G: Updated T-Test Results by Factor Grouping Methods

Table G.1. Updated T-Test Results for MAPE Values

Case	WWI Only	AMI Only	AMI and WWI	K-means (n=3)	K-means (n=2)	K-means (n=3) Mod
K-means (n=2) Mod	**0.006**	**0.010**	**0.008**	**0.005**	**0.032**	0.080
K-means (n=3) Mod	**0.013**	**0.018**	**0.019**	**0.015**	**0.042**	NA
K-means (n=2)	0.101	**0.004**	0.490	0.081	NA	NA
K-means (n=3)	**0.020**	**0.021**	**0.024**	NA	NA	NA
AMI and WWI	**0.030**	**0.016**	NA	NA	NA	NA
AMI Only	**0.024**	NA	NA	NA	NA	NA

Values in **bold** are statistically significant at 95% confidence level

Table G.2. Updated T-Test Results for MAE Values

Case	WWI Only	AMI Only	AMI and WWI	K-means (n=3)	K-means (n=2)	K-means (n=3) Mod
K-means (n=2) Mod	**0.001**	**0.014**	**0.008**	**0.010**	**0.035**	0.083
K-means (n=3) Mod	**0.016**	**0.024**	**0.022**	**0.013**	**0.046**	NA
K-means (n=2)	0.125	**0.010**	0.322	0.087	NA	NA
K-means (n=3)	**0.027**	**0.032**	**0.035**	NA	NA	NA
AMI and WWI	**0.041**	**0.029**	NA	NA	NA	NA
AMI Only	**0.034**	NA	NA	NA	NA	NA

Values in **bold** are statistically significant at 95% confidence level

Appendix H: Updated SADPT Estimation Metrics for Sites in Pima County, Arizona

Table H.1. MAPE Values (Original Modified)

Site Code	K-means (n=2) Modified			K-means (n=3) Modified		
	Traditional	AASHTO	Disaggregate	Traditional	AASHTO	Disaggregate
S001	43.4	42.4	52.9	45.1	44.9	60.0
S004	46.8	52.3	66.2	52.4	59.0	78.9
S018	29.7	29.5	38.0	25.5	26.0	30.4
S021	54.9	56.4	28.4	54.9	56.4	28.4
S023	21.4	23.2	42.9	16.0	17.2	17.6
S024	73.1	80.5	111.6	34.4	35.6	31.4
S031	65.2	65.4	63.0	67.3	68.0	68.3
S041	25.8	28.0	35.5	30.0	34.3	54.5
S045	42.2	42.2	45.2	46.0	46.1	56.3
S050	29.1	31.3	55.4	33.1	38.3	69.7
S055	32.7	36.4	44.9	32.7	36.8	49.7
S056	42.8	42.5	47.7	47.6	46.7	60.3
S062	46.4	51.9	68.3	37.8	38.5	36.0
S065	59.7	61.4	56.9	61.9	64.0	61.2
S072	24.5	23.7	30.3	26.1	26.2	47.4
S078	43.9	48.5	70.7	33.6	34.0	35.3
S079	74.9	68.7	73.3	78.3	71.3	79.0
S082	26.6	31.0	50.6	18.4	18.9	21.5
S083	109.9	109.5	80.7	109.9	109.5	80.7
S084	30.7	33.7	39.4	41.6	46.6	60.1
S090	43.3	48.4	64.9	36.1	36.6	36.5
S093	37.9	39.8	45.9	46.3	48.5	65.1
S098	34.4	37.8	56.1	23.7	25.1	28.1
S101	26.8	26.3	25.9	26.2	27.0	29.8
S107	39.4	42.8	14.3	39.4	42.8	14.3

Table H.2. MAE Values (Original Modified)

Site Code	K-means (n=2) Modified			K-means (n=3) Modified		
	Traditional	AASHTO	Disaggregate	Traditional	AASHTO	Disaggregate
S001	3.3	3.2	4.0	3.4	3.4	4.5
S004	9.0	10.1	12.7	10.1	11.4	15.2
S018	16.6	16.5	21.3	14.3	14.6	17.0
S021	31.2	32.1	16.1	31.2	32.1	16.1
S023	19.2	20.9	38.5	14.4	15.5	15.8
S024	84.9	93.4	129.5	40.0	41.4	36.5
S031	9.4	9.4	9.1	9.7	9.8	9.8
S041	9.0	9.8	12.4	10.5	12.0	19.0
S045	2.7	2.7	2.9	2.9	2.9	3.6
S050	19.8	21.3	37.7	22.5	26.0	47.4
S055	4.8	5.4	6.6	4.8	5.4	7.4
S056	15.4	15.2	17.1	17.1	16.7	21.6
S062	16.2	18.1	23.9	13.2	13.5	12.6
S065	5.5	5.7	5.3	5.8	5.9	5.7
S072	19.1	18.5	23.7	20.4	20.5	37.0
S078	45.5	50.3	73.3	34.9	35.3	36.6
S079	4.8	4.4	4.7	5.0	4.5	5.0
S082	56.0	65.2	106.5	38.8	39.7	45.3
S083	190.5	189.8	139.9	190.5	189.8	139.9
S084	9.2	10.1	11.8	12.4	13.9	17.9
S090	7.2	8.1	10.9	6.0	6.1	6.1
S093	8.7	9.2	10.6	10.7	11.2	15.0
S098	10.8	11.8	17.5	7.4	7.8	8.8
S101	14.7	14.4	14.1	14.3	14.7	16.3
S107	83.0	90.2	30.0	83.0	90.2	30.0

Appendix I: AMI Land Use Model

```
> Model2<-ols_step_both_p(lm(AMI~SchoolFlag+Residential+Park_Rec+ConvSFlag+RetailFlag,data=Pima_Reg_In),pent=0.05,prem=0.2, progress= TRUE)
Stepwise Selection Method
---------------------------

Candidate Terms:

1. SchoolFlag
2. Residential
3. Park_Rec
4. ConvSFlag
5. RetailFlag

We are selecting variables based on p value...

Variables Entered/Removed:

+ SchoolFlag
+ Park_Rec

No more variables to be added/removed.


Final Model Output
------------------

                        Model Summary
-------------------------------------------------------------
R                       0.668       RMSE                0.791
R-Squared               0.446       Coef. Var          72.559
Adj. R-Squared          0.395       MSE                 0.626
Pred R-Squared          0.121       MAE                 0.529
-------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------
              Sum of
             Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression     11.080          2          5.540    8.848    0.0015
Residual       13.775         22          0.626
Total          24.855         24
--------------------------------------------------------------------

                                 Parameter Estimates
-----------------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
-----------------------------------------------------------------------------------------------
(Intercept)     0.954         0.192                4.973    0.000     0.556     1.352
 SchoolFlag     1.581         0.394        0.677     4.009    0.001     0.763     2.399
   Park_Rec    -1.014         0.394       -0.434    -2.570    0.017    -1.831    -0.196
-----------------------------------------------------------------------------------------------

> car::vif(Model2$model)
SchoolFlag   Park_Rec
  1.132549   1.132549
```

Appendix J: WWI Land Use Model

```
> Model2<-ols_step_both_p(lm(WWI~SchoolFlag+Residential+Park_Rec+ConvSFlag+RetailFlag,data=Pima_Reg_In),pent=0.05,prem=0.2, progress= TRUE)
Stepwise Selection Method
---------------------------

Candidate Terms:

1. SchoolFlag
2. Residential
3. Park_Rec
4. ConvSFlag
5. RetailFlag

We are selecting variables based on p value...

Variables Entered/Removed:

+ SchoolFlag

No more variables to be added/removed.


Final Model Output
------------------

                        Model Summary
-------------------------------------------------------------
R                       0.553       RMSE                 0.297
R-Squared               0.306       Coef. Var           32.257
Adj. R-Squared          0.275       MSE                  0.088
Pred R-Squared          0.125       MAE                  0.230
-------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
--------------------------------------------------------------------
              Sum of
             Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression      0.893         1          0.893    10.126    0.0042
Residual        2.029        23          0.088
Total           2.922        24
--------------------------------------------------------------------

                                 Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     1.027         0.068                15.072    0.000     0.886     1.168
 SchoolFlag    -0.443         0.139       -0.553    -3.182    0.004    -0.730    -0.155
-----------------------------------------------------------------------------------------

> car::vif(Model2$model)
Error in vif.default(Model2$model) : model contains fewer than 2 terms
```

Appendix K: January Scaling Factor Land Use Model

```
> Model2<-ols_step_both_p(lm(Jan~SchoolFlag+Residential+Park_Rec+ConvSFlag+RetailFlag,data=Pima_Reg_In),pent=0.05,prem=0.2, progress= TRUE)
Stepwise Selection Method
---------------------------

Candidate Terms:

1. SchoolFlag
2. Residential
3. Park_Rec
4. ConvSFlag
5. RetailFlag

We are selecting variables based on p value...

Variables Entered/Removed:

+ ConvSFlag

No more variables to be added/removed.


Final Model Output
------------------

                        Model Summary
-------------------------------------------------------------
R                       0.473       RMSE                0.091
R-Squared               0.223       Coef. Var           9.142
Adj. R-Squared          0.189       MSE                 0.008
Pred R-Squared          0.068       MAE                 0.071
-------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                               ANOVA
-----------------------------------------------------------------
              Sum of
             Squares        DF    Mean Square      F        Sig.
-----------------------------------------------------------------
Regression      0.055         1          0.055    6.611    0.0171
Residual        0.192        23          0.008
Total           0.247        24
-----------------------------------------------------------------

                                 Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     1.019         0.020                51.156    0.000     0.978     1.060
  ConvSFlag    -0.128         0.050       -0.473    -2.571    0.017    -0.231    -0.025
-----------------------------------------------------------------------------------------

> car::vif(Model2$model)
Error in vif.default(Model2$model) : model contains fewer than 2 terms
```

Appendix L: March Scaling Factor Land Use Model

```
> Model2<-ols_step_both_p(lm(Mar~SchoolFlag+Residential+Park_Rec+ConvSFlag+RetailFlag,data=Pima_Reg_In),pent=0.05,prem=0.2, progress= TRUE)
Stepwise Selection Method
---------------------------

Candidate Terms:

1. SchoolFlag
2. Residential
3. Park_Rec
4. ConvSFlag
5. RetailFlag

We are selecting variables based on p value...

Variables Entered/Removed:

+ ConvSFlag

No more variables to be added/removed.


Final Model Output
------------------

                        Model Summary
-------------------------------------------------------------
R                       0.503       RMSE                0.156
R-Squared               0.253       Coef. Var          15.263
Adj. R-Squared          0.220       MSE                 0.024
Pred R-Squared          0.112       MAE                 0.112
-------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
-----------------------------------------------------------------------
              Sum of
             Squares        DF    Mean Square      F         Sig.
-----------------------------------------------------------------------
Regression      0.188         1          0.188    7.771    0.0105
Residual        0.557        23          0.024
Total           0.745        24
-----------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------
      model     Beta    Std. Error    Std. Beta      t        Sig     lower    upper
-----------------------------------------------------------------------------------------
(Intercept)    0.982         0.034                28.910    0.000    0.911    1.052
  ConvSFlag    0.237         0.085        0.503    2.788    0.010    0.061    0.412
-----------------------------------------------------------------------------------------

> car::vif(Model2$model)
Error in vif.default(Model2$model) : model contains fewer than 2 terms
```

Appendix M: Tuesday Scaling Factor Land Use Model

```
> Model2<-ols_step_both_p(lm(Tue~SchoolFlag+Residential+Park_Rec+ConvSFlag+RetailFlag,data=Pima_Reg_In),pent=0.05,prem=0.2, progress= TRUE)
Stepwise Selection Method
---------------------------

Candidate Terms:

1. SchoolFlag
2. Residential
3. Park_Rec
4. ConvSFlag
5. RetailFlag

We are selecting variables based on p value...

Variables Entered/Removed:

+ SchoolFlag

No more variables to be added/removed.


Final Model Output
------------------

                        Model Summary
-------------------------------------------------------------
R                       0.553       RMSE                0.135
R-Squared               0.305       Coef. Var          13.688
Adj. R-Squared          0.275       MSE                 0.018
Pred R-Squared          0.197       MAE                 0.100
-------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------
               Sum of
              Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression      0.183         1          0.183    10.106    0.0042
Residual        0.416        23          0.018
Total           0.599        24
--------------------------------------------------------------------

                                 Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     1.031         0.031                33.403    0.000     0.967     1.095
 SchoolFlag    -0.200         0.063       -0.553   -3.179    0.004    -0.331    -0.070
-----------------------------------------------------------------------------------------

> car::vif(Model2$model)
Error in vif.default(Model2$model) : model contains fewer than 2 terms
```

Appendix N: T-Test Results by Land Use Factor Grouping Case

Table N.1. Land Use T-Test Results for MAPE Values

Case	WWI Only	AMI Only	AMI and WWI	K-means (n=3) Mod (3V)	K-means (n=3) Mod (7V)	K-means (n=2) Mod (3V)
K-means (n=2) Mod (7V)	**0.016**	**0.002**	**0.017**	**0.025**	0.147	**0.029**
K-means (n=2) Mod (3V)	**0.013**	**0.00003**	**0.014**	**0.024**	0.1	NA
K-means (n=3) Mod (7V)	**0.047**	**0.03**	**0.048**	0.053	NA	NA
K-means (n=3) Mod (3V)	0.125	0.241	0.097	NA	NA	NA
AMI and WWI	**0.048**	0.662	NA	NA	NA	NA
AMI Only	0.445	NA	NA	NA	NA	NA

Values in **bold** are statistically significant at 95% confidence level

Table N.2. Land Use T-Test Results for MAE Values

Case	WWI Only	AMI Only	AMI and WWI	K-means (n=3) Mod (3V)	K-means (n=3) Mod (7V)	K-means (n=2) Mod (3V)
K-means (n=2) Mod (7V)	**0.011**	**0.004**	**0.011**	**0.007**	0.158	**0.03**
K-means (n=2) Mod (3V)	**0.008**	**0.001**	**0.009**	**0.0004**	0.102	NA
K-means (n=3) Mod (7V)	**0.032**	**0.019**	**0.031**	0.053	NA	NA
K-means (n=3) Mod (3V)	**0.015**	**0.002**	**0.015**	NA	NA	NA
AMI and WWI	**0.019**	0.392	NA	NA	NA	NA
AMI Only	0.138	NA	NA	NA	NA	NA

Values in **bold** are statistically significant at 95% confidence level

Appendix O: Land Use SADPT Estimation Metrics for Sites in Pima County, Arizona

Table O.1. MAPE Traditional (Land Use Predicted)

Site Code	K-means (n=2) Mod (7V)	K-means (n=2) Mod (3V)	K-means (n=3) Mod (7V)	K-means (n=3) Mod (3V)	AMI and WWI	AMI Only	WWI Only	Direct Est	Direct Est HSG Sep
S001	39.7	39.7	40.6	25.8	41.2	41.0	41.0	40.6	40.6
S004	29.9	29.9	28.0	42.2	33.6	30.3	37.2	31.0	31.0
S018	29.7	29.7	25.5	26.6	28.5	30.7	33.8	24.7	24.7
S021	54.9	54.9	54.9	54.9	113.4	140.9	108.3	131.0	54.9
S023	21.4	21.4	16.0	57.9	31.2	25.8	32.3	24.3	24.3
S024	39.3	73.1	34.4	30.7	28.9	33.3	28.6	42.1	42.1
S031	65.2	65.2	63.3	27.3	64.5	66.0	66.3	65.3	65.3
S041	25.8	25.8	25.1	73.1	29.6	28.0	31.6	27.1	27.1
S045	42.2	42.2	38.5	39.7	45.5	45.8	48.4	44.6	44.6
S050	20.2	20.2	20.3	42.8	20.6	20.1	22.0	21.2	21.2
S055	39.1	39.1	42.6	46.4	36.1	37.1	36.0	37.8	37.8
S056	42.8	42.8	40.0	29.9	46.4	46.4	48.1	44.0	44.0
S062	46.4	46.4	37.8	20.2	56.9	54.4	60.2	50.0	50.0
S065	56.6	56.6	55.1	24.5	55.7	56.1	57.2	57.6	57.6
S072	24.5	24.5	26.2	39.1	29.2	27.1	28.4	26.1	26.1
S078	43.9	43.9	33.6	56.6	60.3	49.4	59.2	47.1	47.1
S079	66.3	66.3	61.4	43.9	65.0	68.1	67.3	67.7	67.7
S082	26.6	26.6	18.4	43.3	40.8	32.9	44.7	31.9	31.9
S083	109.9	109.9	109.9	109.9	180.9	181.9	170.8	198.0	110.0
S084	30.7	30.7	25.3	66.3	43.9	33.1	41.4	33.4	33.4
S090	43.3	43.3	36.1	29.7	53.4	51.1	56.5	46.8	46.8
S093	37.9	37.9	33.6	21.4	48.4	40.1	46.6	40.7	40.7
S098	34.4	57.9	23.7	37.9	21.5	24.8	21.0	27.5	27.5
S101	26.8	27.3	31.3	65.2	35.2	31.4	35.4	29.1	29.1
S107	39.4	39.4	39.4	39.4	92.7	116.8	88.0	110.1	39.3

Table O.2. MAPE AASHTO (Land Use Predicted)

Site Code	K-means (n=2) Mod (7V)	K-means (n=2) Mod (3V)	K-means (n=3) Mod (7V)	K-means (n=3) Mod (3V)	AMI and WWI	AMI Only	WWI Only
S001	39.8	39.8	40.7	28.0	40.3	40.3	39.3
S004	32.5	32.5	28.0	42.2	37.8	34.5	42.4
S018	29.5	29.5	26.0	31.0	27.4	32.1	35.8
S021	56.4	56.4	56.4	56.4	117.1	147.7	109.2
S023	23.2	23.2	17.2	66.8	33.8	27.8	37.1
S024	40.5	80.5	35.6	33.7	31.1	36.3	30.9
S031	65.4	65.4	63.3	29.8	62.8	67.4	67.2
S041	28.0	28.0	25.5	80.5	30.6	28.9	34.2
S045	42.2	42.2	38.9	39.8	45.1	45.9	49.9
S050	19.0	19.0	20.3	42.5	19.9	20.4	21.8
S055	38.4	38.4	42.7	51.9	39.5	36.7	37.7
S056	42.5	42.5	41.1	32.5	45.1	45.3	44.7
S062	51.9	51.9	38.5	19.0	63.2	59.3	69.1
S065	57.6	57.6	54.4	23.7	54.2	57.1	57.8
S072	23.7	23.7	27.1	38.4	30.3	25.9	29.2
S078	48.5	48.5	34.0	57.6	64.2	53.4	65.5
S079	63.8	63.8	63.3	48.5	63.6	65.7	63.7
S082	31.0	31.0	18.9	48.4	44.2	36.1	50.5
S083	109.5	109.5	109.5	109.5	184.7	186.7	172.2
S084	33.7	33.7	24.8	63.8	48.2	37.1	45.5
S090	48.4	48.4	36.6	29.5	57.9	54.7	62.5
S093	39.8	39.8	35.3	23.2	51.1	42.1	49.5
S098	37.8	66.8	25.1	39.8	24.4	28.7	22.8
S101	26.3	29.8	31.3	65.4	33.9	30.7	35.0
S107	42.8	42.8	42.8	42.8	96.5	122.9	89.4

Table O.3. MAPE Disaggregate (Land Use Predicted)

Site Code	K-means (n=2) Mod (7V)	K-means (n=2) Mod (3V)	K-means (n=3) Mod (7V)	K-means (n=3) Mod (3V)	AMI and WWI	AMI Only	WWI Only
S001	40.3	40.3	38.7	35.5	42.5	40.2	43.6
S004	39.0	39.0	26.0	45.2	45.6	42.9	55.8
S018	38.0	38.0	30.4	50.6	41.5	42.5	48.0
S021	28.4	28.4	28.4	28.4	110.3	144.6	98.2
S023	42.9	42.9	17.6	91.9	57.2	43.5	60.1
S024	53.3	111.6	31.4	39.4	24.6	29.4	21.6
S031	63.0	63.0	61.6	36.0	62.5	62.5	66.2
S041	35.5	35.5	26.2	111.6	42.7	36.6	45.7
S045	45.2	45.2	38.2	40.3	51.8	49.3	58.2
S050	25.4	25.4	22.2	47.7	32.1	29.3	37.4
S055	33.5	33.5	42.8	68.3	41.3	36.4	44.1
S056	47.7	47.7	38.0	39.0	50.6	45.5	52.1
S062	68.3	68.3	36.0	25.4	83.2	72.8	92.3
S065	58.8	58.8	53.3	30.3	55.3	57.4	56.4
S072	30.3	30.3	24.1	33.5	39.8	33.0	48.2
S078	70.7	70.7	35.3	58.8	91.8	71.5	92.4
S079	64.6	64.6	61.6	70.7	67.9	70.2	63.6
S082	50.6	50.6	21.5	64.9	68.6	50.6	78.4
S083	80.7	80.7	80.7	80.7	178.9	176.0	161.3
S084	39.4	39.4	25.6	64.6	61.1	48.8	52.9
S090	64.9	64.9	36.5	38.0	76.8	64.7	82.1
S093	45.9	45.9	34.4	42.9	62.3	48.6	56.4
S098	56.1	91.9	28.1	45.9	35.1	35.5	31.8
S101	25.9	36.0	31.3	63.0	41.5	37.0	40.4
S107	14.3	14.3	14.3	14.3	90.2	119.6	79.6

Table O.4. MAE Traditional (Land Use Predicted)

Site Code	K-means (n=2) Mod (7V)	K-means (n=2) Mod (3V)	K-means (n=3) Mod (7V)	K-means (n=3) Mod (3V)	AMI and WWI	AMI Only	WWI Only	Direct Est	Direct Est HSG Sep
S001	3.0	3.0	3.0	9.0	3.1	3.1	3.1	3.1	3.1
S004	5.8	5.8	5.4	2.7	6.5	5.8	7.2	6.0	6.0
S018	16.6	16.6	14.3	56.0	15.9	17.2	18.9	13.8	13.8
S021	31.2	31.2	31.2	31.2	64.5	80.1	61.6	74.5	31.2
S023	19.2	19.2	14.4	18.1	28.0	23.2	29.0	21.8	21.8
S024	45.7	84.9	40.0	9.2	33.5	38.7	33.2	48.9	48.9
S031	9.4	9.4	9.1	14.9	9.3	9.5	9.5	9.4	9.4
S041	9.0	9.0	8.8	84.9	10.3	9.8	11.0	9.5	9.5
S045	2.7	2.7	2.5	3.0	2.9	2.9	3.1	2.8	2.8
S050	13.8	13.8	13.8	15.4	14.0	13.7	14.9	14.4	14.4
S055	5.8	5.8	6.3	16.2	5.3	5.5	5.3	5.6	5.6
S056	15.4	15.4	14.3	5.8	16.6	16.7	17.2	15.8	15.8
S062	16.2	16.2	13.2	13.8	19.9	19.0	21.1	17.5	17.5
S065	5.3	5.3	5.1	19.1	5.2	5.2	5.3	5.3	5.3
S072	19.1	19.1	20.5	5.8	22.8	21.2	22.1	20.4	20.4
S078	45.5	45.5	34.9	5.3	62.5	51.2	61.3	48.9	48.9
S079	4.2	4.2	3.9	45.5	4.1	4.3	4.3	4.3	4.3
S082	56.0	56.0	38.8	7.2	86.0	69.3	94.2	67.3	67.3
S083	190.5	190.5	190.5	190.5	313.6	315.3	296.2	343.3	190.7
S084	9.2	9.2	7.6	4.2	13.1	9.9	12.4	10.0	10.0
S090	7.2	7.2	6.0	16.6	8.9	8.5	9.5	7.8	7.8
S093	8.7	8.7	7.7	19.2	11.1	9.2	10.7	9.4	9.4
S098	10.8	18.1	7.4	8.7	6.7	7.8	6.6	8.6	8.6
S101	14.7	14.9	17.1	9.4	19.2	17.1	19.3	15.9	15.9
S107	83.0	83.0	83.0	83.0	195.4	246.2	185.4	232.2	82.9

Table O.5. MAE AASHTO (Land Use Predicted)

Site Code	K-means (n=2) Mod (7V)	K-means (n=2) Mod (3V)	K-means (n=3) Mod (7V)	K-means (n=3) Mod (3V)	AMI and WWI	AMI Only	WWI Only
S001	3.0	3.0	3.1	9.8	3.0	3.0	3.0
S004	6.3	6.3	5.4	2.7	7.3	6.6	8.2
S018	16.5	16.5	14.6	65.2	15.3	18.0	20.0
S021	32.1	32.1	32.1	32.1	66.6	84.0	62.1
S023	20.9	20.9	15.5	20.9	30.4	25.0	33.3
S024	47.0	93.4	41.4	10.1	36.1	42.2	35.9
S031	9.4	9.4	9.1	16.3	9.0	9.7	9.7
S041	9.8	9.8	8.9	93.4	10.7	10.1	11.9
S045	2.7	2.7	2.5	3.0	2.9	2.9	3.2
S050	12.9	12.9	13.8	15.2	13.5	13.8	14.8
S055	5.7	5.7	6.3	18.1	5.8	5.4	5.6
S056	15.2	15.2	14.7	6.3	16.2	16.2	16.0
S062	18.1	18.1	13.5	12.9	22.1	20.7	24.2
S065	5.3	5.3	5.1	18.5	5.0	5.3	5.4
S072	18.5	18.5	21.2	5.7	23.7	20.3	22.8
S078	50.3	50.3	35.3	5.3	66.6	55.4	67.9
S079	4.1	4.1	4.0	50.3	4.0	4.2	4.0
S082	65.2	65.2	39.7	8.1	93.0	76.0	106.4
S083	189.8	189.8	189.8	189.8	320.3	323.7	298.6
S084	10.1	10.1	7.4	4.1	14.4	11.1	13.6
S090	8.1	8.1	6.1	16.5	9.7	9.2	10.5
S093	9.2	9.2	8.1	20.9	11.8	9.7	11.4
S098	11.8	20.9	7.8	9.2	7.6	9.0	7.1
S101	14.4	16.3	17.1	9.4	18.5	16.7	19.1
S107	90.2	90.2	90.2	90.2	203.3	259.2	188.5

Table O.6. MAE Disaggregate (Land Use Predicted)

Site Code	K-means (n=2) Mod (7V)	K-means (n=2) Mod (3V)	K-means (n=3) Mod (7V)	K-means (n=3) Mod (3V)	AMI and WWI	AMI Only	WWI Only
S001	3.0	3.0	2.9	12.4	3.2	3.0	3.3
S004	7.5	7.5	5.0	2.9	8.8	8.3	10.7
S018	21.3	21.3	17.0	106.5	23.2	23.8	26.9
S021	16.1	16.1	16.1	16.1	62.7	82.2	55.9
S023	38.5	38.5	15.8	28.7	51.3	39.1	54.0
S024	61.9	129.5	36.5	11.8	28.6	34.1	25.0
S031	9.1	9.1	8.9	19.7	9.0	9.0	9.5
S041	12.4	12.4	9.1	129.5	14.9	12.8	15.9
S045	2.9	2.9	2.4	3.0	3.3	3.2	3.7
S050	17.3	17.3	15.1	17.1	21.8	19.9	25.4
S055	5.0	5.0	6.3	23.9	6.1	5.4	6.5
S056	17.1	17.1	13.6	7.5	18.1	16.3	18.7
S062	23.9	23.9	12.6	17.3	29.1	25.5	32.3
S065	5.5	5.5	4.9	23.7	5.1	5.3	5.2
S072	23.7	23.7	18.8	5.0	31.1	25.8	37.7
S078	73.3	73.3	36.6	5.5	95.1	74.1	95.8
S079	4.1	4.1	3.9	73.3	4.3	4.5	4.0
S082	106.5	106.5	45.3	10.9	144.4	106.6	165.1
S083	139.9	139.9	139.9	139.9	310.2	305.2	279.6
S084	11.8	11.8	7.6	4.1	18.2	14.6	15.8
S090	10.9	10.9	6.1	21.3	12.8	10.8	13.7
S093	10.6	10.6	7.9	38.5	14.3	11.2	13.0
S098	17.5	28.7	8.8	10.6	11.0	11.1	9.9
S101	14.1	19.7	17.1	9.1	22.6	20.2	22.0
S107	30.0	30.0	30.0	30.0	190.1	252.1	167.9