7-1-2022

Enhanced Methods for Utilization of Data to Support Multi-Scenario Analysis and Multi-Resolution Modeling

Syed Ahnaf Morshed
smors005@fiu.edu

Recommended Citation

Morshed, Syed Ahnaf, "Enhanced Methods for Utilization of Data to Support Multi-Scenario Analysis and Multi-Resolution Modeling" (2022). *FIU Electronic Theses and Dissertations*. 5100.
https://digitalcommons.fiu.edu/etd/5100

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

ENHANCED METHODS FOR UTILIZATION OF DATA TO SUPPORT MULTI-SCENARIO ANALYSIS AND MULTI-RESOLUTION MODELING

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CIVIL ENGINEERING

by

Syed Ahnaf Morshed

2022

To: Dean John L. Volakis
College of Engineering and Computing

This dissertation, written by Syed Ahnaf Morshed, and entitled Enhanced Methods for Utilization of Data to Support Multi-Scenario Analysis and Multi-Resolution Modeling, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Albert Gan

B M Golam Kibria

Xia Jin

Priyanka Alluri

Mohammed Hadi, Major Professor

Date of Defense: July 1, 2022

The dissertation of Syed Ahnaf Morshed is approved.

Dean John L. Volakis
College of Engineering and Computing

Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2022

DEDICATION

I dedicate this dissertation to my beloved parents, Eti and Morshed, my lovely sisters, Nawmee and Nisha, and my late uncle, Dr. Mustafizur Rahman, for their unconditional love, endless support, and encouragement.

ACKNOWLEDGMENT

First and foremost, all of my praise is to the Almighty Allah (SWT)!

I would like to strongly thank my advisor, Dr. Mohammed Hadi, for his guidance, patience, and for his confidence in me. It was my honor to work with such a well-respected advisor in the transportation domain. He was generous with the time he spent advising and training me on conducting high quality research, and on developing a variety of skills necessary to succeed in my professional career. He also played a major role in helping me become better overall. I wish I could tell him how influential he has been in my life.

I experienced a great teamwork environment in the Integrated Intelligent Transportation Systems (IITS) research lab. I would like to extend my deep appreciation to Kamar Amine, who supported me through and through. I am also grateful to my committee members, Dr. Albert Gan, Dr. Xia Jin, Dr. Priyanka Alluri, and Dr. Golam Kibria, for serving on my committee, and for their valuable comments, suggestions, and interest in my research. I will be always thankful to my family especially, my aunt, Dr. Nurun Nahar, and my cousin, Dr. Nahid Rahman, for their continued support, care, and encouragement.

ABSTRACT OF THE DISSERTATION

ENHANCED METHODS FOR UTILIZATION OF DATA TO SUPPORT MULTI-SCENARIO ANALYSIS AND MULTI-RESOLUTION MODELING

by

Syed Ahnaf Morshed

Florida International University, 2022

Miami, Florida

Professor Mohammed Hadi, Major Professor

The success of analysis and simulation in transportation systems depends on the availability, quality, reliability, and consistency of real-world data and the methods for utilizing the data. Additional data and data requirements are needed to support advanced analysis and simulation strategies such as multi-resolution modeling (MRM) and multi-scenario analysis.

This study has developed, demonstrated, and assessed a systematic approach for the use of data to support MRM and multi-scenario analysis. First, the study developed and examined approaches for selecting one or more representative days for the analysis, considering the variability in travel conditions throughout the year based on cluster analysis. Second, this study developed and analyzed methods for using crowdsourced data

to estimate origin-destination demands and link-level volumes for use as part of an MRM with consideration of the modeling scenario(s).

The assessment of the methods to select the representative day(s) utilizes statistical measures, in addition to measures and visualization techniques that are specific to traffic operations. The results of the assessment indicate that the utilization of the K-means clustering algorithm with four clusters and spatio-temporal segregation of the variables demonstrated superior performance over other tested approaches, such as the use of the Gaussian Mixture clustering algorithm and the use of different segregation levels.

The study assessed methods for the use of third-party crowdsourced data from StreetLight (SL) as part of the Origin-Destination Matrix Estimation (ODME), which identifies the method resulting in the closest origin-destination demands to the original seed matrices and real-world link counts. The results of the study indicate that Method 3(b) produced the best performance, which utilized combined data from demand forecasting models, crowdsourced data, and traffic counts. Additionally, this study examined regression models between crowdsourced data and count station data developed for link-level estimation of the volumes. This study also examined the accuracy and transferability of the link-level estimation of the volumes to determine if the crowdsourced data combined with available volume data at several locations can be used to predict missing or unavailable volumes in different locations on different days and times within the network. Regression models produced low errors than the default SL estimates when hourly or daily traffic volumes were taken into account. For similar traffic conditions, the models predicted directional traffic volume close to the real-world value.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

ABBREVATIONS AND ACRONYMS

ADT	Average Daily Traffic
AADT	Annual Average Daily Traffic
AMS	Analysis, Modeling and Simulation
ANN	Artificial Neural Network
APE	Absolute Percentage Error
ATR	Automatic Traffic Recorders
AVI	Automatic Vehicle Identification
BIC	Bayesian Information Criterion
CH	Calinski-Harabasz
DB	Davies-Bouldin Index
EB	Eastbound
EBL	Eastbound Left
EBR	Eastbound Right
EBT	Eastbound Through
FDOT	Florida Department of Transportation
FHWA	Federal Highway Administration
FL	Florida
GB	Gradient Boosting
GCN	Graph Convolutional Networks
GIS	Geographic Information System
GMM	Gaussian Mixture Model

GPS	Global Positioning Services
KNN	K-Nearest Neighbors
LBS	Location-Based Service
LSTM	Long Short-Term Memory
MAD	Mean Absolute Deviation
MADT	Monthly Average Daily Traffic
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ME	Mean Error
MnDOT	Minnesota Department of Transportation
MOE	Measures of Effectiveness
MPH	Miles per Hour
MPE	Mean Percentage Error
MRM	Multi-Resolution Modeling
MSD	Mean Signed Difference
NB	Northbound
NBL	Northbound Left
NBR	Northbound Right
NBT	Northbound Through
NN	Neural Network
O-D	Origin-Destination
ODME	Origin-Destination Matrix Estimation

PE	Percentage Error
PCA	Principal Component Analysis
PCS	Permanent Count Station
PTR	Permanent Traffic Recorders
RF	Random Forest
RSME	Root Mean Squared Error
RTMC	Regional Transportation Management Centers
SERPM	Southeast Florida Regional Planning Model
SB	Southbound
SBL	Southbound Left
SBR	Southbound Right
SBT	Southbound Through
SC	Silhouette Coefficient
SF	Seasonal Factors
SL	StreetLight
SVM	Support Vector Machine
TAZ	Traffic Analysis Zone
t-SNE	t-Distributed Stochastic Neighbor Embedding
TSM&O	Transportation Systems Management and Operations
TT	Travel Time
USDOT	United States Department of Transportation
VDOT	Virginia Department of Transportation

VMT	Vehicles Miles Traveled
VPD	Vehicles per Day
VPH	Vehicles per Hour
WB	Westbound
WBL	Westbound Left
WBR	Westbound Right
WBT	Westbound Through
WCSS	Within Clusters Sum of Squares
WIM	Weight in Motion

CHAPTER 1

INTRODUCTION

1.1. Background

Transportation system analysts have used analysis, modeling, and simulation (AMS) tools of different types and resolutions to assess the system performance under different conditions and alternative improvements. The use of AMS in transportation decision-making systems provides a detailed analysis of alternatives and a thorough refinement of strategies and plans. Such a process is beneficial for assessing traffic performance based on the various measures of effectiveness (MOE) of transportation systems, as it provides detailed levels of assessment. Despite requiring a time-consuming rendition and recourse-intensive activity period, transportation engineers and planning practitioners often use traffic simulation models for detailed temporal and spatial analysis of traffic to assess different MOEs of the transportation system and network. However, it is critical that analysts or practitioners scope, develop, and calibrate the models to existing travel conditions and validate them to emulate real-life scenarios within the transportation network. Deficiencies in scoping, development, calibration, and validation of the models risk the accuracy of the modeling results and the decision-making process. Therefore, accurately assessing the performance of any transportation system effectively requires checking the data and model quality.

AMS tools are essential for evaluating potential solutions for different transportation decision-making processes. Based on the level of details, traffic simulation models are categorized into three different resolution levels: macroscopic, microscopic, and mesoscopic. Macroscopic models use macroscopic traffic models, parameters, and measures such as traffic flow, average speed, and density to model traffic operations. Microscopic models simulate individual vehicle dynamics and driver behavior and utilize microscopic parameters such as car-following, lane changing, and gap acceptance (Spiegelman et al., 2010). Although macroscopic models render faster processing time and are easier to develop and calibrate, microscopic simulation models tend to be more accurate and precise if well calibrated and validated. On the other hand, mesoscopic models combine the properties of both microscopic and macroscopic models. In most cases, they use aggregated speed-density (macroscopic) relationships but depict the motions of individual vehicles (microscopic) (De Palma et al., 2002), although some mesoscopic models use low fidelity car-following and lane selection models. Mesoscopic simulation models provide less fidelity than microsimulation tools but are more computationally efficient, making them ideal to use in the iterative process required for dynamic traffic assignment, particularly for larger networks. Depending on network size and the types of analyses required, all types of models are potentially valuable for transportation analysis. A multi-resolution modeling (MRM) framework that combines different modeling resolutions has found significant interest in recent years to answer questions related to traffic operations and advanced strategies. The term “resolution” in AMS refers to the degree of detail and precision in the representation of real-world conditions in a model (Army Modeling and Simulation Office, 2020). MRM in traffic simulation is an integrated framework that combines microscopic, mesoscopic, and macroscopic levels of traffic flow modeling to provide a unified analysis for different scenarios.

Another important concept that has been considered in recent years is multi-scenario modeling that considers the variations in travel conditions due to differences in demand, capacity, weather, and incidents. These factors are not usually accounted for in modeling practice. Many improvement alternatives have different impacts depending on traffic conditions. With that, creating scenario-based models by identifying and labeling days of the year based on travel conditions can provide a better understanding of the impacts of these alternatives, even if only one scenario is selected for the analysis. However, there are constraints associated with scenario-based modeling, specifically with the required extra costs and time.

An important constraint that limits this type of modeling is the unavailability of a complete data set for different days of the year, preventing the identification of scenarios and representative days. Thus, research is needed for the development of methods for the cost-effective support for multi-scenario modeling, or at least for the selection of a representative day in the year for the modeling. The Traffic Analysis Toolbox Volume 3 developed by the Federal Highway Administration (FHWA) (Wunderlich et al., 2019) recommends the use of multi-scenario analysis instead of the current practices of using one scenario. Accordingly, it also suggests clustering the data to produce scenarios that represent operating conditions seen in the real world. The emergent crowdsourced data from social media, smartphones, and devices (Morshed et al. 2021) installed in vehicles can be combined with other available data to address problems related to missing and extensive data needs in multi-scenario modeling and MRM. Automated data collection via crowdsourcing created a paradigm shift in the urban mobility domain. Important information that can be obtained based on the crowdsourced data includes network-wide traffic information such as travel times, origin-destination (O-D) demand matrices, and the routes used by the drivers between the O-D pairs.

1.2. Problem Statement

There is a wide range of simulation and modeling tools available in the market with different resolutions and sets of strengths. The use of these tools requires detailed data from multiple sources. With the availability of this data, multi-scenario analysis can be combined with multi-resolution analysis to provide a powerful transportation system analysis framework. This study research methods to address important knowledge gaps in multi-scenario modeling and MRM in the transportation domain considering the availability of data from multiple sources.

1.2.1. Identified Gaps in Multi-Resolution Modeling (MRM)

MRM frameworks possess varied theories and applications with different simulation granularity, representation (vehicle, cell, and flow), and data requirements (Zhou et al., 2022). MRM tools, when utilized, must provide better assessments of traffic conditions and an increased understanding of the performance of different alternatives other than single resolution models. For MRM, traffic count data from different sources are used in developing O-D matrices using algorithms referred to as O-D matrix estimation (ODME) algorithms. However, there are questions related to the quality of the resulting O-D matrices. In addition, there are often issues related to data unavailability and missing data within the study area. For freeways, there are data available from traffic sensors.

However, there is a general lack of data for arterial streets that hinders multi-scenario modeling and MRM. Accurate and more detailed volume data combined with the ODME can aid both multi-scenario and MRM analysis, respectively. In addition to traffic counts, O-D matrices based on crowdsourced data can be used to estimate the O-D matrices. The crowdsourced data can also be used to validate the path selection resulting from the traffic assignment in the MRM. Additionally, network-wide traffic flow information can work as a benchmark to help evaluate the performance of MRM.

Existing transportation AMS tools vary widely in their implementation and data requirements. Combining data from various sources is a potential approach to obtaining good quality O-D matrices. Sources of data for use in ODME can include automatic vehicle re-matching technologies and automatic vehicle location technologies that track or identify vehicles as they move between the origins and destinations (Zhang et al. 2020). An important source is crowdsourced data, which is generally based on automatic vehicle location technologies. In many cases, this data serve to collect partial trips that do not cover the full trip, which means that the data only identify trips between points located on the paths between the origins and the destinations. There is a need for methods for the best use of O-D matrices estimated using crowdsourced data or vehicle re-matching technologies as inputs to the ODME process. Existing research studies do not address combining crowdsourced data with count data and initial O-D matrices from other sources to estimate O-D matrix.

1.2.2. Identified Gap in Multi-Scenario Modeling

An important step in scoping traffic analysis projects is to process and use real-world data to identify the operational scenarios and the representative days for these scenarios for AMS model development. Most existing traffic analysis studies use one set of inputs to the utilized tool to represent one operational scenario for each analyzed peak period. In most cases, analysts take the averages of traffic volume, speed, travel time, and possibly other measures for the weekdays and use these averages in the development and calibration of the simulation models. In other cases, the analysts use the averages only from data collected from the peak months of the year rather than the whole year. Averaging measures based on data collected from different days of the year could result in a synthetic day that does not really occur in the real world (Wunderlich et al., 2019, Dowling et al., 2004).

The identification of traffic patterns that best represent the traffic conditions needed for an AMS effort is critical to the success of these projects. In some efforts, particularly those associated with traffic management and operations, the analysts should perform the analysis for different operational conditions of the year, including different recurrent congestion levels, incident conditions, and bad weather conditions (Vasudevan and Wunderlich, 2013). However, even if the scope of the analysts is only to analyze normal no-event conditions, there is a need to analyze the variations in traffic conditions to determine the best day(s) that represent the traffic conditions. In this regard, the FHWA Traffic Analysis Toolbox Volume 3 (Wunderlich et al., 2013) recommended the use of cluster analysis to categorize the traffic conditions in an entire year into a number of clusters and determine a representative day for each of these clusters. Such identification will allow the selection of the best operational conditions to represent the whole spectrum of traffic conditions present in the network for the purpose of analysis. However, the above-mentioned reference does not provide detailed guidance on how to perform clustering analysis to identify the travel conditions. There are many essential decisions to take into consideration when conducting the cluster analysis to obtain the representative day(s). This study identified three knowledge gaps requiring further research in this regard. First, there are various clustering algorithms and methods; each have their strengths and weaknesses. Therefore, it is important to identify the performance of these algorithms in clustering data obtained from typical traffic monitoring systems. Second, it is expected that higher resolutions of the input variables to the clustering analysis can improve the results from the analysis. For example, if the traffic volume used as input is averaged over the whole corridor of the study and for the whole analysis peak period, then important traffic patterns in time and space might be lost. However, how the segregation of the input variables in time and space can improve the quality of the clustering is yet to be understood. Third, it is also necessary to investigate methods needed for measurement and evaluation of the aggregation levels of the performance results from the clustering of traffic conditions.

1.3. Research Goal and Objectives

Considering the identified gaps as presented in the previous section, the goal of this project is to identify the methods needed to use data from multiple sources to support improved multi-scenario analysis and MRM. The specific objectives related to the goal of this study are listed below.

- Develop a method to identify traffic scenarios representing traffic patterns and representative days of these travel patterns considering the day-to-day variations in traffic conditions.
- Develop methods for integrating crowdsourced data into the ODME process required for the MRM for the modeled scenario(s).
- Check the reliability and transferability of the use of crowdsourced data obtained from a third-party vendor in estimating segment-level daily and hourly volumes in support of AMS and MRM.

1.4. Dissertation Organization

This section presents the organization of this document, including an overview of the chapters in this document.

Chapter 2 presents a synthesis and a detailed review of past research related to the objectives and tasks of this study. The literature review is categorized into three different parts, which consist of multi-scenario modeling, ODME, and crowdsourced data.

Chapter 3 focuses on investigating methods for traffic patterns and representative day selection for multi-scenario analysis. This chapter explores different clustering algorithms and their performance based on measures reported in statistical and traffic engineering literature.

Chapter 4 discusses the utilization of crowdsourced data in the ODME process. This chapter reports the results from using ODME methods in combination with crowdsourced data from a third-party vendor and regional demand forecasting model.

Chapter 5 includes an analysis of the accuracy of segment-level estimation of volumes using crowdsourced data from a third-party vendor.

Chapter 6 presents the conclusions based on the results from the research and recommendations for future studies.

CHAPTER 2

LITERATURE REVIEW

2.1. Literature Synthesis

The literature review is categorized into three major categories based on the research objectives. The categories of literature review are multi-scenario modeling, ODME, crowdsourced data. A chart showing the synthesis of the literature review is shown in **Figure 2-1.**

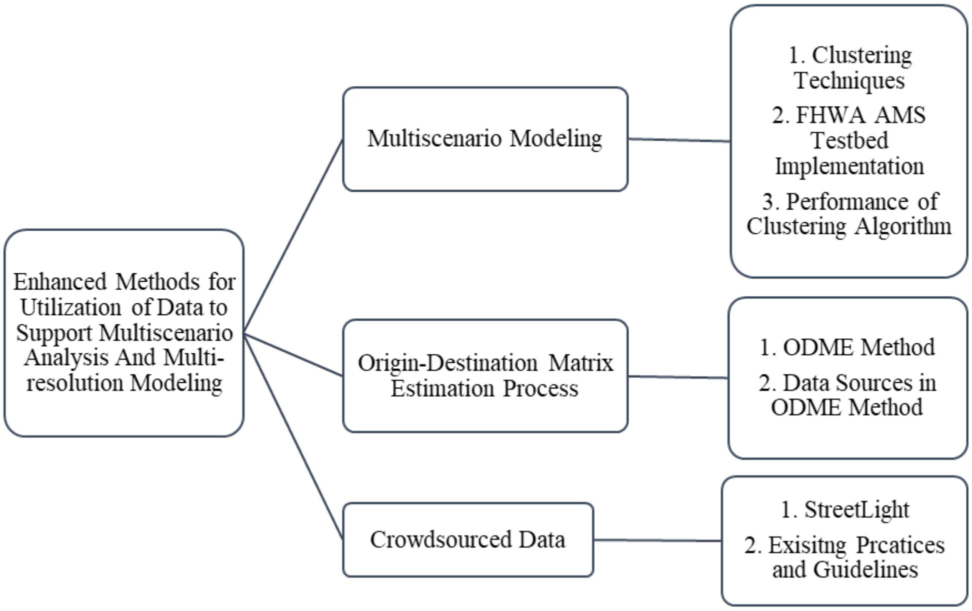

Enhanced Methods for Utilization of Data to Support Multiscenario Analysis And Multi-resolution Modeling
Multiscenario Modeling
Origin-Destination Matrix Estimation Process
Crowdsourced Data
1. Clustering Techniques
2. FHWA AMS Testbed Implementation
3. Performance of Clustering Algorithm
1. ODME Method
2. Data Sources in ODME Method
1. StreetLight
2. Exisitng Prcatices and Guidelines

Figure 2-1 Literature review synthesis

2.2. Multi-Scenario Modeling

The FHWA Traffic Analysis Toolbox Volume 3 (Wunderlich et al., 2019) presented a procedure for multi-scenario analysis, which involves the use of clustering analysis for the identification of modeling scenarios. Clustering analysis is characterized as an unsupervised machine-learning technique that segments the objects under consideration into clusters, with the objects within each cluster having closer relations compared with objects in other clusters (Hastie et al., 2009). Clustering methods mainly utilize a dissimilarity measure to cluster the objects (Xia and Chen, 2007b). Although the use of a clustering approach in the transportation engineering field has been limited due to the absence of data, the demand for using this approach in the recent past has gained popularity to identify multi-scenario prediction for traffic analysis, modeling, and simulation (AMS). The goal of performing data clustering is to form a set of objects into subsets so that the objects in the clusters formed are in close proximity compared to objects assigned to different clusters (He et al., 2011).

Traffic data are mostly spatial and temporal in nature and possess high correlations. The focal point of implementing a clustering technique is to determine spatial and temporal patterns among the traffic measures, such as traffic volume, speed, and travel time. Ku et al. (2016) proposed a data-driven model consisting of K-means clustering and a subsequent deep learning-based neural network model to predict the missing counts of traffic data from the road segments. The implementation of the K-means clustering has been used to categorize the traffic flow operation regimes based on traffic density and speed data aggregated in 15-minute intervals. Similarly, nested clustering techniques have been used to identify the operating scenarios of freeways (Xia and Chen, 2007a). The K-means and Fuzzy clustering were used to forecast traffic demand and analyze temporal and spatial travel behaviors using data collected by license plate recognition readers (Park, 2002). Other studies implemented clustering approaches such as Fuzzy C-means clustering and spectral clustering in congestion pattern recognition on urban roads and analyzing traffic state variations (Chen et al. 2017, Zhang et al. 2017). The K-means clustering based on historical data showed that travel time can be related to external factors such as weather conditions and traffic incidents (Nath et al., 2010; Chen et al., 2001; Wei et al., 2007; Wu et al., 2004, Al-Kaisy et al. 2022). Another study showed the possibility of implementing clustering techniques to compute the probabilistic distribution of travel time variability on urban arterials (Hans et al., 2014).

However, the most detailed and sophisticated implementation of clustering techniques in transportation engineering is the FHWA's AMS testbed effort (FHWA 2013a, FHWA 2013b, Vasudevan and Wunderlich, 2013). This effort involved six testbeds (San Mateo (US 101), Pasadena, Dallas, San Diego, Phoenix, and Chicago testbeds) that pilot-tested the use of AMS for assessing active traffic management and dynamic mobility applications. Table 2-1 summarizes the clustering analysis methods used in the six testbeds of the FHWA study mentioned above.

Table 2-1 Clustering analysis methods used in the six FHWA testbeds. (Source: FHWA 2013a, FHWA 2013b, Vasudevan and Wunderlich 2013)

Testbed Location	Clustering Method	Multi-scenario Parameter
Dallas	K-means clustering	Clustering of traffic data based on Vehicle Miles Traveled (VMT), travel time (TT), incident severity and precipitation
San Diego	K-means clustering	Incident duration, demand, travel time, and incident impact on delay in the clustering
Pasadena	K-means clustering	VMT, travel time, the total number of incidents, total duration of incidents
Phoenix	Hierarchical clustering & K-means clustering	Hourly travel speed, precipitation, incident frequency, traffic counts, and travel speed
Chicago	Two-step joint K-means clustering	Weather patterns identified based on the precipitation type (Rain, Snow, and Clear)
San Mateo	K-means clustering	Travel time, VMT, weather, and incident frequency

There is limited information related to selecting appropriate clustering technique(s), clustering parameters, and an optimal number of clusters. The K-means clustering method is a widely used technique that clusters and analyzes large datasets. However, its application is limited to datasets consisting of quantitative variables, which utilizes Euclidian distance as the dissimilarity matrix (Huang, 1998). Unlike similarity-based clustering techniques such as K-means clustering, model-based clustering is more flexible, which can eradicate problems like data uncertainty. The mixture distribution type of clustering techniques involves data fitting and implementation of the conditional probabilities of data points, which naturally assign probabilistic labels to the clustered data. One of the most widely used mixture models for clustering is the Gaussian Mixture Model

(Bishop 2007). Table 2-2 presents the difference between K-means clustering and GMM clustering.

Table 2-2 Comparisons between GMM and K-means clustering techniques (Wang et al. 2011; Huang, 1999; Bishop, 2007)

GMM Clustering Technique	K-means Clustering Technique
GMM accounts for covariance, which determines the shape of the distribution.	K-means uses simple distance-from-cluster-center to assign cluster membership, which leads to poor representation of the datasets.
GMM performs soft classification of data, i.e., flexible classification.	It performs hard classification of data i.e., rigid classification.
GMM can handle very oblong clusters.	Cluster models must be circular: K-means has no built-in way of accounting for oblong or elliptical clusters, i.e., if we take the same data and transform it, the cluster assignments end up becoming muddled.
GMM contains a probabilistic model that finds probabilistic cluster assignments.	The K-means model has no intrinsic measure of probability or uncertainty of cluster assignments.

The FHWA guideline (Wunderlich et al. 2019) included an example of a stepwise method for identifying the clusters using the K- means method, as follows:

1) Data Sorting: This step involves identifying a critical location in the network (e.g., bottleneck location) and organizing available recurring and non-recurring data.
2) Clustering Number: This step involves specifying an initial cluster number (e.g., K as 4).
3) Day Partitioning: In this section, there is a systematic division of days into each preliminary cluster.
4) Centroid Calculation: This step calculates the centroid or mean of each cluster based on the attribute of the clusters. Assumingly, the attributes are equally weighed.

5) Euclidean Distance Calculation: This step calculates the Euclidean distance of each data to the centroid of all clusters using the following equation:

$$xdis_y = \sqrt{(x_1 - ymean_1)^2 + \ (x_2 - ymean_2)^2 + \cdots + (x_a - ymean_a)^2} \quad (2\text{-}1)$$

where

$xdis_y$ = Distance of data x to the mean of cluster y;

x_a = Value corresponding to the attribute a for data x; and

$ymean_a$ = mean value for attribute a for cluster y.

6) Day Assignment: This step involves the re-assignment of each day to the cluster with the closest centroid.

7) Identifying Stopping Criterion: If there is no change in the day assignment step, then a stopping criterion takes place to identify the optimal number of clusters. Otherwise, the procedure is iterated starting at Step 4.

Several internal clustering validation metrics have been proposed in statistics literature, such as Silhouette Coefficient (SC), Calinski-Harabasz Index (CH), and Davies-Bouldin Index (DB), to measure the goodness-of-fit in the clustering methods (Liu et al., 2013). However, these measures have an issue in that they are dependent on the dimensionality of the data, meaning that for larger numbers of parameters, these measures assign datapoints as outliers even though the datapoints could be a part of a cluster (Platzer, 2013). Another method that does not have this issue is the t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization method. The t-SNE method allows visualizing the difference between different aggregation levels of the investigated dataset (Azimi and Zhang, 2010). The t-SNE is a non-linear dimensionality reduction technique that uses an unsupervised learning algorithm. It attempts to place similar data points close to each other while preserving the global structure of the dataset (Lowry, 2014). Researchers have also proposed methods for assessing the results from cluster analysis based on traffic engineering concepts. For example, Azimi and Zhang (2010) observed the distribution of the performance measures of different clusters by plotting speed versus flow (fundamental diagram) data to identify the performance of the clusters.

2.3. Origin-Destination Matrix Estimation Process

An important and challenging component of multi-resolution analysis is the estimation of origin-destination (O-D) demand matrices for the analysis. Recognizing that the O-D demand matrices produced by most existing travel demand models cannot provide the accuracy required in multi-resolution analysis, the analysts have used ODME processes to refine the O-D demand matrices. The ODME processes are mainly performed to better match real-world traffic volumes when assigning the O-D demands to the network, considering initial O-D matrices from existing sources such as demand forecasting models. Based on static or dynamic assignment of the O-D demands to the paths between the O-D pairs, the O-D matrices are subjected to optimization procedures by minimizing the differences between the link volumes resulting from the assignment and the measured values. Sources of traffic volume data generally include permanent and portable count stations installed by various planning agencies, road sensors installed by traffic management agencies, traffic count tubes, or a combination of these sources .

Many researchers have developed ODME models to estimate O-D matrices based on an initial seed matrix combined with measurements such as traffic counts. Table 2-3 summarizes examples of such methods developed to estimate the O-D demand.

Table 2-3 Developed methods to predict the O-D matrices

Author	Methodology
Lowry (2014)	The O-D centrality method is based on graph theory, which is used in this paper as a regression parameter directly in a demand model (linear regression) to estimate AADT spatially with the help of GIS.
Shi et al. (2020)	The methodology used in this paper was to predict O-D flows by implementing Long Short-Term Memory (LSTM) and multi-perspective Graph Convolutional Networks (GCN) processes.
Ashok (1996)	The authors proposed an offline (historical time-series data) and online estimation and prediction (proposed sequential model) of time dependent O-D flows.
Cascetta (1993)	This method developed a dynamic estimator using a (discrete) time varying traffic count on a general network to obtain time varying O-D flows or average O-D flows. The results showed consistent and significant estimates of true O-D flows over 15-min. intervals.
Kikuchi and Tanaka (2000)	The authors proposed a back-propagation artificial neural network (ANN) model that has been used to estimate a ramp-to-ramp O-D table.
Mussone and Matteucci (2013)	This paper performed O-D estimation using a multilayer feed forward neural network (NN) and principal component analysis (PCA). NN is robust even when data contains wrong information. This method also considered missing data.
Chang and Wu (1993)	This paper proposed a nonlinear dynamic system model with the Extended Kalman Filtering procedure. It also considered interrelations between O-D distributions and observed flows under congested conditions.
Zhou et al. (2003)	This research developed a dynamic O-D demand-based method. The objective function of the proposed model is to minimize the deviation between observed link flows and estimated link flows. Similarly, the method also worked on minimizing deviation between target demand and estimated demand volume.
Dixon and Rilett (2000)	The authors assumed that AVI readers are available at the boundaries of their network and used the data to estimate the O-D demand. This study evaluated the generalized least squares and Kalman Filter algorithms on a freeway section with on- and off-ramps and found that utilizing AVI has improved ODME estimation.

Some of the utilized techniques in the ODME include Bayesian Inference (Van Der Zijpp, 1997), Generalized Least Squares (Asakura et al., 2000) and Maximum Likelihood (Parry and Hazelton, 2012). Zhou and Mahmassani (2006) and Cantelmo et al. (2014) presented a customized ODME using the least-square estimation models based on traffic counts, point-to-point travel time collected from Automatic Vehicle Identification (AVI) data, and seed O-D matrices into a multi-objective optimization framework. Carrese et al. (2017) and Jamali (2014) proposed an ODME model based on travel time data from probe vehicles data and partial counts from AVI sensors, respectively.

Technologies such as AVI and probe data can provide partial and sample O-D demands that can be used as part of the O-D matrix estimation process. AVI data is used to validate and adjust the initial O-D matrices from the demand models. An example is the use of AVI data based on Bluetooth readers to verify and adjust the O-D patterns and trip lengths (Corradino Group, 2013). Antoniou et al. (2004) also incorporated AVI data into the ODME and found that the quality of the ODME results improved with the use of AVI data and link counts. Barceló et al. (2010) utilized simulated data to measure the reliability of AVI through Bluetooth and Wi-Fi data to estimate dynamic O-D matrices using the linear Kalman Filter approach. Alibabai and Mahmassani (2008) experimented with using turning movement counts as observation in dynamic ODME and found that intersection turning movement counts have considerable benefits in matching observed counts over link volumes.

It is evident from the literature review that there is a need for using supplemental data in ODME methods to improve the quality of the resulting matrices (Rodrigues et al., 2017). An important type of data that became available in recent years is the crowdsourced probe vehicle data provided by third-party vendors. However, there are no existing studies that investigate crowdsourced data with other data sources for use in the ODME process. Crowdsourced data are designed to collect partial trips that do not cover the full trip, which means that the data only identify trips between points located on the paths between the origins and destinations. In addition, only a small percentage of vehicles are tracked as part of crowdsourced data. As a result, vendors like Streetlight (SL), which is one of the most widely used providers of crowdsourced data, report a relative number of trips between the origins and the destinations rather than the actual number of trips. This relative number of trips is reported as an index, or an "SL Index," which can be expanded to estimate the O-D matrix. The expansion of the O-D matrix can be done by the user, although the vendor also provides estimates of the expanded O-D matrices. There are questions regarding the accuracy of the O-D matrices obtained from the third-party vendors since they cover a relatively small sample size that may not be representative of the full population. Therefore, combining data from various sources with the crowdsourced data is a potential approach to obtaining good quality O-D matrices.

2.4. Review of Crowdsourced Data Utilization

Crowdsourced data, such as online web services (Tostes et al. 2013), social media data (Ni et al. 2014 and Morshed et al. 2021), cellphones, and hotspots (Pereira et al. 2015, Demissie et al. 2016), are often used to estimate traffic volumes. In particular, crowdsourced data through mobile devices have the potential for use in estimating link volume and O-D demand matrices (Sanchez et al. 2014). Third-party vendors such as

StreetLight (SL) utilize GPS-based mobile devices or location-based services (LBS) data (i.e., data collected from devices installed in vehicles to estimate link traffic volumes, O-D matrices, and Annual Average Daily Traffic (AADT)), among other data. Other vendors like Wejo aggregate connected vehicle data obtained from automobile manufacturers. The aggregated data is used to estimate various parameters, including directional traffic volumes, average speeds, and O-D demands. In this study, SL data is used as a source of crowdsourced data. Thus, a detailed discussion of this data is presented in the following section.

2.4.1. Overview of Vendor-Provided Data

Crowdsourced data from mobile devices is challenging due to difficulties in detecting the locations of vehicles, which is caused by a variable location sampling rate. Additionally, errors in location tracing through mobile data can result in faulty transportation mode detection. The absence of ground truth data can also reduce the validation and trust in the collected data.

To tackle the challenges associated with crowdsourced data, sampling and filtering of the collected data is essential. For example, Rodrigues et al. (2017) utilized classification algorithms such as Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), and clustered algorithms such as weighted K-means to sample and filter the data. SL (StreetLight Insight 2020 Whitepaper Version 1), one of the main vendors of crowdsourced data, uses supervised machine learning models that are adopted to improve the estimation of different traffic data-related features such as O-D demand matrices, traffic volumes, and AADT.

Recently, various public agencies have become interested in the prospect of using traffic parameters estimated based on crowdsourced data vendors. SL, a popular crowdsourced data platform, provides estimates of the AADT and Monthly Average Daily Traffic (MADT). However, the SL estimates rely heavily on the data points sampled from smartphone applications and GPS devices, which may be subject to potential bias and coverage issues. To estimate the traffic flow parameters, SL uses machine-learning models that use mobile data, in combination with demographics data derived from census, permanent traffic recorders (PTR), and OpenStreetMap data, among other data sources. Prior to implementing the machine-learning model, a gradient-boosting model is used to estimate a two-month "reference period" average daily traffic (ADT), which is used instead of determining the AADT based on the whole year. The ADT is calculated for September and August since these two months were found to exhibit relatively high and consistent penetration rates nationwide. The vendor also claims that the two-month ADT reference model outperforms a full-year AADT reference model and helps reduce the time required to run an analysis based on volume. Afterwards, the estimated traffic volume is adjusted based on data collected from 1700 PTR sites across 20 states in the U.S. In specific locations, PTR data is used to create MADT metrics to assess the monthly variation of trip volumes. MADT obtained based on PTR data are used to fine-tune the monthly seasonal adjustment models. The system utilizes the k-fold technique, leaving out one state from a set of 20 and validating the estimated results of trip volume with the left-out state based on novel zones created within the network. Additionally, historic counts are also used over the years to fine-tune the MADT in the software. Later, the data is normalized along several different parameters to create a relative measure of the trips called the SL Index. The SL

Index is a normalized mobile parameter updated each month based on the ratio of the mobile sample available for a location to the total population (collected from the census block). In other words, in the calculation of the SL Index, the data are scaled based on the available SL devices and residents in those locations.

To create the O-D matrices, SL uses mobile navigation data that is updated every month. The trips are calibrated externally based on high-quality vehicle count sensors, surveys, and other externally validated sources. Ultimately, the trips are normalized to avoid sample bias within the data.

2.4.2. Experience with Crowdsourced Data

Several public agencies have explored and tried to validate the estimated traffic volumes obtained from SL. However, there are no clear guidelines on how to utilize the SL metrics and how to evaluate its performance in an accurate way. Table 2-4 shows various approaches taken by public agencies to assess the performance of SL metrics, along with the usability of the data for different applications.

Table 2-4 Existing guidelines on StreetLight data

Public Agency (DOT)	Purpose	Benchmark Data	Performance Metrics
Virginia	Evaluation of SL AADT, SL O-D Trips, SL traffic counts, turn counts and truck volumes at intersection (2020)	Traffic Count Database System (City of Virginia Beach 2019); O-D Trip based on Electronic Toll System (VDOT 2018)	Percentage Error (PE) & Absolute Percentage Error (APE)
Minnesota	Evaluation of SL AADT and Average Hourly Volume (2017)	MnDOT 69 permanent monitoring sites and 7837 short-duration count stations	MAPE, MAD, MSD
Minnesota	Evaluation of SL AADT (2020)	Permanent automatic traffic recorders (ATR), Permanent	MAPE, MPE, Mean error (ME)

		WIM, Permanent RTMC, Permanent Wavetronix	
Oregon	Evaluation of SL AADT (2019)	Automatic Traffic Recorders	PE & APE
Georgia	Calculate O-D Matrix Indices, Freight patterns (2019)	N/A	N/A
Ohio	Estimation of daily truck volume	N/A	N/A
USDOT	Bicycle safety analysis (2020)	Permanent Bicycle Count Station	MAPE, MAE

Traffic volume data are essential for traffic planning, especially for evaluating the ratio of volume and capacity, origin–destination (O-D) matrix estimation, and detecting bottleneck roads (Hobbs, 2016). Additionally, transportation agencies can utilize network-wide traffic flow information to undertake proactive traffic control policy in order to respond to traffic congestion (Zhu et al., 2010). Traditionally, traffic volume is measured by installing permanent road sensors, popularly known as automatic traffic recorders (ATRs), which collect counts over a period of a year. However, factors such as deployment cost, maintenance cost and insufficient coverage across the entire road network limit the availability of network-wide data to estimate traffic volume. The emergence of crowdsourced data created an opportunity to address this problem. Inclusion of big data from crowdsourced platforms such as SL is an innovative approach to strengthen the formulation of transportation decision-making strategies based on traffic volume estimation. The methodology and findings of SL volume estimation and validation reported by Virginia, Minnesota, and Oregon Department of Transportation (DOT) are discussed in the next section.

2.4.2.1. Virginia Department of Transportation (VDOT) Experience

A VDOT's study (Yang et al., 2020) aimed to formulate a guideline on using SL metrics by measuring its performance in different application contexts. The report assessed the quality of the SL metrics in six testing contexts covering the AADT, O-D trips, traffic flow on road links, turning movements at intersections, and truck traffic. The benchmark data sources were continuous count stations, toll transaction data, and the VDOT's internal traffic estimations. The analysis showed that the AADT estimates based on the SL index had relatively small absolute percentage errors compared to other outputs such as the O-D trips, traffic counts on roadway links, turning movement counts at intersections, and truck traffic volumes, which exhibited relatively higher and consistent errors. One important finding from the study was that lower volume levels estimated based on the SL metrics generated higher errors. Additionally, using multi-periods, such as multiple days, weeks, or months, rather than individual periods as the input for estimating traffic measures in SL, resulted in reduced errors, especially in low-volume traffic segments.

The VDOT report found a linear trend between SL Index Estimates of AADT and ground truth AADT, as shown in Figure 2-2. The Mean of Absolute Percentage Error (MAPE) measured for both 2017 and 2018 showed that for low volume segments (0-10,000 vph), the highest errors are 18.2% and 10.2%, respectively. After observing a linear trend between the benchmark AADT and SL Index estimates of AADT, a linear regression model was developed between the SL Index and hourly sensor data to estimate traffic volume. However, significant errors of up to 25% have been observed in the estimated traffic volume from the SL index. High errors were mostly observed in the road segments with a low volume of less than 500 vehicles per hour (vph). Figure 2-3 plots the fitted regression model, along with the resulting percentage errors.

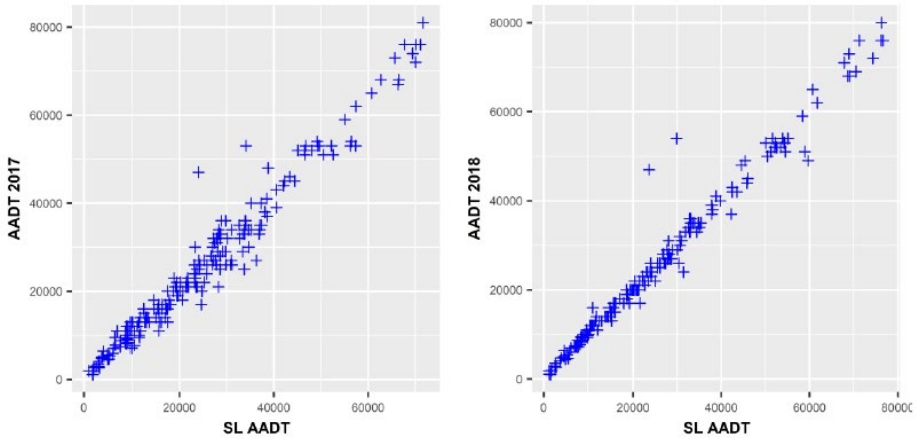


Figure 2-2 Relationship between the SL AADT Estimate and VDOT AADT in 2017 and 2018 (Yang et al., 2020)

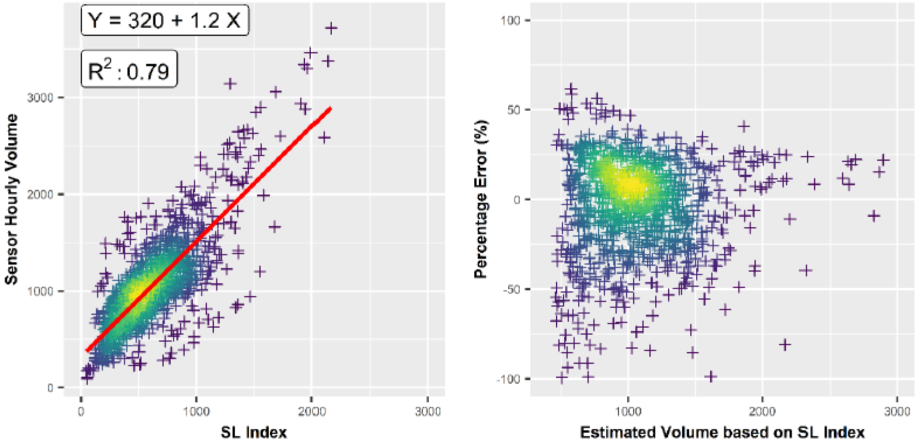


Figure 2-3 Linear regression based on SL Index and hourly sensor data (Yang et al., 2020), Minnesota Department of Transportation (MnDOT) Experience

2.4.2.2. Minnesota Department of Transportation (MnDOT) Experience

The MnDOT (Turner and Koeneman, 2017) evaluated the accuracy of SL AADT estimates, which were compared against volume data from 442 permanent continuous counter locations in 2017 and 2019. Additionally, the MnDOT conducted an analysis on several hundreds of low-volume sites for short-duration counts in 2019. Typically, short-duration count stations generate erroneous estimates of AADT. As a result, the MnDOT conducted the evaluation study to address the uncertainty of SL AADT estimates based on these two benchmark data sources (the permanent continuous counter locations and the short-duration count stations). The findings of the study also showed that the SL estimation of AADT improved significantly in 2019, compared to 2017, for moderate to high volume ranges (AADT greater than 10,000 vph). The MAPE decreased from approximately 42% to 10% for high volume locations, and 68% to 34% for low volume locations in the year 2019, compared to 2017. Figure 2-4 shows plots of SL estimates between the AADT and AADT benchmark from several data sources. The plots indicated that the SL estimates for high volume regions (volumes higher than 10,000 vph) underestimate the volumes compared to benchmark data but overestimate the volumes for the low volume regions.

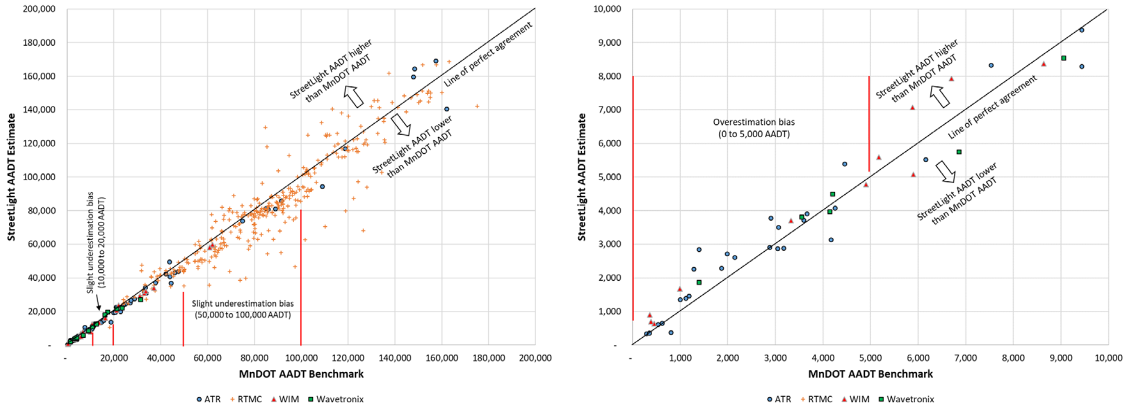


Figure 2-4 SL Estimates versus AADT Benchmark (Turner, 2020)

The MnDOT also explored the accuracy of SL estimates compared to short duration counts and found high errors ranging between 86% and 90% for AADTs less than 1,000 vph. It was evident that the SL estimates do not work well with annualized short-duration counts.

2.4.2.3. Oregon Department of Transportation (ODOT) Experience

The Oregon DOT (Roll, 2019) used AADT from the year 2017, which is based on Automatic Traffic Recorders (ATRs) data to measure the accuracy of the SL estimates of AADT. Out of the 180 ATRs in the region, 173 ATRs were used in the comparison. The results showed that the median and mean absolute percentage error is 18% and 26%, respectively. As in previous cases, low volume regions (between 0-1,000 vph) exhibited high discrepancies for SL volume estimation, as shown in Figure 2-5.

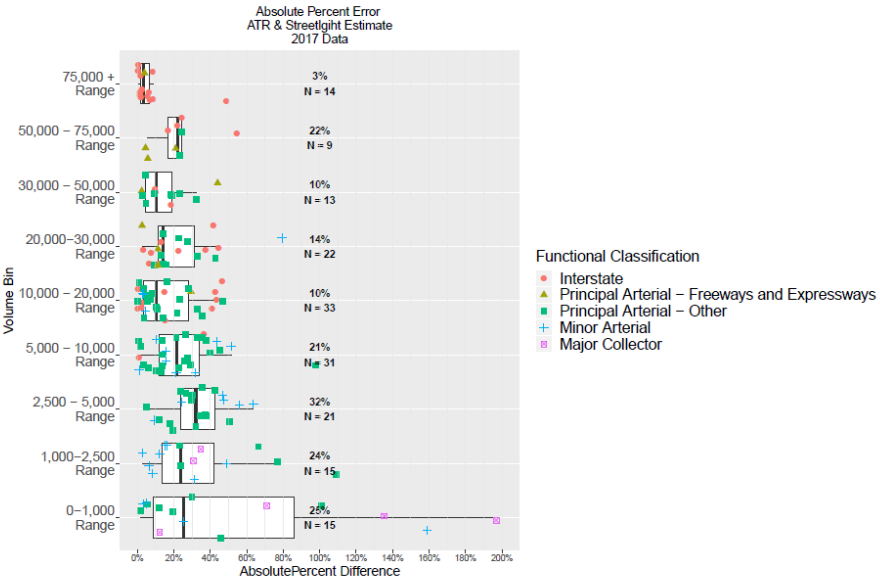


Figure 2-5 Absolute Error of the AADT between Benchmark Data and StreetLight Estimates (Roll, 2019)

2.4.3. Using Crowdsourced Data to estimate O-D Matrix

A real-world application of the SL Index to estimate the O-D matrix is presented in a recent VDOT report on Guidelines for Using SL Data for Planning Tasks (StreetLight White Paper, 2018). In this report, the O-D trips were derived by combining the SL Index and actual O-D trips collected from toll transaction data archived by the Electronic Toll System on the I-66 Expressway (Northern Virginia). However, the resulting O-D estimates, especially for O-D pairs with estimated hourly trips below 600, generated high errors (Absolute Percentage Error, APE) as large as 60% for the analyzed O-Ds. To reduce the estimation errors, SL indices were averaged across multiple hours of different days (5 or 15 days) for low volume conditions, as shown in Figure 2-6. The mean absolute percentage error (MAPE) for 5-day and 15-day aggregation were 62.9% and 58.08%, respectively.

Depending on project purposes, the aggregation can be based on the metrics of multiple days, weeks, or months.

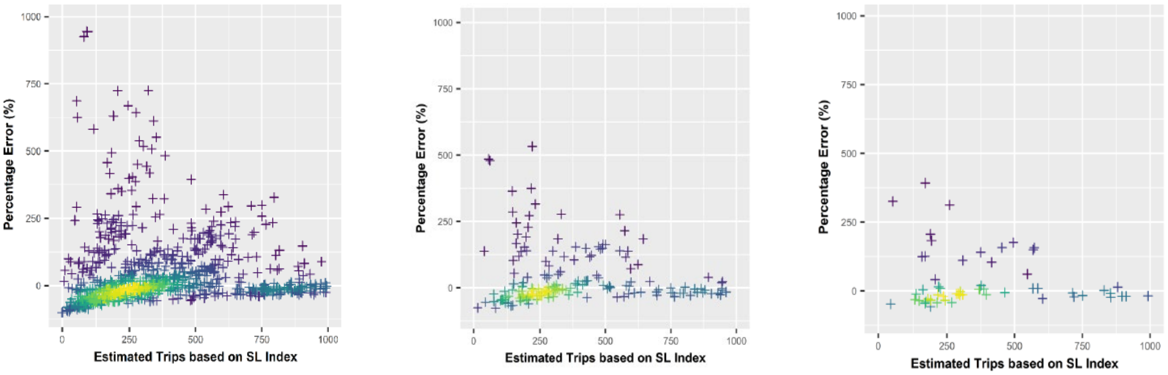


Figure 2-6 Distribution of percentage error in estimated trips based on SL Index for (a) Original SL index; (b) 5-day average SL Index; (c) 15-day SL Index, (Yang et al., 2020)

2.5. Summary

As mentioned earlier, the literature review is synthesized into three major sections. It is evident from the literature that there is a need for data enhancement from multiple sources such as detector data and crowdsourced data to develop methods in order to conduct efficient and accurate multi-scenario analysis and MRM.

There is no clear guidance in the FHWA Traffic Analysis Toolbox Volume 3 (Wunderlich et al., 2019) on how to select the optimal cluster numbers, clustering technique, clustering parameters, measuring clustering performance, etc., to measure day-to-day variations of traffic data. The FHWA Traffic Analysis Toolbox Volume 3 suggested different clustering techniques such as K-means, GMM, etc., to conduct multi-scenario analysis for non-recurrent days. Theoretically, the GMM clustering technique has advantages over the K-means clustering technique. However, in practice, these two methods were not compared to find the suitable clustering technique for multi-scenario analysis. The FHWA’s AMS testbed effort is the most detailed and sophisticated implementation of clustering techniques in transportation engineering. However, limited information is available related to clustering parameters, optimal number of clusters, and input data to measure the day-to-day variation. Additionally, the FHWA testbed only considered non-recurrent days for clustering purposes.

Many researchers have developed ODME models to estimate O-D matrices based on an initial seed matrix combined with measurements such as traffic counts. It is evident from the literature review that there is a need for using supplemental data in ODME methods to improve the quality of the resulting metrices. ODME methods such as Bayesian Inference, Generalized Least Squares, Maximum Likelihood, and Customized Least-Square Estimation have been explored by researchers to produce O-D matrices close to the real-world values. Researchers have used technologies such as AVI and probe data, which provided partial and sample O-D demands that can be used as part of the O-D matrix estimation process. There was no instance found in the literature review that shows the implementation of crowdsourced data in the ODME process.

To check accuracy, several public agencies have explored and tried to validate the estimated traffic volumes obtained from crowdsourced data provided by third-party vendors such as StreetLight. Different DOTs such as Virginia, Oregon, Minnesota, etc., found that there is a linear trend between SL and ground truth data. Another major finding was that estimated traffic volumes based on SL had relatively small absolute percentage errors for locations with high volume (>1000 vph) compared to locations with smaller traffic volume. The VDOT reported that the AADT estimates based on SL had relatively small absolute percentage errors compared to other outputs such as the O-D trips, traffic counts on roadway links, turning movement counts at intersections, and truck traffic volumes, which exhibited relatively higher and consistent errors. A literature review found that there is limited information or resources regarding the integration of crowdsourced data with the ODME process.

CHAPTER 3

SELECTION OF TRAFFIC ANALYSIS DAY(S) UTILIZING CLUSTERING ANALYSIS

An important step in scoping traffic analysis projects is how real-world data is collected, processed, and analyzed to identify the operational scenarios and the representative days for these scenarios. Most existing traffic analysis studies use one set of inputs to the utilized tool to represent one operational scenario for each analyzed peak period. When there are multiple days of data, the analysts in most cases take the averages of traffic volume, speed, travel time, and possibly other measures from weekdays and use these averages as inputs in the development, verification, and validation calibration of the simulation models. In some cases, to find the averages, analysts only use data collected from the peak months of the year rather than the whole year. Averaging measures based on data collected from different days of the year to develop and calibrate the simulation models results in a synthetic day that does not really occur in the real world.

The identification of traffic patterns that best represent the traffic conditions in the AMS effort is critical to the success of these projects. In some efforts, particularly those associated with traffic management and operations, analysts should perform the analysis for different operational conditions throughout the year, including different recurrent congestion levels, incident conditions, and bad weather conditions. However, even if the scope of the analysts is only to analyze normal non-event conditions, there is a need to analyze the variations in traffic conditions from these days to determine the best day(s) that will represent the traffic conditions. In this regard, the FHWA Traffic Analysis Toolbox Volume 3 (Wunderlich et al., 2019) recommended the use of cluster analysis to categorize the traffic conditions in an entire year into a number of clusters and determine a representative day for each of these clusters. Such identification will allow for the selection of the best operational conditions to represent the whole spectrum of traffic conditions present in the network for the purpose of analysis. However, the above-mentioned reference did not provide detailed guidance on how to perform clustering analysis to identify travel conditions.

There are many decisions that need to be made when conducting the cluster analysis to obtain the representative day(s). This study identified three knowledge gaps to be addressed to support these decisions. First, there are various clustering algorithms and methods, each of which have their own strengths and weaknesses. There is a need to research and identify the abilities and performance of these algorithms in the clustering data obtained from typical traffic monitoring systems. Second, it is expected that a higher resolution of the input variables to the clustering analysis can improve the results from the analysis. However, it is not clear how the segregation of the input variables in time and space can improve the quality of the clustering. For example, if the traffic volume used as input is averaged over the whole corridor of the study and for the whole analysis peak period, then important traffic patterns in time and in space might be lost. However, it is not clear how the segregation of the input variables in time and space can improve the quality of the clustering. Third, there is a need to investigate how the performance of the results from the clustering of traffic conditions is to be measured, as well as how to evaluate the aggregation levels.

This chapter presents the methods utilized to identify the representative days and patterns for use in modeling event-free conditions (conditions with no incidents, construction, or weather events). This chapter also evaluates the use of clustering algorithms and methods and assesses the quality of the resulting clusters using various performance metrics. Initially, this chapter compares the results from the two widely used clustering algorithms – the K-means and Gaussian Mixture Model (GMM). Then, it compares the use of input variables aggregated for an entire facility and peak period versus using segregated input variables that are segregated in time and space. The chapter identifies performance assessment techniques for evaluating the clustering results in terms of their abilities to distinguish between different traffic patterns throughout the year. The evaluation of the performance is conducted using metrics reported in statistical studies as well as metrics and visualization techniques based on traffic engineering concepts. In addition, the best representative days obtained based on the clustering as identified based on the assessment is compared with the average day used in common modeling practices, specifically in terms of their abilities to represent the event-free traffic conditions in the year.

3.1. Utilized Data

The utilized case study in this chapter is an I-95 corridor segment in Broward County, Florida, and Palm Beach County, Florida. This segment has a length of 34.3 miles and 23 interchanges. The traffic volume and speed data were collected from point traffic detectors installed and operated by the Transportation System Management and Operations (TSM&O) program of the Florida Department of Transportation (FDOT). The study is based on time-variant data for the northbound (NB) direction during the PM peak period (4:30 PM to 6:30 PM), which was aggregated at 15-minute intervals. This study focused on determining distinguished traffic patterns in event-free or "normal" weekdays. Thus, the data for these days were segregated from the data for "abnormal" weekdays that have non-recurring events like incidents, weather events, and/or construction. However, the investigation can be extended to include non-recurrent event days in the clustering process.

The attributes of each data point used as inputs to the analysis are date, time interval, location, traffic volume in vehicle per hour (vph), and travel time rate in seconds per mile (sec/mi). The volume and speed data are collected using detectors located at an average of half-mile intervals. The traffic management center calculates the travel time based on the speed data collected from the point detectors. These estimates are used in this study to calculate the travel time rates in seconds per mile. The volume and travel time measures are referred to as "key measures" in this study since they are used by the cluster algorithms to group the data.

3.2. Investigated Clustering Algorithms

This study compares the use of the K-means and the GMM algorithms for the identification of traffic conditions for analysis. The K-means clustering partitions the dataset objects into "k" different clusters through an iterative process using simple distance from the cluster's center to assign the cluster membership (Na. et al., 2010). The GMM is based on the Expectation-Maximization algorithm that uses Gaussian distribution in clustering. Unlike the K-means algorithm, which places a circle or a hyper-sphere around each cluster, the directions, and lengths of the axes of GMM's density contours are characterized by an ellipsoid. In theory, this is an advantage of the GMM algorithm since it can accommodate oblong clusters of any shape. If the shape of optimal clusters is not circular, then the data points of two different clusters can overlap each other when using the K-means algorithm due to the rigid (circular) nature of the algorithm. However, it is not clear if this provides an actual advantage when the GMM algorithm is applied to traffic data instead of the more widely used K-means algorithm.

3.3. Investigated Aggregation Levels

There is no clear guidance on how the data obtained from different locations along the analyzed facility can be aggregated in time and space. In investigating the impact of using segregated volume and travel time data in time and space, the study compared four levels of the aggregation of the data.

- *No segregation of the inputs:* The clustering is based on the average values of the traffic volume and travel time over the entire 34.3-mile facility and over the entire peak period (4:30 PM to 6:30 PM) without segregating the data of the facility into data for different segments in space and/or different time intervals, resulting in only two input variables: one for volume, and one for travel time.
- *Spatial segregation of the inputs:* With this level, the clustering is based on the average values of the traffic volumes and traffic travel times for three segments instead of the averages for the full length of the facility. The three segments were a result of dividing the 34.3-mile facility into sub-segments based on the differences

in the congestion patterns and traffic demands between the three segments. This segregation was done considering that utilizing the average values for the whole segment can dilute the information required for identifying different traffic patterns. The spatial segregation resulted in three inputs to clustering for each of the two key measures instead of having one average input per measure for the full segment. The three segments are referred to in this study as Segment-1 or Location 1 (Hallandale Beach Blvd. to I-595), Segement-2 or Location 2 (Davie Rd. to Cypress Creek Rd.), and Segment-3 or Location 3 (Atlantic Blvd. to Linton Blvd.). The lengths of the first two segments are about 8 miles each, while the length of the third segment is about 17 miles. Segment-2 has the highest average traffic flow (8,206 vph), and Segment-3 has the lowest average volume (5,788 vph) among the three segments.

- *Temporal segregation of the inputs:* With this segregation level, the two-hour peak period is divided into four 30-minute time slices, resulting in four inputs to clustering for each of the two key measures instead of one input per measure reflecting the whole period. This level does not involve spatial segregation of the data.
- *Spatial and temporal segregation of the inputs:* This level of segregation involves combining both the spatial and temporal segregation, as mentioned above. The measures are segregated both in time (four 30-minute periods) and space (three segments), resulting in a total of 12 inputs to clustering for each of the two key measures. This level presents the highest resolution of the key measures investigated in this study. The twelve input variables were given acronyms based

on the location and the time periods. For example, at location 1, volume and travel time of the first 30-minute time slice of the analysis period is referred to as avg_vol1.1 and avg_TT1.1, respectively. For location 2 and time slice 3, the variables are referred to as avg_vol2.3 and avg_TT2.3.

3.4. Selection of the Number of Clusters

Irrespective of the algorithm and the aggregation level used in clustering, an important consideration in cluster analysis is to determine the best number of clusters to use in the analysis. The Bayesian Information Criterion (BIC) and Elbow methods are typically used to determine the optimum number of clusters when using the GMM and K-means clustering algorithms, respectively (Burnham et al., 2004; Brownlee, 2019). The BIC statistics is calculated based on logistic regression where a score is minimized to select the optimal number of clusters (Brownlee, 2019). However, for a small sample size, the BIC is not an effective measure for selecting the optimal number of clusters (Burnham et al., 2004). The Elbow technique is a plot that depicts the total within clusters sum of squares (WCSS) for each number of clusters (k). The value of k is selected for the analysis as the point in the graph where the decrease in the WCSS stops being significant as the value of k increases (Marutho et al., 2018). Table 3-1 shows the BIC score points for the different clustering aggregation levels and different number of clusters when using the GMM algorithm. The lowest BIC score indicates the optimal number of clusters and is shaded in Table 3-1. Table 3-1 shows that the optimal number of clusters for no segregation is 3, spatial segregation is 3, temporal segregation is 2, and spatio-temporal segregation is 4. Interestingly, the Elbow method used with the K-means algorithm suggests the same optimal number of clusters (3 for no segregation, 3 for spatial segregation, 2 for temporal segregation, and 4 for spatio-temporal segregation). The higher number of clusters when using the spatio-temporal aggregation level is an indication that this clustering is able to identify more distinctive patterns in the data compared to other segregation levels.

Table 3-1 Optimal number of clusters using BIC score for GMM clustering algorithm with different aggregation levels

Aggregation Level	BIC Score for Number of Clusters		
	2	3	4
No Segregation	-57	-117	-102
Spatial Segregation	-120	-125	-50
Temporal Segregation	-420	-340	-230
Spatio-Temporal Segregation	-375	-1200	-2000

3.5. Assessment of Clustering Quality

After performing the clustering, this study evaluated the quality of clustering using the performance measures reported in the statistics literature and traffic engineering literature, as mentioned earlier.

3.5.1. Performance Measures from Statistical Studies

In this study, the t-SNE visualization method, discussed in the review of literature, is used to visualize the performance of the investigated clustering algorithms and aggregation levels (Al Mamun et al., 2021). The utilization of the t-SNE as a statistical performance measure avoids the limitations of the internal clustering validation metrics, such as the SC, CH, and DB, as discussed in the review of literature.

3.5.2. Performance Assessment Techniques Derived based on Traffic Engineering Literature

In addition to the t-SNE visualization method recommended in the statistics literature, this study explores the use of measures reported in traffic engineering literature to quantify the quality of clustering of the collected traffic data.

3.5.2.1. Sum of Representative Day Distances

The FHWA toolbox III methodology involves identifying a single representative day for each cluster of the time-variant data as the individual day with the minimum distance between the values of the key measures (such as volume and travel time) from the values of these measures for all days in the cluster (Wunderlich et al., 2019). The Representative Day Distance is adopted in this study to obtain a variable called the Sum of the Representative Day Distances to assess how far the traffic conditions of the representative days of the clusters are, which were obtained using a specific clustering approach from the conditions of all days in the year. The method used to calculate the Sum of Representative Day Distances is shown below.

In calculating this Sum of Representative Day Distance, the difference ($\dot{m}_{k,i,n}$) between the value of a certain input variable in the representative day and the value observed on a particular day, expressed as a percentage of the representative day value, is calculated as,

$$\dot{m}_{k,i,n} = \frac{\sqrt{(m_{k,i,n} - m_{k,n})^2}}{m_{k,n}} \tag{3-1}$$

where

$\dot{m}_{k,i,n}$ = difference between the value of a key measure *k* in the representative day and the value observed on a particular day *i* in cluster *n* (percentage);

$m_{k,\,i,\,n}$ = value of a key measure *k* on day *i* in cluster *n*; and

$m_{k,\,n}$ = value of a key measure *k* in the representative day of cluster *n*.

Next, for each cluster, the distance between each individual day in the cluster and the corresponding representative day is calculated as the summation ($d_{i,n}$) of all the differences of all the input variables calculated in the previous step:

$$d_{i,n} = \sum_k \dot{m}_{k,i,n} \tag{3-2}$$

where

$d_{i,n}$ = distance between each individual day *i* in cluster *n* and the corresponding representative day.

Finally, the Sum of Representative Day Distance ($\bar{d}$) for each aggregation level is calculated as the average of all the distances $d_{i,n}$ between each individual day and the representative day of all clusters, calculated in the previous step:

$$\bar{d} = \frac{\sum_i d_{i,n}}{N_{days}} \tag{3-3}$$

where

$\bar{d}$ = sum of Representative Day Distance; and

N_{days} = total number of clustered days.

3.5.2.2. Fundamental Diagram

A macroscopic fundamental diagram is used to visualize the relationship between traffic flow and speed (Geroliminis and Sun, 2011). The horizontal and vertical axes represent the volume (vehicles per 30 minutes) and speed (miles per hour, mph), respectively. In this study, the data points of speeds and volumes of the identified representative days are highlighted on the fundamental diagram to visualize the change in travel conditions between the identified days relative to the traffic conditions for all days in the year.

3.5.2.3. Heat Map

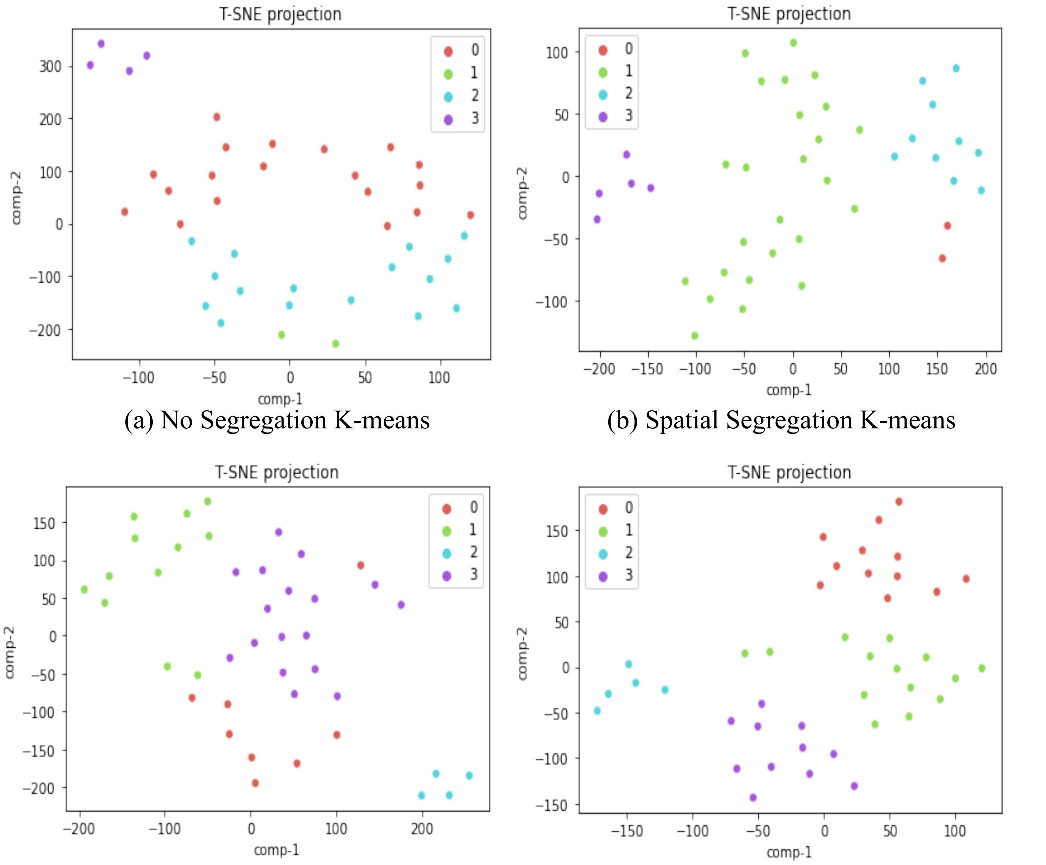
The heat map is a two-dimensional comprehensive visualization technique that shows the magnitude of a phenomenon as variation in color by hue or intensity (Zhao, 2014). Heat maps are used to visualize the traffic patterns generated for each cluster generated using a given clustering approach. Each point on the horizontal axis in the utilized maps represents a day belonging to a given cluster. In the heat maps, the key measures within each cluster are visualized through changing color schemes, with the red color indicating high traffic volume or high travel time; green representing low traffic volume or low travel time, and yellow representing medium traffic volume and travel time.

3.6. Clustering Results

This section describes the results of clustering using different algorithms and aggregation levels. The clustering results are assessed based on performance assessment techniques, which are identified based on statistics literature and traffic engineering literature. In addition, the best representative days obtained based on the clustering and the assessment is compared with the average day used in common modeling practices in terms of their abilities to represent the event-free traffic conditions in a year.

3.6.1. Assessment of Clustering based on Measures from Statistical Literature

Figure 3-1 shows the t-SNE projections for different aggregation levels for the same clustering algorithm (the K-means algorithm) and same number of clusters (four). In Figure 3-1, the data points in the four clusters are displayed in different colors (red for Cluster 1, green for Cluster 2, blue for Cluster 3, and purple for Cluster 4, as indicated by the legends in the figures) for the comparison of the results of clustering using different aggregation levels. Figure 3-1(d) shows that the best clustering is obtained with the spatio-temporal segregation level. Compared to other clustering levels shown in Figure 3-1(a), Figure 3-1(b) and Figure 3-1(c), the spatio-temporal segregation level in Figure 3-1(d) shows a clearer separation of the data points of different clusters, indicating better clustering of the data points.



(a) No Segregation K-means

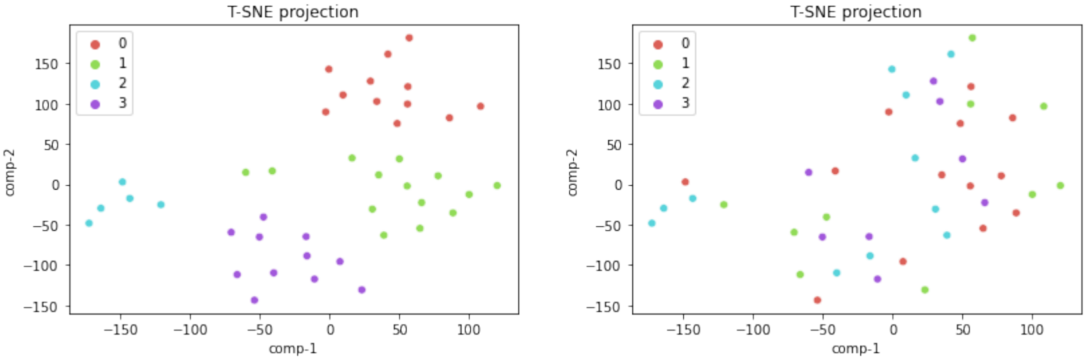
(b) Spatial Segregation K-means

(c) Temporal Segregation K-means

(d) Spatio-temporal Segregation K-means

Figure 3-1 t-SNE Projections to compare the different aggregation levels

Second, with the results from using the two clustering algorithms, the K-means and GMM are compared in Figure 3-2. In this case, the same number of clusters (four) and the same segregation level (spatio-temporal segregation) were used. Figure 3-2(a), which represents the K-means algorithm, shows evenly clustered data, while Figure 3-2(b), which represents the GMM algorithm, shows significant overlaps in the data points from different clusters with no clear separation of the data points for different clusters. In the upcoming sections, the GMM algorithm is ruled out based on the results of the projections in Figure 3-2(b).

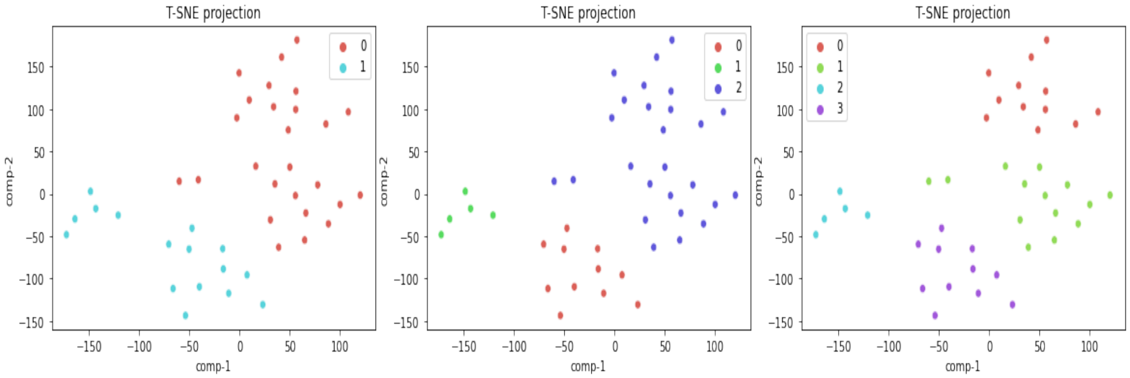


(a) Spatio-temporal K-means

(b) Spatio-temporal GMM

Figure 3-2 t-SNE Projections to compare the K-means and GMM algorithms

Finally, the use of different numbers of used clusters are compared in Figure 3-3 using the same clustering algorithm (K-means) and the same aggregation level (Spatio-Temporal Segregation). The x-axis and y-axis, referred to as comp-1 and comp-2, indicate the relative distances between the data points. The wider the range of the distances between the data points in the extremities within the cluster, the poorer the clustering. This property is called perplexity. Typically, the range of distance of the extreme data points should be closer to the total number of data points to ensure good quality clusters. Figure 3-3(a) indicates that using two clusters results in a large number of data present in each cluster covering wide ranges (Chatzimparmpas et al., 2020). Figure 3-3(b), which represents three clusters, shows better results than two clusters. Figure 3-3(c), which represents four clusters, shows even better results compared to two and three clusters. In the upcoming sections, the utilization of two clusters is ruled out, and further results are based on only using three and four clusters.



(a) K-means with 2 clusters (b) K-means with 3 clusters (c) K-means with 4 clusters

Figure 3-3 t-SNE Projections to compare the use of different number of clusters

Based on the analysis of the results of the t-SNE projections, it can be concluded that the K-means algorithm using the spatio-temporal aggregation level with four clusters provides the most distinctive clustering results with a better ability to group days that share similar travel conditions together.

3.6.2. Assessment of Clustering based on Techniques Derived based on Traffic Engineering Literature

This section compares different approaches to clustering based on selected measures with techniques derived from traffic engineering literature. These techniques include the Sum of Representative Day Distances, the heat map, and the fundamental diagram, as described earlier.

3.6.2.1. Sum of Representative Day Distances

Table 3-2 presents the calculated Sum of Representative Day Distance according to Equations 3-1 through 3-3 for each aggregation level when using the K-means with four clusters. Table 3-2 indicates that the segregation with the spatio-temporal segregation performed the best among all other segregation levels in providing the lowest Sum of Representative Day Distance (230%). This value is significantly lower than the Sum of Representative Day Distance when using the average day, which is 313%, indicating better representation of the traffic conditions with the utilization of the representative days from clustering.

Table 3-2 Sum of Representative Day Distances with different segregation levels using K-means clustering with four clusters

Segregation Level	Sum of Representative Day Distances
No Segregation	281%
Spatial Segregation	253%
Temporal Segregation	261%
Spatio-Temporal Segregation	230%
Average (All-Days)	313%
Average (Feb-Mar)	381%

3.6.2.2. Fundamental Diagram

The macroscopic fundamental diagrams were plotted to examine the advantage of using clustering to select representative days for modeling compared to using the average values. For this purpose, this section presents the diagrams for event-free days for Segement-2, which connects the Davie Rd. interchange and Cypress Creek Rd. and has a length of about eight miles. Figure 3-4 shows a color-coded traffic fundamental diagram with each color representing the data points of one of the four clusters obtained when using the spatio-temporal level of segregation. The data for the representative day for each cluster are highlighted using numbers from 1 to 4. These numbers indicate the 30-minute interval of the data with "1" indicating the data point for the first interval in the peak period, and "4" indicating the data point for the last interval in the peak period. The diagram indicates that both Cluster 1 and Cluster 4 are congested, but Cluster 4 is less congested, and the congestion in the representative day of Cluster 4 starts clearing in the last 30 minutes. Cluster 2 and Cluster 3 have relatively high speeds (55 to 65 mph versus 35 to 45 mph for the congested intervals of Clusters 1 and 4). However, Cluster 3 has much higher demands than Cluster 2, with volumes close to the capacity of the segment. It is interesting to see that although Clusters 1 and 4 have significantly lower speeds than Cluster 3, their volumes are also significantly lower, indicating that these clusters are on the congestion side of the fundamental diagram with the measured volumes constrained by the capacity of the bottleneck segments.

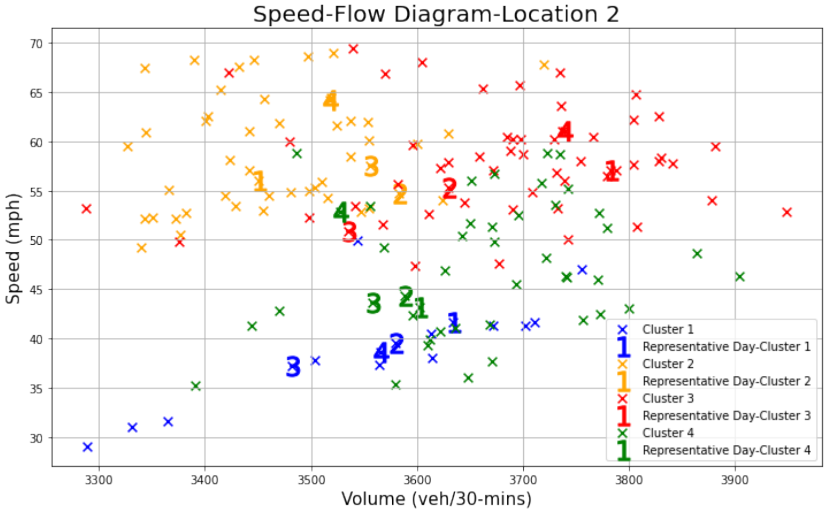


Figure 3-4 Traffic fundamental diagram with highlighted representative day data points for segment 2 for four time slices

Figure 3-5 shows the same fundamental diagram as Figure 3-4, but this time with the highlighted data points with the interval numbers reflecting the average value for all days in the year and the average values for the peak season (February and March). The average speed ranges from 52 mph to 55 mph when averaging the data points through the year, and from 47 mph to 54 mph when averaging the data for the peak season. However, the actual speed of a significant proportion of the actual data points is below 45 mph. Averaging the speed and volume values masked the fact that such days present since the low speeds are averaged with high speeds. As indicated in Figure 3-4 , the speed ranges from 35 mph to 65 mph when using clustering, which indicates better coverage of the operational conditions.

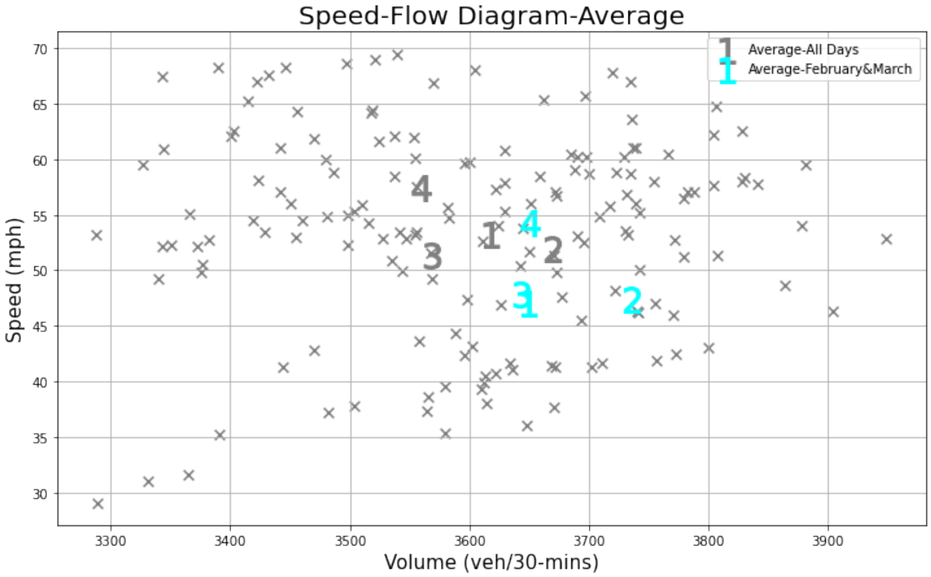


Figure 3-5 Traffic fundamental diagram with highlighted average day data points for segment 2 for four time slices

3.6.2.3. Heat Maps

Heat maps based on the key measures, such as volume and travel time, are constructed for location 2 out of the three locations in the network. The vertical solid black borders divided the heat maps by the clusters, which was a result of the clustering with the spatiotemporal segregated data, as shown in Figure 3-6 and Figure 3-7. The horizontal lines divide the heat map by the four investigated time slices. On the x-axis, each column corresponds to a day within a cluster, with the last two columns representing the average value of all days and the average value of the days in the peak months (February and March), respectively. The representative day of each cluster is displayed in both figures with a red box surrounding it.

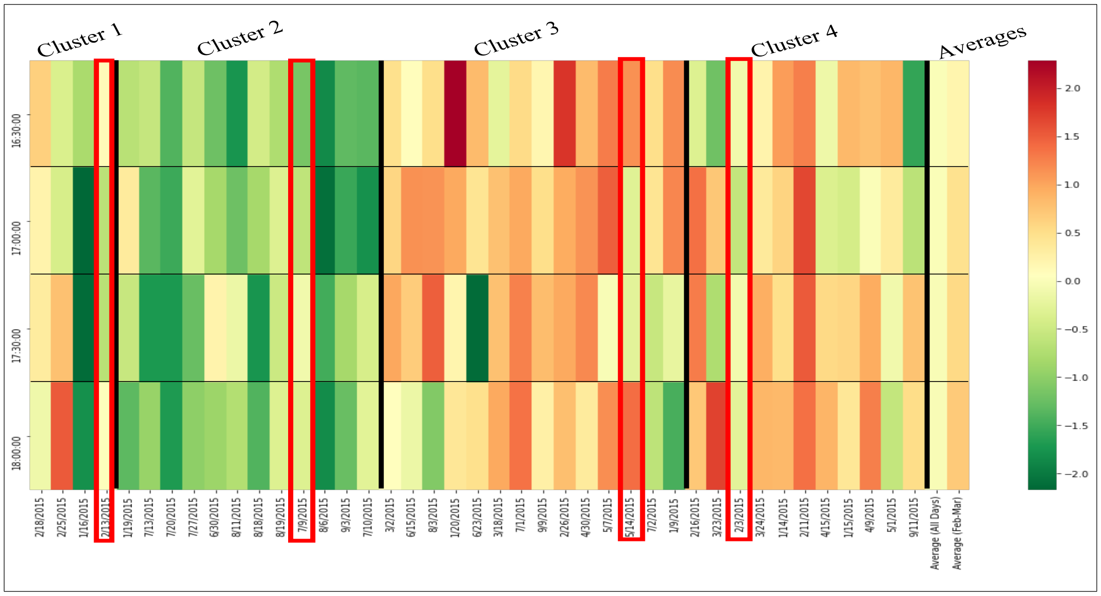


Figure 3-6 Heat Map of location 2 displaying traffic volume

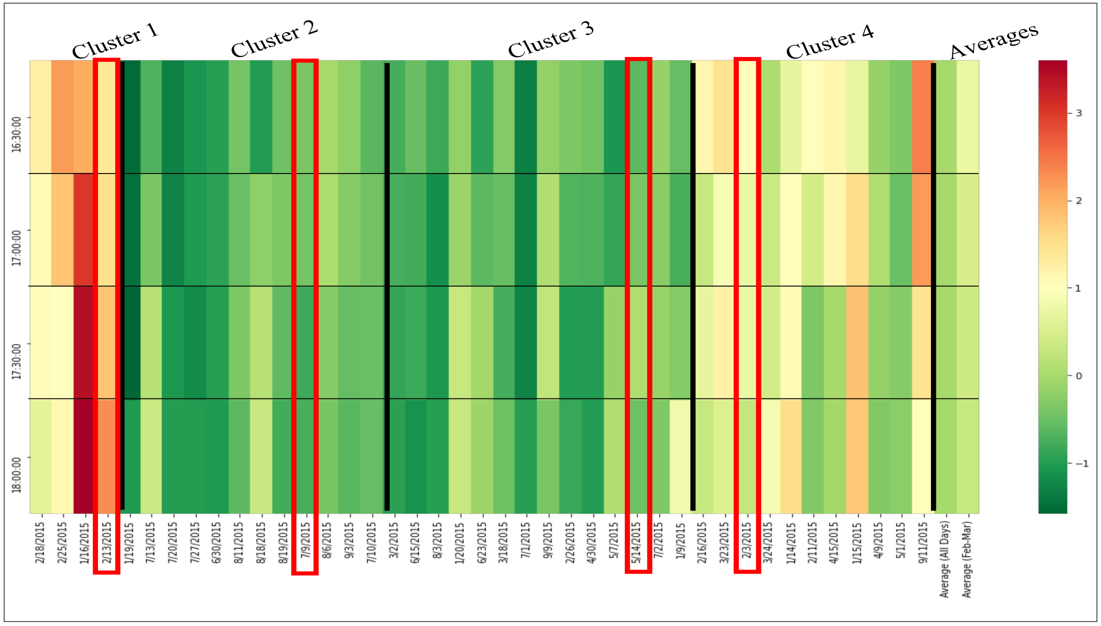


Figure 3-7 Heat Map of location 2 displaying travel time

It is evident from the heat maps that Cluster 1 groups together the days with volumes that are lower than capacity and high travel times indicating congested conditions constrained by capacity. Cluster 2 consists of days with the lowest volumes and lowest travel times, whereas Cluster 3 has high volumes and low travel times, indicating conditions that are close but below the capacity. Finally, Cluster 4 has higher volumes (higher than Cluster 1 and Cluster 2) and lower than that of Cluster 3, and higher travel times compared to Cluster 2 and Cluster 3, indicating that this cluster represents traffic volumes above the capacity of the segments, but not as bad as the conditions of Cluster 1.

It is clear from the last two columns in Figure 3-6 and Figure 3-7 that both methods of averaging clearly level out the variation in volume and travel time, as significant variation can be observed in each of the four resulting clusters. In comparison to the identified clusters, the averages overestimate or underestimate the volumes and travel times. Hence, when modelers choose an average day to represent a certain network condition, different travel patterns are diluted to become a singular travel condition that incorrectly characterizes the modeled network. Clustering that uses spatio-temporal segregation effectively separates the different travel conditions within the network.

Given that there are four identified representative days (one for each resulting cluster) for normal (event-free) days, the next issue to address is which one of these four days will be used in the modeling. Table 3-3 shows the representative day for each cluster and the percentage of days for each of the four clusters resulting from different levels of segregation in time and space. The analyst can examine the number of days in each cluster and the congestion level in the representative day to determine which day to model. As stated earlier in this section, Cluster 1 and Cluster 4 represent the congested conditions on the corridor. Table 3-3 indicates that Cluster 1, which has the highest level of congestion, as indicated earlier, represents 10% of the data points, while Cluster 4 represents 27% of the data points. Depending on the analysis scope, the analyst may decide to model the representative day for Cluster 1, the representative day for Cluster 4, or more than one representative day, if multi-scenario analysis is to be modeled. In other words, the distribution of days in different clusters aids the analyst in examining the number of days in each cluster and the congestion level in the representative day to determine which day to model.

Table 3-3 Number of days in each resulting clusters

Clustering Method	Cluster Components	Representative Day	Percentage of days
Spatio-Temporal	1	2/13/2015	4(10%)
	2	7/9/2015	12(29%)
	3	5/14/2015	14(34%)
	4	2/3/2015	11(27%)
Temporal	1	7/13/2015	8(20%)
	2	4/9/2015	12(29%)
	3	2/25/2015	4(10%)
	4	3/2/2015	17(41%)
Spatial	1	9/9/2015	24 (59%)
	2	7/10/2015	2(5%)
	3	2/13/2015	4(10%)
	4	8/19/2015	11(27%)
No Segregation	1	2/16/2015	19 (46%)
	2	7/20/2015	2(5%)
	3	7/13/2015	16(39%)
	4	2/25/2015	4(10%)

3.7. Summary

The K-means clustering method with four clusters and spatiotemporal segregation level exhibited the best results from the statistical measures (t-SNE plot) and traffic engineering technique (Sum of Representative Day Distances, heat map and fundamental traffic diagram) points of view. Zero use of spatial segregations of the road segments or temporal segregation of the peak period into intervals, along with a lower number of clusters was less effective in clustering the data into distinctive patterns that account for the variations in traffic conditions along the roadway segments and within the peak period considering the day-to-day variations throughout the year. The study also showed that despite its theoretical advantage, the GMM clustering was less effective than the K-means clustering in identifying the traffic patterns in this study.

The results presented in this chapter also clearly show that the use of an average day of the year or the peak season is not acceptable in allowing an effective simulation model's development and calibration. In addition to the fact that averaging volume and travel time data results in synthetic days that do not occur in the real world, such averaging results in diluted congestion levels. The analysis of the case study indicates that a large percentage of the days (the days in Clusters 1 and 4, which constitute about a third of the days) have more congestion levels than those of the averages. Thus, for example, the use of the averages for making highway designs may result in the under-design of the facilities.

This chapter also introduces various techniques to identify the quality of the results of cluster analysis utilized in selecting the representative days. It is clear that the use of standard statistical measures of clustering quality such as the t-SNE plot is not sufficient to provide the full picture of the quality of the resulting traffic patterns and the associated representative days. Measures identified in this study based on traffic engineering concepts, including the Sum of Representative Day Distances, heat map and fundamental traffic diagram, were demonstrated to be critical in assessing the clustering result quality and their ability to represent traffic conditions throughout the year.

CHAPTER 4

UTILIZING CROWDSOURCED DATA IN ORIGIN DESTINATION MATRIX ESTIMATION (ODME) PROCESS

Origin-destination demand estimation is a vital part of MRM. The resulting traffic O-D demand matrices are used as inputs to static and dynamic traffic assignment models that are important components of the MRM. Often, in current practices, the traffic O-D demands used as inputs to the traffic assignment are extracted for a subnetwork from the larger networks modeled in existing regional travel demand models. However, modelers are often challenged with the inadequate quality of the O-D demand matrices obtained from demand forecasting models, which result in large errors in the link counts resulting from the assignment compared to real-world counts (Hadi et al. 2022). This implies the need for further refinement of the ODME procedures and the use of additional data sources as inputs to these procedures to allow these models to produce results that are acceptable for simulation modeling. In many cases, modelers have used ODME procedures that utilize a combinations of traffic counts and initial O-D demand matrices obtained from the demand models to obtain better estimations of the O-D demand matrices. Still, questions remain about the quality of the O-D matrices resulting from the ODME procedures. This chapter explores the possibility of including crowdsourced data sources into the ODME process and evaluates the quality of the resulting O-D matrices.

4.1. Utilized ODME Procedures

Existing ODME procedures are provided with both commercially available and open-source demand forecasting and assignment tools. These procedures can update initial O-D demands, such as those generated by the demand forecasting models to improve their quality based on collected traffic count data and sometimes, other performance measures. The procedures allow the O-D demand to produce link volumes that better correspond to real-world traffic counts when used in traffic assignment. The ODME procedures in these tools generally utilize optimization algorithms to estimate the O-D matrices by minimizing the errors between the model outputs. This optimization results in improved O-D demand matrices compared to the O-D matrices generated by demand forecasting models. Input variables such as link and/or turning movement counts, as well as initial O-D matrices, are often used to seed the optimization process alongside other variables, such as the attractions and production demands per zone. Some of the latest tools provide the advantage of using additional measures such as travel time, queue lengths, and densities as an input to the ODME process.

Modelers often ignore the inclusion of important inputs to the ODME due to time and cost constraints, despite the proven ability to improve the ODME results. Modelers usually perform ODME procedures with or without a seed O-D demand matrix. Although the quality of the ODME output vastly depends on the availability of high-quality initial O-D demand matrices (Lin, 2006), modelers often choose to ignore the use of O-D matrices as seed matrix and rely on count data only. A common practice is to combine traffic volume measurements with initial O-D matrices obtained from demand forecasting models that are used as seed matrices. Also, partial O-D demand matrices obtained from automated roadside readers, such as Bluetooth readers and license plate readers, vehicle tracking using crowdsourced data such as global positioning system (GPS) data, and other data sources, are used as seed matrices to the O-D demand estimation (Hadi et al. 2022).

The ODME algorithms produce O-D demands that minimize deviations from the counts and seed matrices. In the ODME process, analysts can assign relative weights on different variables included in the optimization objective function to reflect the level of confidence in the data used to estimate the variables. In the mesoscopic simulation tool utilized in this study, weight ratio can be assigned, which reflects the relative weight of the seed matrix to traffic counts in the optimization process. For instance, analysts may assign a lower weight ratio if there is a higher confidence in the quality of the traffic counts compared to the quality of the seed O-D demand matrices.

This study investigated the performance of twelve different variations of the ODME processes. These variations use different combinations of initial O-D matrices from the regional demand forecasting model and crowdsourced data from a third-party vendor. The ODME procedures used in this study are provided with the PTV VISUM modeling tool (PTV Group 2021a). VISUM can be used to provide both macroscopic and mesoscopic levels of modeling as part of an MRM implementation. VISUM has two default ODME algorithms, the Least Square and the TFlowfuzzy methods (PTV Group 2021a). The least-square method minimizes the squared distance between the assignment link volume value and the count value. In addition, it also allows the analysts to limit the deviation from the initial O-D matrix that is used as an input to the ODME process by minimizing the squared distance between the initial and modified trip demand values (Hadi et Al., 2022). This method has other advantages, such as short run time, adaptability to large networks, and simplicity. Hadi et al. (2022) reported better results when using the least square algorithm for a case study compared to the results from using the TFlowfuzzy algorithm. Thus, the least square method is used in this analysis.

4.2. Network Preparation

The case study network is located in downtown West Palm Beach, Florida, which was originally extracted as a subnetwork from the regional demand forecasting model. The regional model is referred to as the Southeast Florida Regional Planning Model (SERPM 7.0) (FSUTMSOnline 202l). The extracted subnetwork consisting of the base geometry and initial O-D matrices for the year 2015 was exported to VISUM in a previous project conducted for Palm Beach County (Palm Beach County, 2020; Hadi et al., 2022). The already exported network to the VISUM model in the above-mentioned project was used in this study. The study area consists of 35 signalized and 38 unsignalized intersections. For completeness and future usage, the previously developed VISUM model includes additional segments and intersections outside the boundaries of downtown West Palm Beach, as indicated by the red lines on the map in Figure 4-1. Throughout the study, this case study network is referred to as the “West Palm Beach” network.

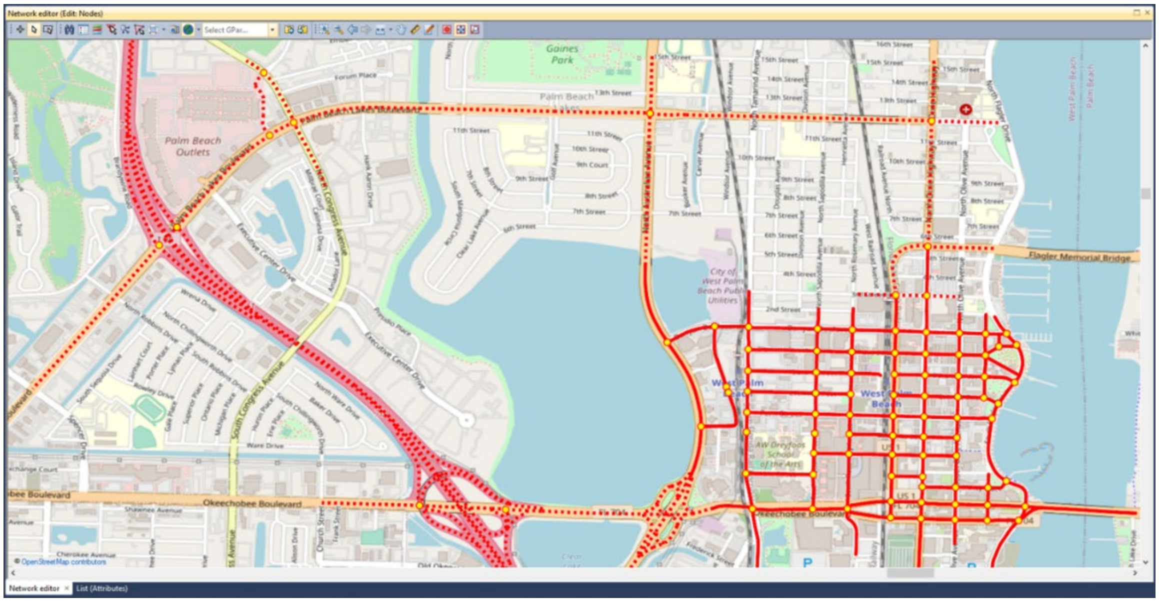


Figure 4-1 Screenshot of Expanded West Palm Beach area in VISUM (Palm Beach County 2020) (Hadi et al., 2022).

4.3. Integration of Crowdsourced Data

In this study, StreetLight (SL) is used as the provider of the crowdsourced data. There are 93 traffic analysis zones (TAZs) in the West Palm Beach network that have been used in the case study. Figure 4-2 shows these zones as highlighted in the SL analysis dashboard display.

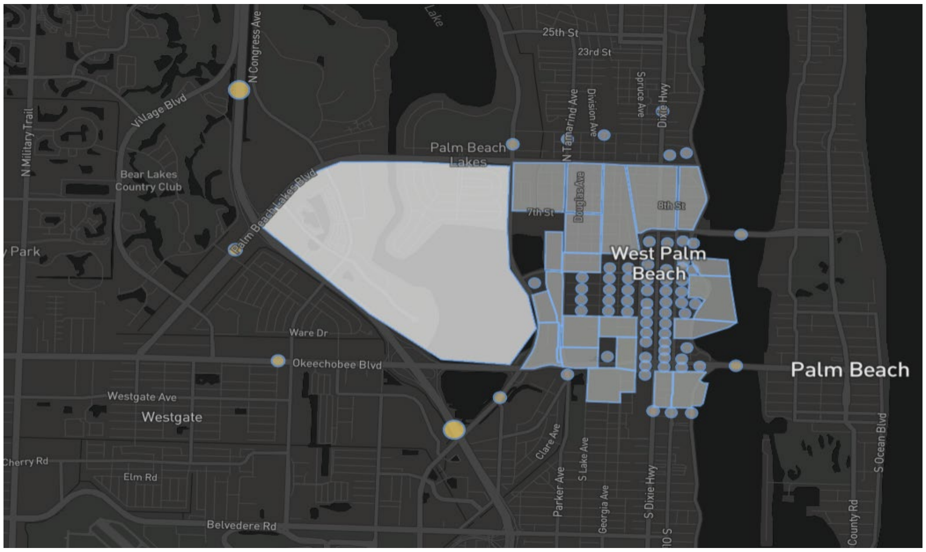


Figure 4-2 StreetLight analysis dashboard of the West Palm Beach network

After selecting the corresponding zones in the SL platform, the SL index value for each O-D pair in the network is obtained from the analysis report generated by the platform. As described earlier, the SL index is a normalized mobile parameter updated each month based on the ratio of the mobile sample available for a location to the total population (collected from the census block). Since the utilized mobile units represent only a fraction of the total numbers of vehicles, the absolute number of trips between the zones cannot be estimated without combining the SL index with other information. However, the SL Index can be used in combination with other data to provide an estimate of the number of trips between any two locations.

In this study, to compute the number of trips between an origin zone and a destination zone, the "SL index proportions" are computed. The SL index proportions is the ratio of the SL index for each O-D pair to the sum of the SL Index values between the origin and all destinations.

In addition to providing the SL Index values, StreetLight provides the users with an option to collect estimated demands from a specific origin zone to a destination zone. The vendor calculates these demands by expanding the mobile data using count data obtained from other sources. These demands referred to as the "SL Expanded" in this study are obtained using a proprietary machine-learning algorithm that uses the SL Index values and ground truth data as inputs (StreetLight Insight 2020 Whitepaper Version 1).

4.3.1. O-D Matrix Development Methods

This study explores twelve variations of the O-D matrix estimation procedure with and without the use of crowdsourced data. As stated earlier, the least square algorithm was utilized to perform the ODME procedure. The ODME procedure used in this study is an assignment-based procedure which utilized the static traffic assignment option in this study. A comparative analysis is performed between all of the investigated methods to identify the best way to develop an O-D matrix. This section describes each method and the use of different weight ratios on the inputs. The methods can be categorized into four different types – Category 1 through Category 4. Methods in Category 1 use an initial O-D matrix from the regional demand model (SERPM 7.0) with relative weights on the initial matrix of 0, 1 and 10. Methods in Category 2 are produced by using an initial O-D matrix that is generated by multiplying the sum of the production trips of each origin as estimated by SERPM 7.0 by the proportion to each destination as estimated by the SL Index proportions, since the SL index values do not provide the actual demands. Thus, the demand forecasting model is used to estimate the generation, while the SL Index is used to estimate the distribution. Methods in Category 3 are similar to the methods in Category 2, but instead of using the trip generation based on SERPM 7.0, the trip generation is based on the results from the Category 1 method with a relative weight of ten. The rationale is that the refined O-D matrix based on both count data and an initial O-D matrix from SERPM 7.0 can provide better trip generation than using trip generation based on the original O-D matrix obtained from the SERPM 7.0 model. Similar to the methods in Category 2, the SL Index proportions are used to distribute the trips to destination in methods from Category 3. The methods in Category 4 use the O-D matrix expanded by the vendor (StreetLight), which is referred to as the SL Expanded. The twelve methods within these four categories are described in the following text.

Method Category 1: The seed matrix used in this category is the initial O-D matrix obtained from SERPM 7.0.

Method 1: This is a Category 1 method with no ODME (uses the O-D matrix obtained from SERPM 7.0 in the assignment process).

Method 1(a): In this method, the ODME procedure is implemented with a relative weight on the seed matrix of zero, indicating that the ODME procedure will estimate the O-D matrix based on the link count only. In other words, this method does not consider the impact of the seed matrix.

Method 1(b): Method 1(b) implements the ODME procedure with a relative weight of ten on the O-D matrix obtained from SERPM 7.0 (the seed matrix) to put more emphasis on the seed matrix than the count deviation.

Method 1(c): Method 1 (c) is similar to Method 1(b), but with a relative weight of one to put equal weight on the count and seed matrix in the ODME process.

Method Category 2: As explained earlier, the methods in this category use the trip generation from demand forecasting model (SERPM 7.0) and calculated trip distributions based on the SL Index.

Method 2: This is a Category 2 method with no ODME (uses an O-D matrix with trip generation from the O-D matrix obtained from SERPM 7.0 and trip distribution based on the SL Index proportion in the assignment process).

Method 2(a): This is a Category 2 method with a relative weight of ten on the initial O-D matrix in the ODME process.

Method 2(b): This is a Category 2 method with a relative weight of one on the initial O-D matrix in the ODME process.

Method Category 3: This Category is similar to the Method Category 2 but with the use of trip generation based on the O-D matrix resulting from Method 1(b) rather than based on the original O-D matrix from SERPM 7.0. As with the methods in Category 2, the SL index proportion is used for trip distribution.

Method 3: This is a Category 3 method with no ODME (uses an O-D matrix with trip generation from the O-D matrix obtained from Method 1b and trip distribution based on the SL Index proportion in the assignment process).

Method 3(a): This is a Category 3 method with a relative weight of ten on the initial O-D matrix in the ODME process.

Method 3(b): This is a Category 3 method with a relative weight of one on the initial O-D matrix in the ODME process.

Method Category 4: The utilized matrix in Method 4 is the default SL expanded O-D matrix provided by the vendor (i.e., matrices estimated by SL without any ODME procedure implemented on the resulting O-D matrix).

Method 4: This is a Category 4 method with no ODME (uses the SL Expanded O-D matrix in the assignment process).

Method 4 (a): This is a Category 4 method with a relative weight of ten on the initial SL Expanded O-D matrix in the ODME process.

4.3.2. Performance Measures of O-D Matrix Estimation Methods

Instead of a single performance indicator, three performance indicators were used to evaluate the performance of different methods in this study. The following three performance metrics (Equations 4-1 to 4-3) were used to evaluate the accuracy of the O-D matrix development methods:

1. Root Mean Squared Error (RMSE): RMSE measures the error of the prediction model. The distance between the predicted values and the benchmark values is squared and averaged over the sample dataset. Since the errors are squared before it is averaged, RMSE puts a high weight to large deviations compared to the

absolute mean error measures. As a result, RMSE is a good indicator when large errors are undesirable.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\acute{x}_i - x_i)^2}{n}} \tag{4-1}$$

where,

RMSE = root mean squared error;

$\acute{x}_i$ = benchmark traffic volume for link/turn movement *i*;

x_i = estimated traffic volume for link/turn movement *i*; and

n = number of estimate-to-benchmark comparisons.

2. Mean Absolute Error (MAE): MAE is the average of the absolute distance between each estimated data value and the benchmark value. The MAE is a linear score which puts equal weight on the individual differences in the average. A small MAE indicates that the data values are close to the benchmark data. The unit used in determining the

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |x_i - \acute{x_\iota}| \tag{4-2}$$

where

MAE = mean absolute error;

$\acute{x}_i$, x_i and n are as previously defined in Equation (4-1).

3. Mean Absolute Percentage Error (MAPE): MAPE is the average of the absolute percentage errors of the predicted value. MAPE is scale-independent and widely used due to easy interpretation of the deviations from the benchmark data.

$$MAPE\ (\%) = \frac{100}{n} \times \sum_{i=1}^{n} \frac{|x_i - \acute{x_\iota}|}{\acute{x}_i} \tag{4-3}$$

where

MAPE = mean absolute percentage error;

$\acute{x}_i$, x_i and n are as previously defined in Equation (4-1).

Table 4-1 shows an overall comparison of the performance measures computed based on Equations 4-1 to 4-3 for the different ODME estimation methods. The error values (performance measures) are rounded off in the table due to space constraints. These performance measures are used to determine the deviation of the link volumes from the real-world link counts referred to as "Link," and the deviation of the turn movement volumes from the real-world counts referred to as "Turn" in Table 4-1.

Table 4-1 Performance measures computed for link and turn volumes

Method				RMSE		MAE		MAPE (%)	
ID	**Seed Matrix**	**Relative Weight**	**ODME**	**Link**	**Turn**	**Link**	**Turn**	**Link**	**Turn**
1	None	N/A	No	769	394	513	249	138	252
1(a)	None	0	Yes	68	76	46	39	20	50
1(b)	SERPM	10	Yes	87	85	52	48	20	60
1(c)	SERPM	1	Yes	903	741	593	379	170	356
2	SERPM x SL Index	N/A	No	156	108	92	59	33	61
2(a)	SERPM x SL Index	10	Yes	1172	779	601	341	116	211
2(b)	SERPM x SL Index	1	Yes	1172	779	601	341	116	211
3	Production of 1(b) x SL Index	N/A	No	99	96	67	53	30	60
3(a)	Production of 1(b) x SL Index	10	Yes	708	616	331	271	90	170
3(b)	Production of 1(b) x SL Index	1	Yes	71	84	47	43	20	50
4	SL Expanded	N/A	No	315	303	166	159	42	94
4(a)	SL Expanded	10	Yes	142	99	77	55	27	60

Method 1(a) and Method 1(b) produced the best results from the first category as it exhibited lower deviations from traffic link volume and turn counts. Method 1(a) has a relative weight of zero, which means the ODME process puts maximum weight on the counts by not taking the seed matrix into consideration. Method 1(b) has a relative weight of ten, which is the maximum weight explored in the ODME processes. Method 1 consists of the initial O-D matrix and no ODME process was applied to it. The initial O-D matrix exhibited high deviation from traffic link volume and turn counts. In Method 1(c), a relative weight of one was implemented to put relatively higher emphasis on the counts than the seed matrix in the ODME process. However, Method 1(c) produced the poorest result among all of the methods in the category with Method ID 1. In the next category, i.e., with Method ID 2, the initial O-D matrix as used in Method 1 is combined with SL index proportions to observe improvements in the O-D methods. Method 2 with no ODME produced results with high discrepancies in the second category like Method 1 in the first category. Method 2(a) and Method 2(b) utilized the ODME process with a relative weight of ten and one, respectively. However, both methods equally produced poor performance with high deviation from the link and turn counts. The next category consists of methods where the seed matrix is developed by multiplying the trip production from Method 1(b) with SL Index proportions. Method 1(b) produced good performance from the first category of the O-D matrix, and as a result, the SL data was integrated into developing Method 3. The seed matrix of Method 3(a) and 3(b) is developed by using a first ODME process with a seed O-D matrix from the regional model combined with the SL Index, respectively. The result is used to estimate the trip production for each zone, and a new O-D matrix is calculated based on the SL Index proportions. Finally, a weight of 10.0 and 1.0 is put on this seed matrix, which is assigned in a second ODME process. Method 3(b) exhibited minimum deviations in link and turn counts. In the fourth category with Method ID 4, the SL Expanded O-D was used with and without the ODME. However, using the

SL Expanded as a seed matrix did not produce better performance than the other methods explored in the previous categories.

4.3.3. Deviations from the Available O-D Matrices

After carefully considering the results of the investigation of the deviations from the ground truth traffic counts of the ODME estimation methods presented in the previous section, this study then compares the deviations for the available O-D matrices resulting from some of the investigated methods. As stated in the previous section, Method 1(b) and Method 3(b) exhibited minimum deviations from the link volume and turn counts and was also similar to the deviation obtained with Method 1(a) when no seed matrix was used in the ODME. These methods are candidates to be recommended for use in the ODME. However, further analysis is needed to determine the degree of deviation from the available O-D matrices of the investigated methods.

To illustrate the trend in the deviation of demands between the O-D pairs of individual methods, each method is compared to the O-D matrix from the demand model and the O-D matrix based on the SL Expanded matrix. Table 4-2 to Table 4-4 show the distribution of mean absolute errors (MAE) of the O-D pair demands among the three matrices. The errors are categorized into seven intervals, as follows: errors that are equal to 0, errors that are between 0 and 50 vehicles ($0<MAE<50$), errors that are between 50 and 100 vehicles ($50\leq MAE<100$), errors that are between 100 and 300 vehicles ($100\leq MAE<300$), errors that are between 300 and 500 vehicles ($300\leq MAE<500$), errors that are between 500 and 1,000 vehicles ($500\leq MAE<1000$), and errors that are greater than

1,000 (≥1000). There are certain O-D pairs that do not have trips between them (i.e., the volume is zero). For that reason, the MAE values are zero in such cases.

Table 4-2 The deviation of the O-D matrix with no ODME or relative weight 0 from different methods to SERPM 7.0 and SL Expanded O-D matrices

(a)Method 1(a) – Relative Weight = 0 (no Seed Matrix)

Method 1 (a)	Relative to SERPM 7.0		Relative to StreetLight Expanded	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage	Frequency Count	Frequency Percentage
≥1000	1	0.01%	0	0.00%
500≤MAE<1000	6	0.07%	5	0.06%
300≤MAE<500	10	0.12%	6	0.07%
100≤MAE<300	38	0.44%	33	0.39%
50≤MAE<100	79	0.92%	51	0.60%
0<MAE<50	4671	54.59%	4656	54.42%
MAE = 0	3751	43.84%	3805	44.47%
MAE Considering All Errors	4805	56.16%	4751	55.53%
Total Cells	8556	100%	8556	100%

(b) Method 2 – Production from SERPM 7.0 and Distribution based on SL Index Proportions, No ODME

Method 2	Relative to SERPM 7.0		Relative to StreetLight Expanded	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage	Frequency Count	Frequency Percentage
≥1000	0	0.00%	1	0.01%
500≤MAE<1000	3	0.04%	1	0.01%
300≤MAE<500	3	0.04%	5	0.06%
100≤MAE<300	43	0.50%	22	0.26%
50≤MAE<100	106	1.24%	61	0.71%
0<MAE<50	5146	60.14%	1513	17.68%
MAE = 0	3255	38.04%	6953	81.26%
MAE Considering All Errors	5301	61.96%	1603	18.74%
Total Cells	8556	100%	8556	100%

(c) Method 3 -Production based on Method 1b Results and Distribution based on SL Index Proportions, No ODME

Method 3	Relative to SERPM 7.0		Relative to StreetLight Expanded	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage	Frequency Count	Frequency Percentage
≥1000	1	0.01%	0	0.00%
500≤MAE<1000	4	0.05%	2	0.02%
300≤MAE<500	6	0.07%	2	0.02%
100≤MAE<300	39	0.46%	22	0.26%
50≤MAE<100	99	1.16%	54	0.63%
0<MAE<50	5162	60.33%	1531	17.89%
MAE = 0	3245	37.93%	6945	81.17%
MAE Considering All Errors	5311	62.07%	1611	18.83%
Total Cells	8556	100%	8556	100%

(d) Method 4 – SL Expanded, No ODME

Method 4	Relative to SERPM 7.0	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage
≥1000	1	0.01%
500≤MAE<1000	6	0.07%
300≤MAE<500	5	0.06%
100≤MAE<300	31	0.36%
50≤MAE<100	52	0.61%
0<MAE<50	5232	61.15%
MAE = 0	3229	37.74%
MAE Considering All Errors	5327	62.26%
Total Cells	8556	100%

Table 4-3 The Deviation of the O-D matrix from methods with relative weight 10 compared to the SERPM 7.0 and SL Expanded O-D matrices

(a)Method 1(b) – Seed Matrix: Method 1, Relative Weight = 10

Method 1 (b)	Relative to SERPM 7.0		Relative to StreetLight Expanded	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage	Frequency Count	Frequency Percentage
≥1000	0	0.00%	0	0.00%
500≤MAE<1000	2	0.02%	3	0.04%
300≤MAE<500	4	0.05%	5	0.06%
100≤MAE<300	42	0.49%	37	0.43%
50≤MAE<100	91	1.06%	63	0.74%
0<MAE<50	4627	54.08%	4547	53.14%
MAE = 0	3790	44.30%	3901	45.59%
MAE Considering All Errors	4766	55.70%	4655	54.41%
Total Cells	8556	100%	8556	100%

(b) Method 2(a) – Seed Matrix: Method 2, Relative Weight = 10

Method 2 (a)	Relative to SERPM 7.0		Relative to StreetLight Expanded	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage	Frequency Count	Frequency Percentage
≥1000	1	0.01%	1	0.01%
500≤MAE<1000	4	0.05%	1	0.01%
300≤MAE<500	6	0.07%	3	0.04%
100≤MAE<300	49	0.57%	29	0.34%
50≤MAE<100	102	1.19%	68	0.79%
0<MAE<50	5960	69.66%	1500	17.53%
MAE = 0	2434	28.45%	6954	81.28%
MAE Considering All Errors	6122	71.55%	1602	18.72%
Total Cells	8556	100%	8556	100%

(c) Method 3(a) – Seed Matrix: Method 3, Relative Weight = 10

Method 3 (a)	Relative to SERPM 7.0		Relative to StreetLight Expanded	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage	Frequency Count	Frequency Percentage
≥1000	2	0.02%	0	0.00%
500≤MAE<1000	5	0.06%	2	0.02%
300≤MAE<500	9	0.11%	2	0.02%
100≤MAE<300	44	0.51%	15	0.18%
50≤MAE<100	99	1.16%	51	0.60%
0<MAE<50	5964	69.71%	1541	18.01%
MAE = 0	2433	28.44%	6945	81.17%
MAE Considering All Errors	6123	71.56%	1611	18.83%
Total Cells	8556	100%	8556	100%

(d) Method 4(a) – Seed Matrix: Method 4, Relative Weight = 10

Method 4 (a)	Relative to SERPM 7.0		Relative to StreetLight Expanded	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage	Frequency Count	Frequency Percentage
≥1000	1	0.01%	0	0.00%
500≤MAE<1000	5	0.06%	0	0.00%
300≤MAE<500	6	0.07%	0	0.00%
100≤MAE<300	27	0.32%	3	0.04%
50≤MAE<100	86	1.01%	25	0.29%
0<MAE<50	5196	60.73%	1526	17.84%
MAE = 0	3235	37.81%	7002	81.84%
MAE Considering All Errors	5321	62.19%	1554	18.16%
Total Cells	8556	100%	8556	100%

Table 4-4 The deviation of the O-D matrix from methods with relative weight 1 compared to the SERPM 7.0 and SL Expanded O-D matrices

(a) Method 1(c) – Seed Matrix: Method 1, Relative Weight = 1

Method 1 (c)	Relative to SERPM 7.0		Relative to StreetLight Expanded	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage	Frequency Count	Frequency Percentage
≥1000	0	0.00%	0	0.00%
500≤MAE<1000	7	0.08%	3	0.04%
300≤MAE<500	13	0.15%	8	0.09%
100≤MAE<300	44	0.51%	47	0.55%
50≤MAE<100	88	1.03%	56	0.65%
0<MAE<50	5902	68.98%	3058	35.74%
MAE = 0	2502	29.24%	5384	62.93%
MAE Considering All Errors	6054	70.76%	3172	37.07%
Total Cells	8556	100%	8556	100%

(b)   Method 2(b) – Seed Matrix: Method 2,Relative Weight = 1

Method 2 (b)	Relative to SERPM 7.0		Relative to StreetLight Expanded	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage	Frequency Count	Frequency Percentage
≥1000	1	0.01%	0	0.00%
500≤MAE<1000	6	0.07%	3	0.04%
300≤MAE<500	11	0.13%	3	0.04%
100≤MAE<300	51	0.60%	40	0.47%
50≤MAE<100	96	1.12%	49	0.57%
0<MAE<50	5958	69.64%	1516	17.72%
MAE = 0	2433	28.44%	6945	81.17%
MAE Considering All Errors	6123	71.56%	1611	18.83%
Total Cells	8556	100%	8556	100%

(c) Method 3(b) – Seed Matrix: Method 3, Relative Weight = 1

Method 3 (b)	Relative to SERPM 7.0		Relative to StreetLight Expanded	
MAE by Error Range (Vehicle per PM Peak Period)	Frequency Count	Frequency Percentage	Frequency Count	Frequency Percentage
≥1000	1	0.01%	0	0.00%
500≤MAE<1000	5	0.06%	3	0.04%
300≤MAE<500	7	0.08%	2	0.02%
100≤MAE<300	49	0.57%	33	0.39%
50≤MAE<100	93	1.09%	59	0.69%
0<MAE<50	5155	60.25%	1514	17.70%
MAE = 0	3246	37.94%	6945	81.17%
MAE Considering All Errors	5310	62.06%	1611	18.83%
Total Cells	8556	100%	8556	100%

Table 4-2 shows the MAE of the resulting O-D matrix compared to the SERPM 7.0 matrix and SL Expanded matrix for the ODME methods. Method 1(a), where no initial seed matrix was used, produced high overall deviation from both the SERPM 7.0 and SL Expanded matrices of about 56% and 55%, respectively. When using the SL index to distribute the trips generated based on the SERPM 7.0 model or the results from Methods 2 and 3, the MAE, compared to the SL Expanded matrix, dropped significantly, from 55% to about 19%, while the error, compared to the SERPM 7.0 matrix, increased from 56% to 62%. This indicates that improvement in the MAE relative to SL Expanded is much higher than the increase in the MAE relative to the SERPM 7.0 when using the SL Index proportion to distribute the generated trips.

Table 4-3 shows that Method 1(b), which uses the SERPM 7.0 O-D matrix as an initial O-D matrix with a relative weight of 10 in the ODME, resulted in high deviation from both SERPM 7.0 (about 55%) and SL Expanded (about 54%). However, in terms of large errors (MAE >300), Method 1(b) produced a higher frequency of errors compared to Method 1(a). Method 1(b) has only six large error counts, compared to 17 large errors in Method 1(a), when compared to the default SERPM 7.0 O-D matrix. This indicates that Method 1(a) is further away from the existing O-D matrix developed by the regional demand forecasting model (SERPM 7.0). Interestingly, when the SL Index is incorporated to distribute the generated traffic and then conducting an ODME with a relative weight 10, as implemented in Methods 2(a) and 3(a), the deviation from the O-D matrix developed from SERPM 7.0 increased to about 71%, while the deviation from the SL Expanded matrix dropped to 19%. This indicates that Method 2(a) and Method 3(a) are similar to SL estimates but are distant from the O-D matrix generated by SERPM 7.0, compared to Method 1(b).

Table 4-4 shows that Method 1(c), which uses the SERPM 7.0 O-D matrix as an initial O-D matrix with relative weight one in the ODME without using SL data resulted in higher deviation from the SERPM 7.0 O-D matrix (about 71%) and lower deviation from the SL Expanded (about 37%). In terms of large error frequencies (MAE >300), Method 1(c) produced 20 error counts. Interestingly, in Method 2(b), the deviation from the O-D matrix developed from SERPM 7.0 increased to about 72%, while the deviation from the SL Expanded matrix dropped to 19%. This indicates that Method 2(b) is similar to SL estimates but distant from the O-D matrix generated by SERPM 7.0 and similar to Method 1 (c). Method 3(b) generated lower overall deviations from the O-D matrix developed from SERPM 7.0 (62%) and from SL estimates (19%).

4.4. Summary

As stated earlier, based on traffic count deviation, both Method 1(b) and Method 1(c) produced relatively good deviation from the real-world traffic counts, compared to Method 1(a), which does not use a seed matrix and optimizes the O-D matrix. Method 1(b) uses the O-D matrix developed from the SERPM 7.0 model as a seed matrix and a relative weight of 10 in the ODME process. Method 1(b) does not use the crowdsourced data. Method 3(b) further refines the results from Method 1(b) by using an initial matrix resulting from the redistribution of the generated trips according to the O-D matrix calculated in Method 1(b) using the SL Index proportions, and then conducting a second ODME process with weight one on the initial matrix. Although both Method 1(b) and 3(b) resulted in relatively good deviations from the counts (as shown in Table 4-1), the results in Table 4-2 to Table 4-4 show that Method 1(b) produced an overall deviation from the O-D Matrix developed from the SERPM 7.0 model and the SL Expanded by about 55% and 54%, respectively. Method 3(b) produced corresponding deviations of 62% and 19%, respectively, indicating generally less deviation than Method 1(b) when it comes to comparing it with the existing O-D matrix from SERPM 7.0.

To further examine the resulting O-D matrix from the use of Method 3(b), the demands for the O-D pairs with the highest errors from Method 1 (i.e., the O-D matrix from SERPM 7.0 and Method 4, and the O-D matrix from the SL Expanded), are investigated. Figure 4-3 shows that SERPM 7.0 always overestimates the number of trips, while Method 3(b) produces results that fall between the values from SERPM 7.0 and SL Expanded. SERPM 7.0 overestimates the O-D trips because the trips are estimated without considering actual traffic but instead, are based on the data collected from the household survey based on demographic attributes. For instance, a link with a capacity of 2,000 vehicles per link can be assigned much more traffic than reality if only considering the results from the survey. On the other hand, the SL Expanded O-D trips are not perfect either, as it uses partial counts to predict the O-D trips, as discussed earlier. The developed method, Method 3(b), uses data from multiple sources (SERPM 7.0, SL, and traffic counts) to produce results that produce reasonable deviations from real-world conditions.

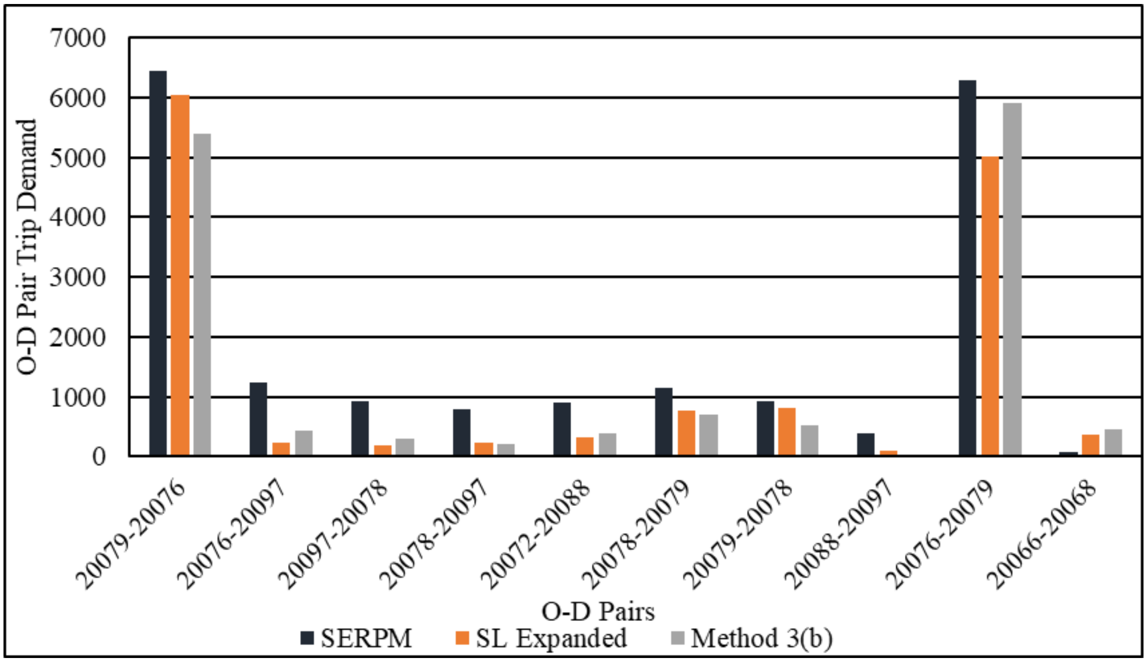


Figure 4-3 O-D Pairs trip demand with highest errors

Table 4-5 to Table 4-8 show the turning movement volumes for four critical intersections resulting from assigning the O-D matrix produced from Method 3(b), compared to those resulting from assigning the SERPM 7.0 and SL Expanded O-D matrices. The four critical intersections are located at Okeechobee Boulevard at the Tamarind Avenue/Parker Avenue intersection, Banyan at the North Tamarind intersection, Banyan Boulevard at the North Dixie Highway intersection, and Lakeview Avenue at the North Dixie Highway intersection in downtown West Palm Beach, which are referred to as Intersection 1, Intersection 2, Intersection 3, Intersection 4, respectively.

The comparative analysis, performed for the four intersections in the network, shows that Method 3(b) in particular produced lower deviations from the real-world turn movement counts. Overall, Method 3(b) exhibited the least deviation from the real-world conditions when compared to the other two examined alternatives.

Table 4-5 Turning movement counts for intersection 1

		Approach												
Intersection 1	Method ID	SBR	SBT	SBL	EBR	EBT	EBL	NBR	NBT	NBL	WBR	WBT	WBL	Total
Counts	Real World	913	596	47	303	1369	382	106	177	103	51	2096	112	6255
	1	839	727	48	303	1350	394	105	187	121	125	2256	91	6546
	3(b)	1041	607	43	298	1351	389	92	191	137	13	2103	92	6357
	4	1058	598	95	270	1401	427	137	201	162	38	2136	105	6628
Percentage Deviation from Real-World Count	1	8%	22%	2%	0%	1%	3%	1%	6%	17%	145%	8%	19%	5%
	3(b)	14%	2%	9%	2%	1%	2%	13%	8%	33%	75%	0%	18%	2%
	4	16%	0%	102%	11%	2%	12%	29%	14%	57%	25%	2%	6%	6%

Table 4-6 Turning movement counts for intersection 2

		Approach												
Intersection 2	Method ID	SBR	SBT	SBL	EBR	EBT	EBL	NBR	NBT	NBL	WBR	WBT	WBL	Total
Counts	Real World	146	503	10	50	472	66	40	399	141	14	1656	178	3675
	1	230	447	0	102	409	58	11	399	701	15	1530	85	3987
	3(b)	140	490	0	40	391	48	0	354	118	15	1588	217	3401
	4	155	510	1	49	386	43	9	336	93	21	1541	198	3342
Percentage Deviation from Real-World Count	1	58%	11%	100%	104%	13%	12%	73%	0%	397%	7%	8%	52%	8%
	3(b)	4%	3%	100%	20%	17%	27%	100%	11%	16%	7%	4%	22%	7%
	4	6%	1%	90%	2%	18%	35%	78%	16%	34%	50%	7%	11%	9%

Table 4-7 Turning movement counts for intersection 3

		Approach									
Intersection 3	Method ID	SBR	SBT	SBL	EBR	EBT	EBL	WBR	WBT	WBL	Total
Counts	Real World	388	354	32	42	95	62	18	263	68	1322
	1	332	263	76	45	160	55	74	338	36	1379
	3(b)	375	380	30	46	94	57	74	174	34	1264
	4	244	357	23	39	113	49	57	288	68	1238
Percentage Deviation from Real-World Count	1	14%	26%	138%	7%	68%	11%	311%	29%	47%	4%
	3(b)	3%	7%	6%	10%	1%	8%	311%	34%	50%	4%
	4	37%	1%	28%	7%	19%	21%	217%	10%	0%	6%

Table 4-8 Turning movement counts for intersection 4

		Approach					
Intersection 4	Method ID	SBR	SBT	NBL	WBT	WBL	Total
Counts	Real World	208	543	327	1086	237	2401
	1	242	680	359	1159	195	2635
	3(b)	262	561	369	1123	203	2518
	4	258	571	369	1121	212	2531
Percentage Deviation from Real-World Count	1	16%	25%	10%	7%	18%	10%
	3(b)	26%	3%	13%	3%	14%	5%
	4	24%	5%	13%	3%	11%	5%

CHAPTER 5

USING CROWDSOURCED DATA FOR SEGMENT-LEVEL VOLUME COUNT ESTIMATION

In addition to providing data that can be used for O-D matrix estimation, crowdsourced data have been proposed to provide estimates of daily and hourly traffic volume counts. Such estimates can be used to provide support to analysis, simulation, and MRM, as volume count data is expensive to collect and in many cases, counts are available only for short periods of time and may not cover all links in the network. This chapter provides a method to utilize crowdsourced data in combination with the real-world data collected from permanent detectors at specific locations to estimate the link volumes. This chapter also discusses the assessment, the accuracy, and the transferability of using the data for link volume estimation.

5.1. Existing Volume Count Data

Data from the same West Palm Beach network presented in Chapter 4 are used in the development and assessment of the method to estimate the volume counts based on SL data in this chapter. There are two permanent count stations (PCS) maintained by the Florida Department of Transportation (FDOT) in the case study network. The first is located near the Flagler Memorial Bridge (Site 0087), while the second is on I-95 (Site 0174), as indicated by the red circles in Figure 5-1. The whole year traffic volumes from these two permanent count stations are used in this study. The year-long data from 2020 is collected from Florida Traffic Online, which is an online tool established and maintained by the FDOT.

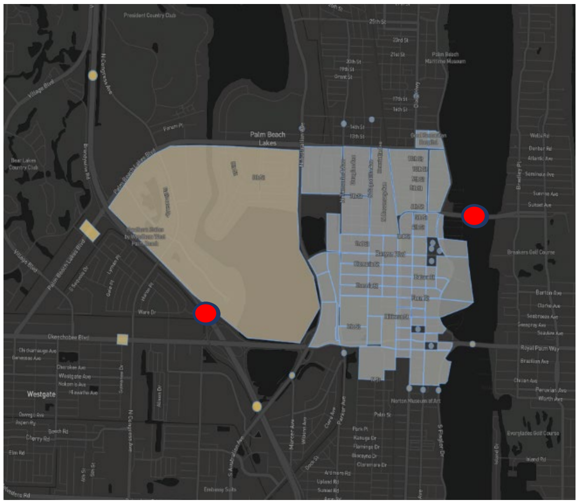


Figure 5-1 PCS locations in StreetLight dashboard

5.2. Seasonal Variations in the Data

The first question to be answered is whether crowdsourced data can provide enough information to determine the seasonal variation in the data. This is important because an analyst may need to conduct multi-scenario analysis or need to estimate the traffic volumes in the peak season of the year based on data collected in other seasons. This section compares seasonal variation in average daily traffic (ADT) from the PCS (used in this case as benchmark data) with the variations based on the SL Expanded volume, in vehicles per day (vpd). The monthly average daily traffic (MADT) of June, July and August are aggregated to represent the summer season ADT and the MADT for December, January and February, which are aggregated to represent the winter season ADT. The traffic demand is expected to be higher during the winter season in West Palm Beach, Florida, compared to the summer.

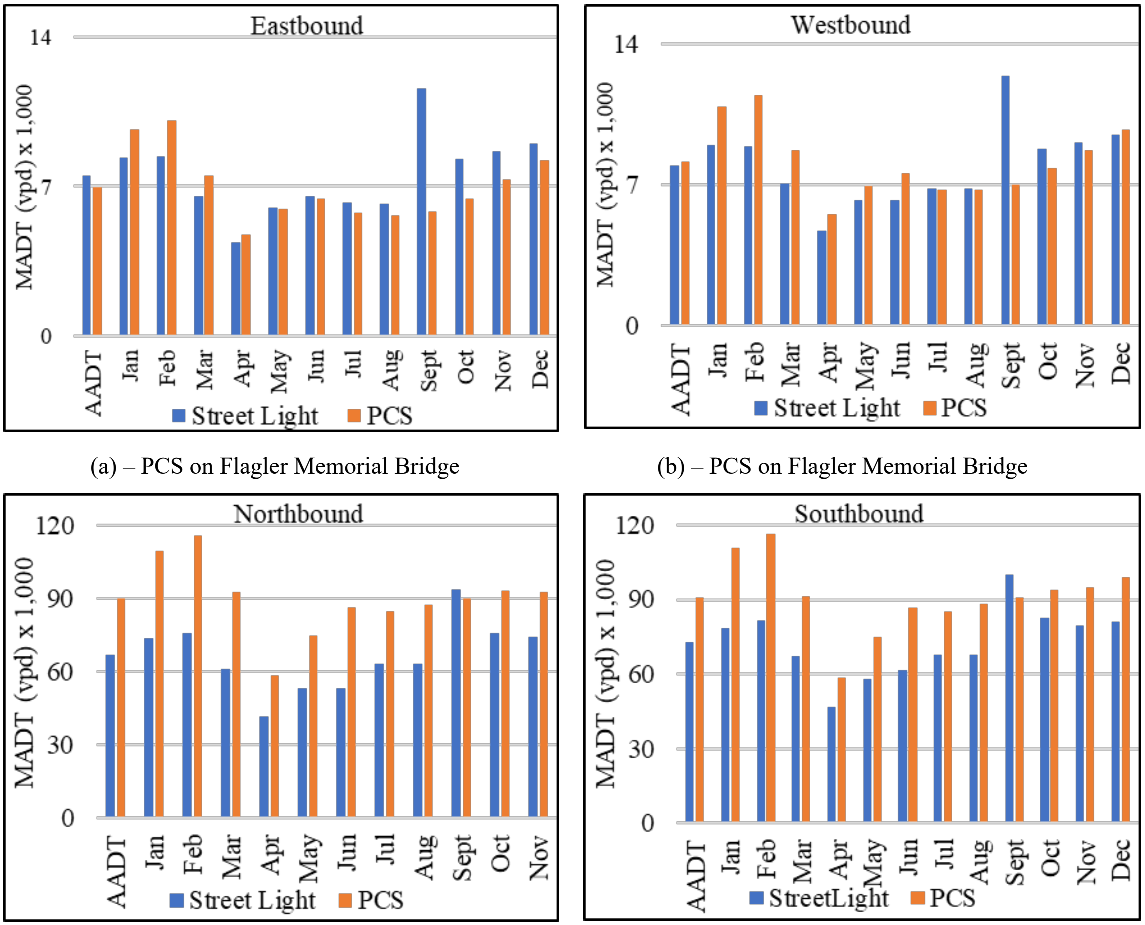
Table 5-1 compares the seasonal variation in the ADT based on the PCS and SL Expanded (SL Exp) data in the two locations where PCSs are present, as discussed earlier. The two locations are the Flagler Memorial Bridge for eastbound and westbound traffic, and the I-95 Congress location for northbound and southbound traffic. As shown in Table 5-1, the seasonal factors (SF) were calculated for PCS and SL Expanded data sources. For instance, the summer SF is calculated by dividing the summer ADT based on the three months (i.e., June, July, and August) over the ADT based on all six months. It is evident that there is a significant difference of ADT in some cases, such as eastbound (EB) (28%) and westbound (WB) (27%) during the summer season.

Table 5-1 Seasonal ADT comparison SL Expanded vs PCS Counts

	Flagler Memorial Bridge				I-95 Congress			
	Eastbound		Westbound		Northbound		Southbound	
	Winter	Summer	Winter	Summer	Winter	Summer	Winter	Summer
PCS (vpd)	28,004	17,823	32,073	21,072	322,701	258,266	326,074	259,769
SL Exp (vpd)	25,969	18,922	27,316	19,873	223,363	179,689	241,207	197,272
PCS SF	0.61	0.39	0.60	0.40	0.56	0.44	0.56	0.44
SL Exp SF	0.50	0.50	0.50	0.50	0.49	0.51	0.48	0.52
Percentage Difference	18%	28%	17%	27%	12%	15%	13%	16%

5.3. Comparison of Monthly Volumes between PCS and SL Expanded Data

Figure 5-2 shows a comparison of the MADT for each month of the year based on the data from PCS with those based on SL Expanded. Interestingly, for the eastbound (EB) and westbound (WB) directions at the Flagler Memorial Bridge location, the MADT estimates based on SL data are similar to the PCS data, except for the months of January, February, and September, as displayed in Figure 5-2 (a) and Figure 5-2 (b). The annual average daily traffic (AADT) for the eastbound direction is 6,961 vpd and 7,507 vpd based on PCS and SL data, respectively. The AADT for the westbound direction is 7,958 vpd and 8,157 vpd based on PCS and SL data, respectively. However, for the northbound (NB) and southbound (SB) directions of the I-95 location, the SL data underestimates the MADT compared to the PCS data in most cases, as seen in Figure 5-2 (c) and (d). In terms of AADT at the I-95 location, high discrepancies are observed for I-95 directions in the I-95. The annual average daily traffic (AADT) for the northbound direction is 90,276 vpd and 66,944 vpd based on PCS and SL data, respectively. The AADT for the southbound direction is 90,836 vpd and 72,720 vpd based on PCS and SL data, respectively. The SL Expanded underestimation of the volume is possibly due to the expansion of the partial data by SL using data collected from other locations that have different characteristics or are less congested than freeway facilities in South Florida.



(a) – PCS on Flagler Memorial Bridge

(b) – PCS on Flagler Memorial Bridge

(c) – PCS on I-95

(d) – PCS on I-95

Figure 5-2 Comparison of MADT based on SL Expanded vs. PCS Counts

5.4. Estimating Daily Volume at PCS locations using Regression Analysis

The comparison in the previous section indicates that there are large discrepancies between the AADT and MADT estimated based on SL Expanded data and PCS data. This section presents and investigates an enhanced methodology to estimate the traffic volumes based on the SL Index values using a regression model to expand the data rather than using the SL Expanded data.

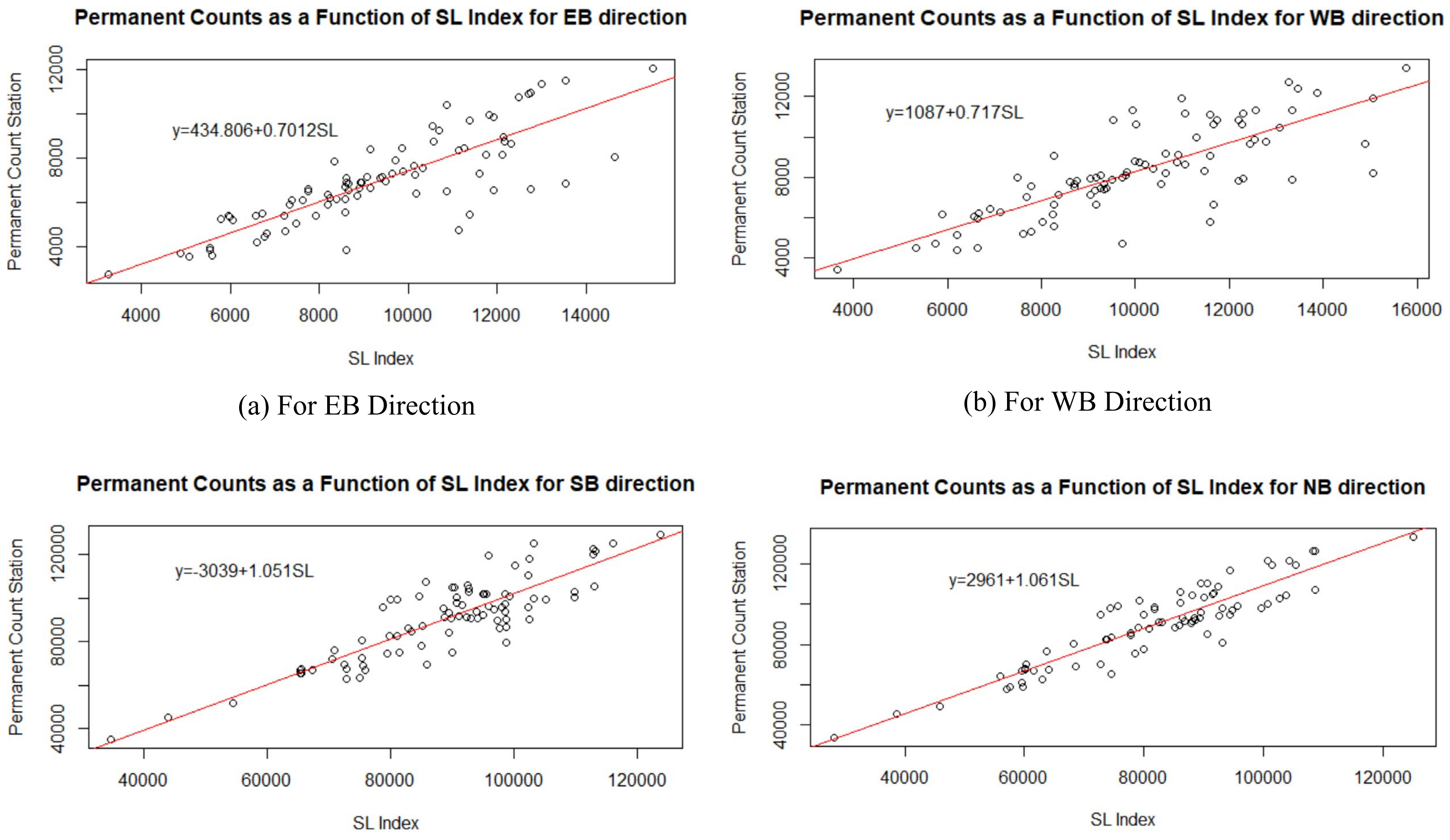
As discussed in Chapter 4, the SL index is a metrics used by SL to estimate features like segment-level volume, O-D trips, and turning movement volumes. The SL index is expanded by Streetlight to full counts based on data from ground truth locations that may not be in the same region or network of the study. Thus, it is expected that expanding the SL index values to full counts based on ground truth counts collected from the same network under investigation can produce better results. To develop and assess a method for SL Index expansion based on local network data, the relationship between the PCS values and the SL Index is fitted in a regression model, with the volumes based on the PCS values as the response or dependent variable (indicated as PCS in the regression equation in the following sections) and SL Index as the explanatory or independent variable (indicated as SL in the regression equations in the following sections). Data aggregated daily from eleven months chosen randomly out of the twelve months in the year 2020 are used as training data, and data from the remaining month is used as testing data in the regression model. For each month, data collected from the PCS and SL Index data for the same locations discussed in the previous section are averaged for each of the seven days in the week over the whole month, resulting in seven data points per month. For instance, all of the Mondays in one certain month are averaged. Similarly, the rest of the days in the month are averaged.

Regression models are fitted for the two different location sites in five variations, including a model for each of the four directions and a model with all four directions combined. The statistical inferences are summarized in Table 5-2. The R^2 value is a goodness-of-fit measure that indicates how much of the deviations in the dependent variable are explained by the independent variable in the fitted model. The higher the R^2, the better and more accurate the model is in predicting the real-world values using SL Index as an input. The fitted regression models, for every direction, have an R^2 greater than 60%, indicating that the models can predict more than 60% of the variation in the PCS values. The model developed with all directions has an R^2 of 97.6%, which is the best fitted model among all of the models. The p-value determines if the explanatory variable is significant in the model. A p-value less than a certain significance level, usually 5%, indicates that the explanatory variable is significant. In this case, all p-values are close to zero, indicating that, for all directions, the SL Index aggregated daily for every month is a significant predictor of PCS values.

Table 5-2 Regression statistical inferences for PCS Counts as a function of SL Index using daily data

Direction	Regression Equation	R^2	p-value
EB	PCS = 434.8060 + 0.7012SL	69.0%	0.000
WB	PCS = 1087.0000 + 0.717SL	62.0%	0.000
SB	PCS = -3039.0000 + 1.051SL	77.8%	0.000
NB	PCS = 2961.0000 + 1.061SL	83.8%	0.000
All Directions	PCS = -2531.0000 + 1.082SL	97.6%	0.000

The regression plots, along with the regression models for the two different location sites in five variations, including a model for each of the four directions and a model with all four directions combined, are shown in Figure 5-3 and Figure 5-4, respectively.



(a) For EB Direction

(b) For WB Direction

(c) For SB Direction

(d) For NB Direction

Figure 5-3 PCS Counts vs. SL Index fitted regression model using daily volumes

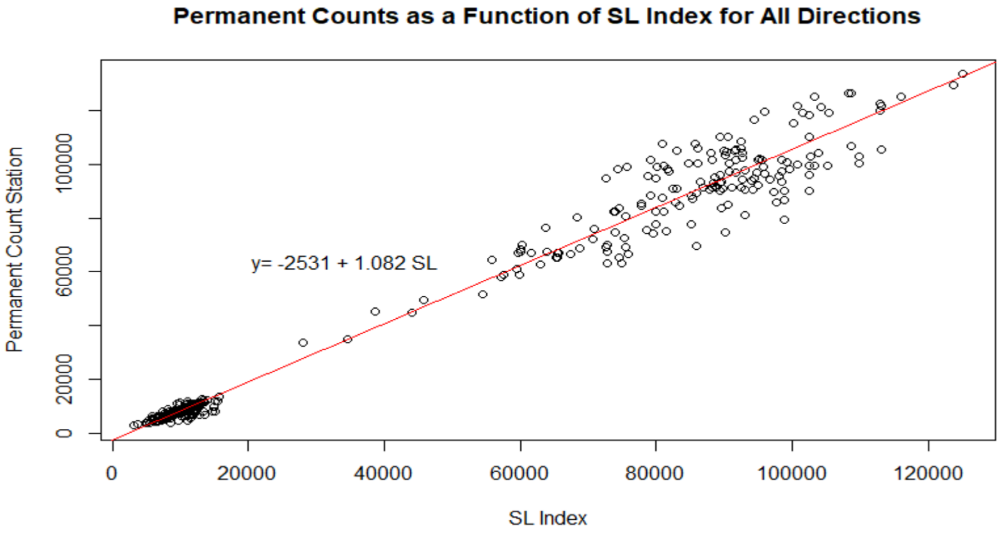


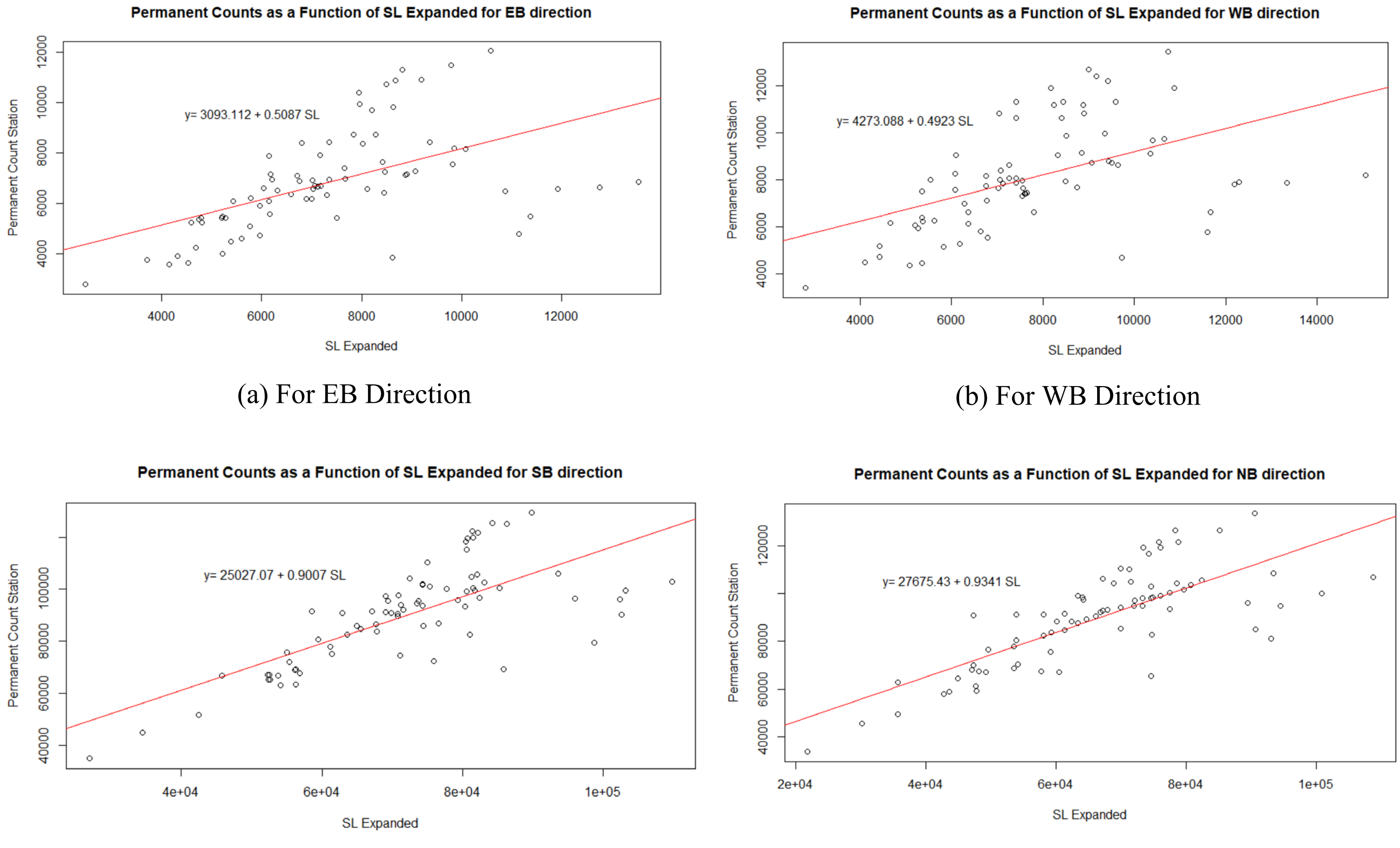
Figure 5-4 PCS Counts vs. SL Index fitted regression model for All Directions using daily volumes

Another approach to expanding the SL data based on the ground truth counts is to develop another regression model to correct the SL Expanded based on PCS data instead of using the SL Index as described above, again at the daily aggregation level, with the volumes based on the PCS values as the response or dependent variable (referred to as PCS in the regression equations) and SL Expanded as the explanatory or independent variable (referred to as SL in the regression equations). The statistical inferences of the fitted models are summarized in Table 5-3. The R^2 values are relatively lower for all five models compared to the models developed earlier using the SL Index. The values are particularly low for the models developed based on the eastbound and westbound data, which are 24% and 29%, respectively. The lower R^2 values indicate that the models developed per direction for SL Expanded are not as accurate to predict the variations in the counts as the models developed based on the SL Index. However, the All Directions model still produced a very good fit with an R^2 of 94.1%.

Table 5-3 Regression statistical inferences for PCS Counts as a function of SL Expanded using daily data

Direction	Regression Equation	R^2	p-value
EB	PCS = 3093.1120 + 0.5087SL	29.0%	0.000
WB	PCS = 4273.0880 + 0.4923SL	23.7%	0.000
SB	PCS = 25027.0000 + 0.9007SL	53.9%	0.000
NB	PCS = 27675.4300 + 0.9341SL	58.4%	0.000
All Directions	PCS = -547.0 + 1.281SL	94.1%	0.000

The regression plots along with the regression models are shown in Figure 5-5 for EB, WB, SB, and NB and in Figure 5-6 for all directions, respectively. Based on the statistical inferences from the above results, regression models developed based on the SL Index for daily volumes was selected for further analysis.



(a) For EB Direction

(b) For WB Direction

(c) For SB Direction

(d) For NB Direction

Figure 5-5 PCS Counts vs. SL Expanded fitted regression model using daily volumes

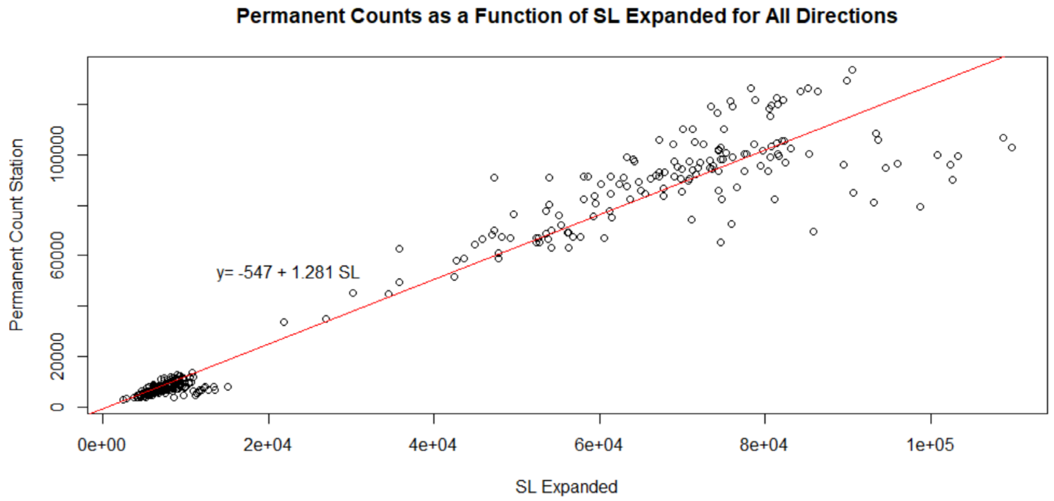


Figure 5-6 PCS Counts vs. SL Expanded fitted regression model for All Directions using daily volumes

The performance of the regression models developed based on SL Index to estimate the daily volumes are compared to the ground truth data collected from the PCS. Table 5-4 shows the MAPE computed for the training data and testing data for the two methods (expansion of SL Index based on local volume data, and the original SL Expanded volume data). As shown in Table 5-4, the regression models developed using the SL Index (for both, each specific direction, and all directions) as an explanatory variable generated low MAPE in each direction. However, the regression models developed using the SL Index in specific directions outperformed the all-direction model in the Flagler Bridge location. For both the Flagler Bridge and the I-95 location, the direction-specific and all-direction models produced low MAPE values. When using the SL Expanded volumes, the MAPE for the testing data reached up to 36% on the I-95 testing location. The volumes estimated using the direction-specific regression models have MAPE values of 5% to 8%. The volumes estimated using the all-direction regression model have MAPE values of 5% to 9%. Manual aggregation of daily volume from the available SL Estimates produced better result than the AADT compared earlier. However, the regression models based on the SL index improved the daily volume predictions in the locations tested.

Table 5-4 MAPE of SL Index regression models for daily volume estimation

Used Model	Measure	Flagler Bridge EB	Flagler Bridge WB	I-95 NB	I-95 SB
SL Estimates (SL Expanded)	MAPE Training Data	22.28%	19.96%	17.62%	20.20%
	MAPE Testing Data	8.00%	17.00%	36.00%	29.00%
Direction-Specific Linear Regression based on SL Index	MAPE: Training Data	12.63%	12.88%	7.53%	8.14%
	MAPE: Testing Data	7.40%	8.50%	6.10%	5.35%
All-Direction Linear Regression based on SL Index	MAPE: Training Data	18.19%	16.05%	5.81%	9.01%
	MAPE: Testing Data	9.43%	7.58%	5.51%	8.87%

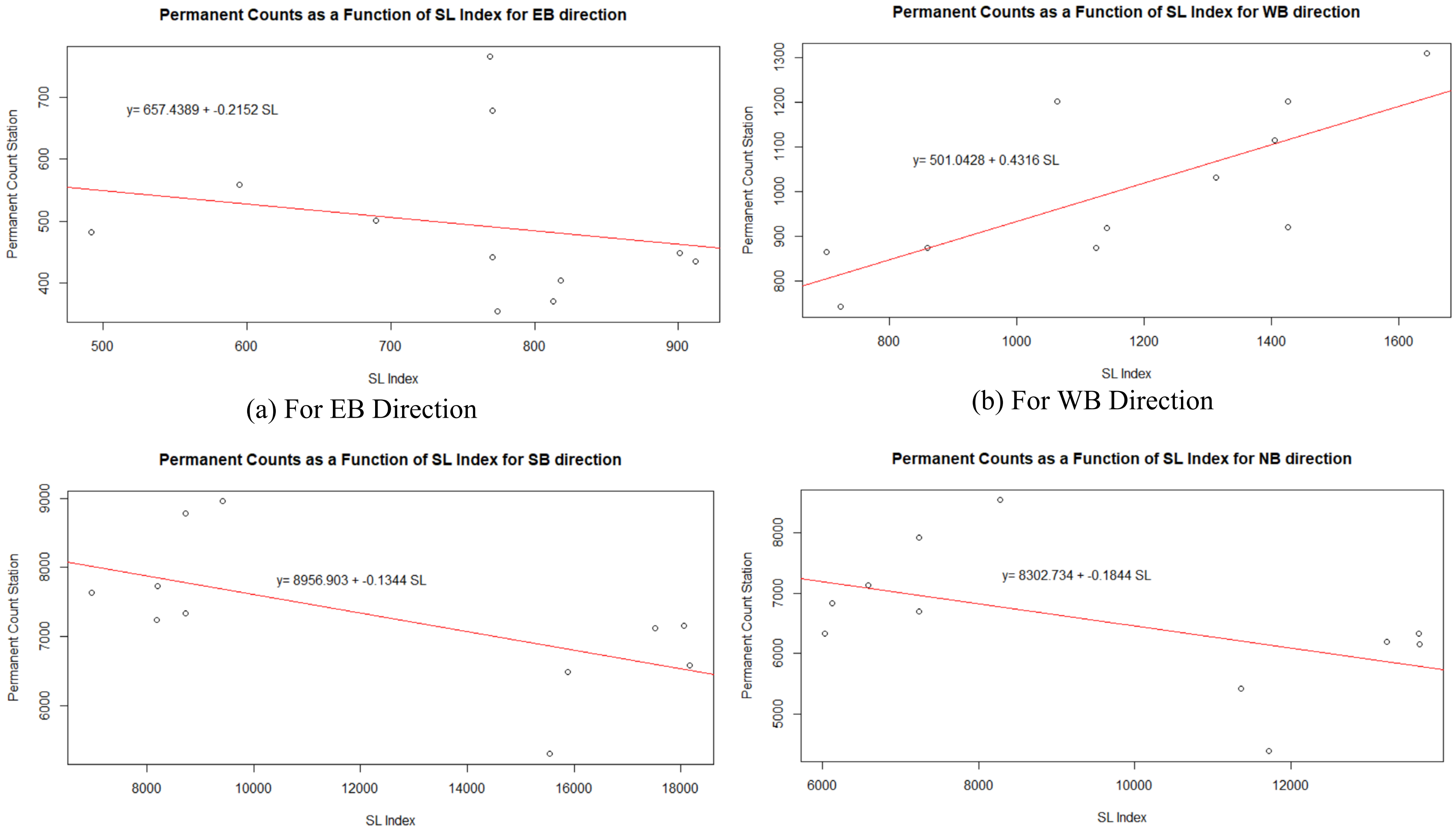
5.5. Estimating Hourly Volume at the PCS locations using Regression Analysis based on SL Index

Considering the improvement in the ADT estimates based on the developed regression models in the previous section, additional work was done to investigate the estimation of the hourly traffic volumes using the same method. Regression models were developed between the hourly SL Index and PCS hourly volumes. The SL and PCS are divided into training datasets, consisting of data for 11 months, and testing dataset, consisting of data in the remaining (one) month. The PCS values are used as the response or dependent variable, and the SL Index as the explanatory or independent variable. The hourly volumes (vph) for all of the days of the week are manually averaged over the whole month. For instance, volumes for all of the Mondays at 4:00 PM are averaged over a certain month.

This study investigated three different variations of the regression models using the hourly volume data. These variations are explained next. In the first variation (Scenario 1), the data is averaged over all Tuesdays (typical weekday) per month from 4:00 PM to 5:00 PM, resulting in a total of 11 data points for the training dataset. Tuesday is selected as a typical weekday, and the timing from 4:00 PM to 5:00 PM is selected as a typical peak hour. The fitted direction-specific regression models did not exhibit good performance as shown in Table 5-5, which summarizes the statistical inferences. The regression plots for Scenario 1 are included in Figure 5-7 and Figure 5-8 for further inferences. The All Directions model has an acceptable R^2 (63%), and low p-value compared to the direction-specific models.

Table 5-5 Regression Statistical Inferences for PCS Counts as a function of SL Index for Tuesday peak hour

Direction	Regression Equation	R^2	p-value
EB	PCS = 657.4389 – 0.2152SL	4.4%	0.540
WB	PCS = 501.0428 + 0.4316SL	55.6%	0.010
SB	PCS = 8956.9030 – 0.1344SL	37.0%	0.050
NB	PCS = 8302.7340 – 0.1844SL	27.0%	0.100
All Directions	PCS = 1203.0 + 0.433SL	62.8%	0.000



(a) For EB Direction

(b) For WB Direction

(c) For SB Direction

(d) For NB Direction

Figure 5-7 PCS Counts vs. SL Index fitted regression model using hourly volumes (Scenario 1)

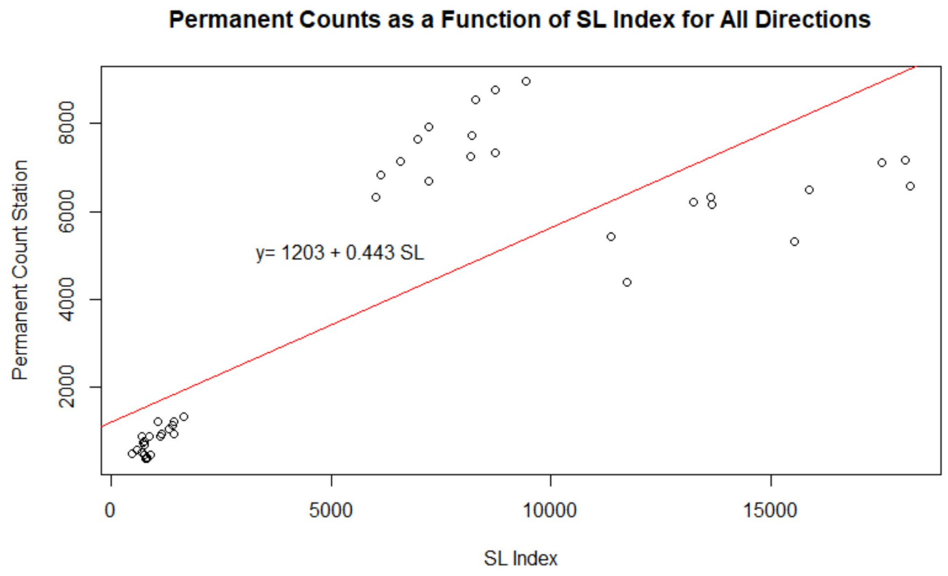
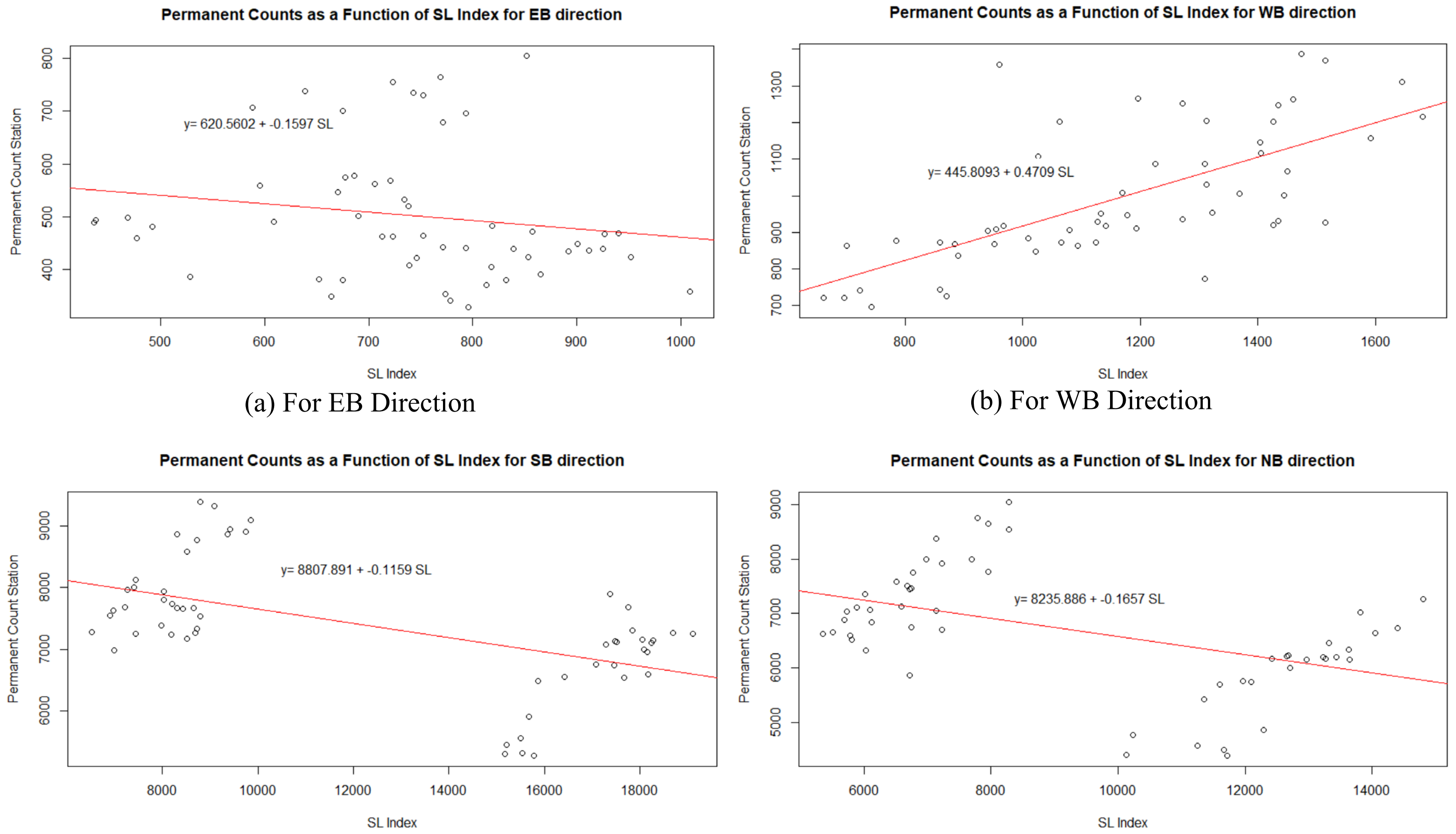


Figure 5-8 PCS Counts vs. SL Index fitted regression model for All Directions using hourly volumes (Scenario 1)

In the second variation (Scenario 2), the hourly volume data for every weekday is averaged over the whole month from 4:00 to 5:00 PM, resulting in a total of 55 data points for the training dataset. The reason weekdays (Monday through Friday) are chosen to perform this analysis is to observe the change in the SL Index data when a larger dataset is considered. Table 5-6 shows that the per direction R^2 values are low, while the R^2 and p-value for the All-Directions model is acceptable. Figure 5-9 and Figure 5-10 exhibit more on the fitted regression models for Scenario 2.

Table 5-6 Regression statistical inferences for PCS Counts as a function of SL Index for Weekday average peak hour

Direction	Regression Equation	R^2	p-value
EB	PCS = 620.5602 – 0.1597SL	2.9%	0.220
WB	PCS = 445.8093 + 0.4709SL	46.3%	0.000
SB	PCS = 8807.8910 – 0.1159SL	29.0%	0.000
NB	PCS = 8235.8860 – 0.1657SL	22.0%	0.000
All Directions	PCS = 1256.0000 + 0.447SL	62.8%	0.000



(a) For EB Direction

(b) For WB Direction

(c) For SB Direction

(d) For NB Direction

Figure 5-9 PCS Counts vs. SL Index fitted regression model using hourly volumes (Scenario 2)

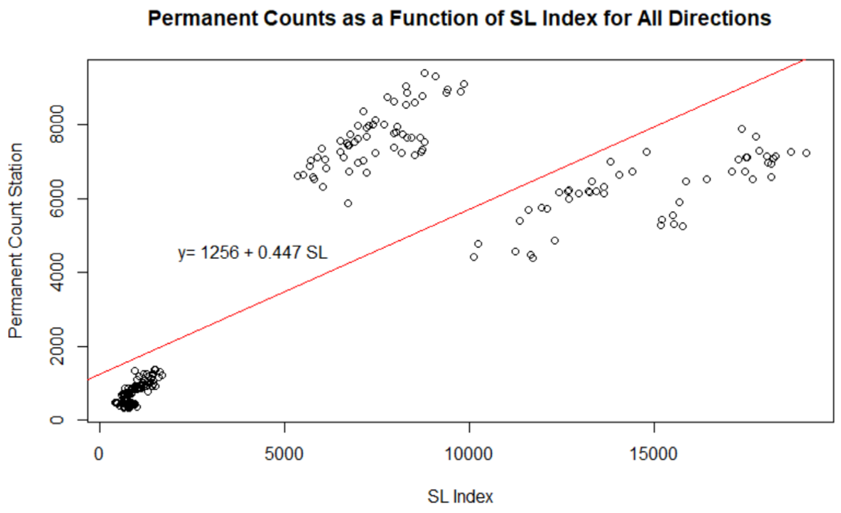
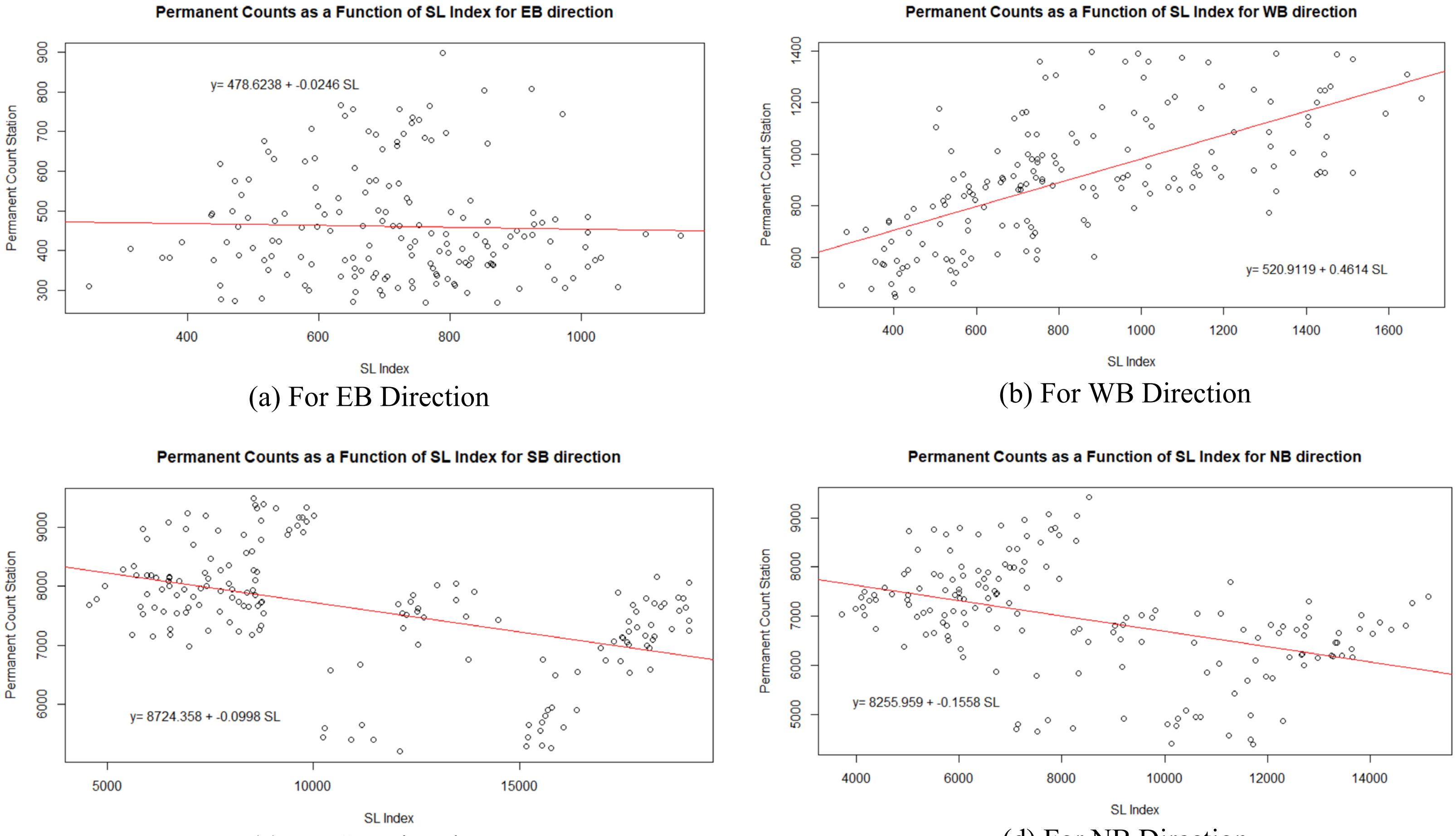


Figure 5-10 PCS Counts vs. SL Index fitted regression model for All Directions using hourly volumes (Scenario 2)

In the third variation (Scenario 3), the analysis period is increased by two hours from the previous scenario to capture more of the peak period. As a result, the weekdays (Monday through Friday) are averaged over the whole month from 4:00 to 7:00 PM, resulting in a total of 165 data (5 days, 3-hour slots, 11 months) points for the training dataset. As was the case with the other two variations, the per direction R^2 values are low, while the R^2 and p-value for the All-Directions model is acceptable, as shown in Table 5-7. Figure 5-11 and Figure 5-12 exhibit the regression plots for the Scenario 3 models to provide more inference to the statistical result.

Table 5-7 Regression statistical inferences for PCS Counts as a function of SL Index for Weekday average peak period

Direction	Regression Equation	R^2	p-value
EB	PCS = 478.6238 – 0.0246SL	0.1%	0.700
WB	PCS = 520.9119 + 0.4614SL	43.0%	0.000
SB	PCS = 8724.3580 – 0.0998SL	21.1%	0.000
NB	PCS = 8255.9590 – 0.1558SL	18.6%	0.000
All Directions	PCS = 1359.0 + 0.493SL	61.4%	0.000



(a) For EB Direction

(b) For WB Direction

(c) For SB Direction

(d) For NB Direction

Figure 5-11 PCS Counts vs. SL Index fitted regression model using hourly volumes (Scenario 3)

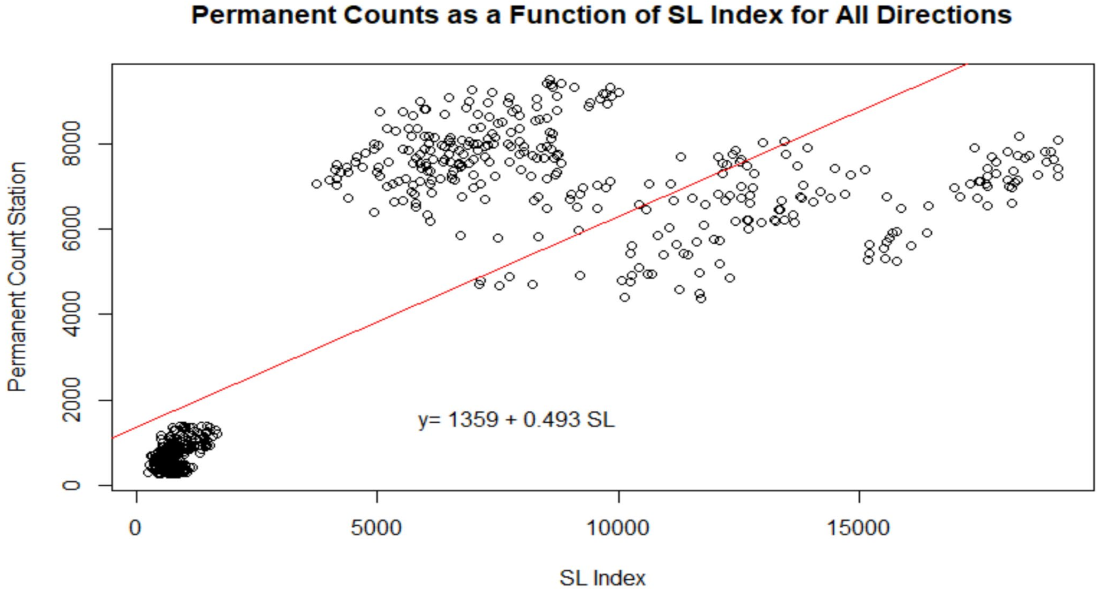


Figure 5-12 PCS Counts vs. SL Index fitted regression model for All Directions using hourly volumes (Scenario 3)

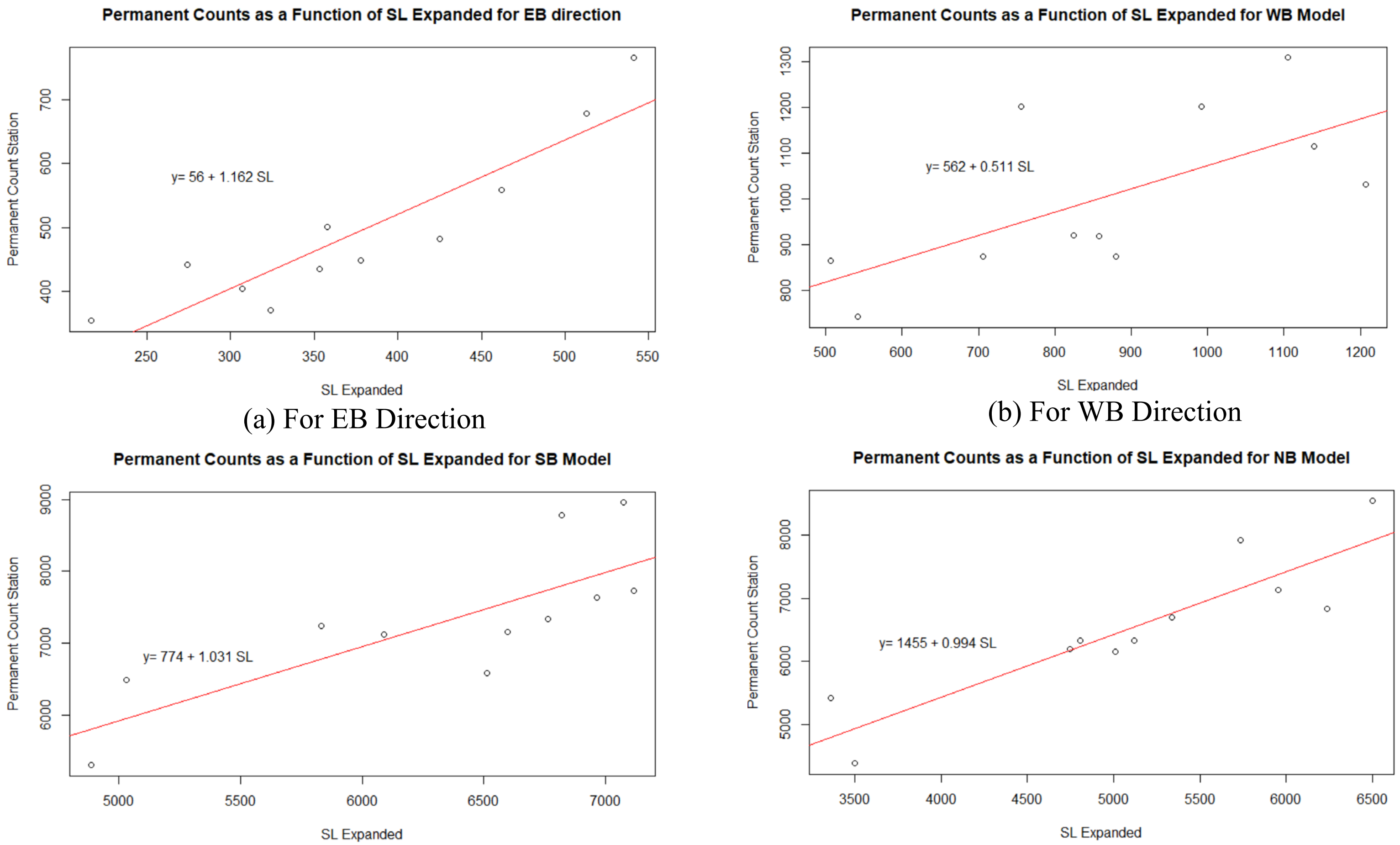
5.6. Estimating Hourly Volume at PCS locations using Regression Analysis based on SL Expanded Volumes

Given the poor prediction of the regression model between the hourly SL Index and hourly PCS, additional exploration is done this time to base the prediction on the hourly output of SL Expanded volumes corrected based on local data. Regression models were developed between the hourly SL Expanded volumes and PCS hourly volumes. The SL and PCS volume data are divided into training datasets, consisting of data for 11 months, and testing dataset, consisting of data in the remaining (one) month. The PCS values are used as the response or dependent variable, and the SL Expanded volumes are used as the explanatory or independent variable. The utilized data is in the form of hourly data for every day of the week, averaged over the whole month, resulting in seven data points for every hour in every month. Similar to the previous regression models, PCS hourly data for every day of the week is manually averaged over the whole month and used as the response variable. As is done in the previous section, this study tried three different variations of the regression models using the hourly volume data.

Contrary to the findings in the model developed based on the SL Index reported in the previous section, the data points when plotting the SL Expanded volumes and PCS volumes exhibit a linear relationship. Table 5-8 summarizes the statistical results obtained from the regression model. According to p-values, all models are significant at a 1% level of significance. Also, the R^2 values are high (i.e., greater than 60% for all models), except for the WB model. The high R^2 values indicate a high predictability of the variations in the PCS data. Similar to the previous models, the All Directions model resulted in the highest R^2 value of approximately 98%. Regression plots and models are shown in Figure 5-13 and Figure 5-14 for further inferences.

Table 5-8 Regression statistical inferences for PCS Counts as a function of SL Expanded volumes for the Tuesday peak hour

Direction	Regression Equation	R^2	p-value
EB	PCS = 56.0 + 1.162SL	83.1%	0.000
WB	PCS = 562.0 + 0.511SL	43.6%	0.027
SB	PCS = 774.0+1.031SL	62.9%	0.004
NB	PCS = 1455.0 + 0.994SL	80.0%	0.000
All Directions	PCS = 86.0+ 1.181SL	97.7%	0.000



(a) For EB Direction

(b) For WB Direction

(c) For SB Direction

(d) For NB Direction

Figure 5-13 PCS Counts vs. SL Expanded fitted regression model using hourly volumes (Scenario 1)

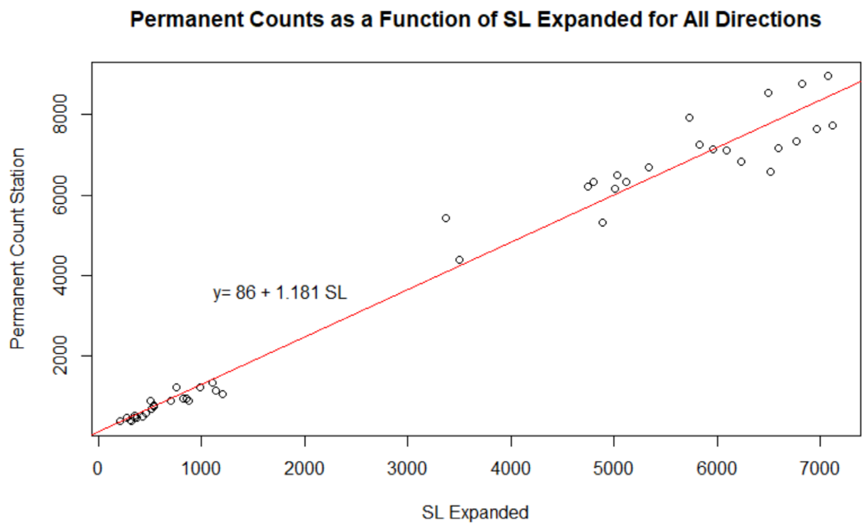
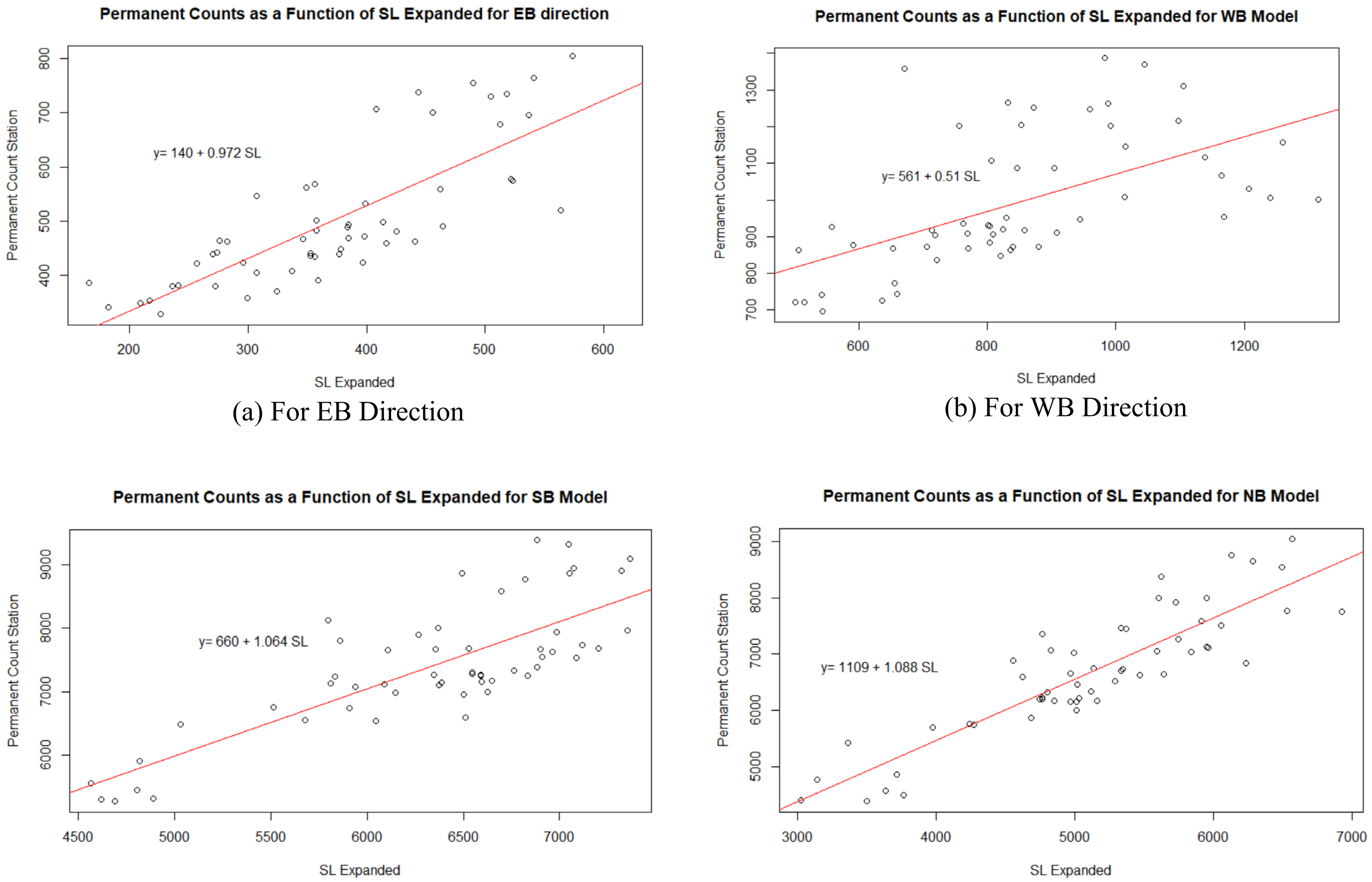


Figure 5-14 PCS Counts vs. SL Expanded fitted regression model for All Directions using hourly volumes (Scenario 1)

As shown in Table 5-9 and Table 5-10 similar pattern of R^2 values were obtained when conducting the regression analysis based on the hourly SL Expanded volumes using weekday average peak hour and weekday average peak period volumes. Although the models produced lower R^2 values compared to SL Expanded Volumes for Tuesday Peak Hour, the All Directions models for all cases show very accurate predictability of variations in volumes. Regression plots and models for Scenario 2 and Scenario 3 are shown in Figure 5-15, Figure 5-16, Figure 5-17 and Figure 5-18 for further inferences.

Table 5-9 Regression statistical inferences for PCS Counts as a function of SL Expanded volumes for Weekday average peak hour

Direction	Regression Equation	R^2	p-value
EB	PCS = 140.0 + 0.972SL	63.5%	0.000
WB	PCS = 561.0 + 0.51SL	32.2%	0.000
SB	PCS = 660.0+ 1.064SL	63.5%	0.000
NB	PCS = 1109.0 + 1.088SL	78.1%	0.000
All Directions	PCS = 78.0+ 1.204SL	97.7%	0.000



(a) For EB Direction

(b) For WB Direction

(c) For SB Direction

(d) For NB Direction

Figure 5-15 PCS Counts vs. SL Expanded fitted regression model using hourly volumes (Scenario 2)

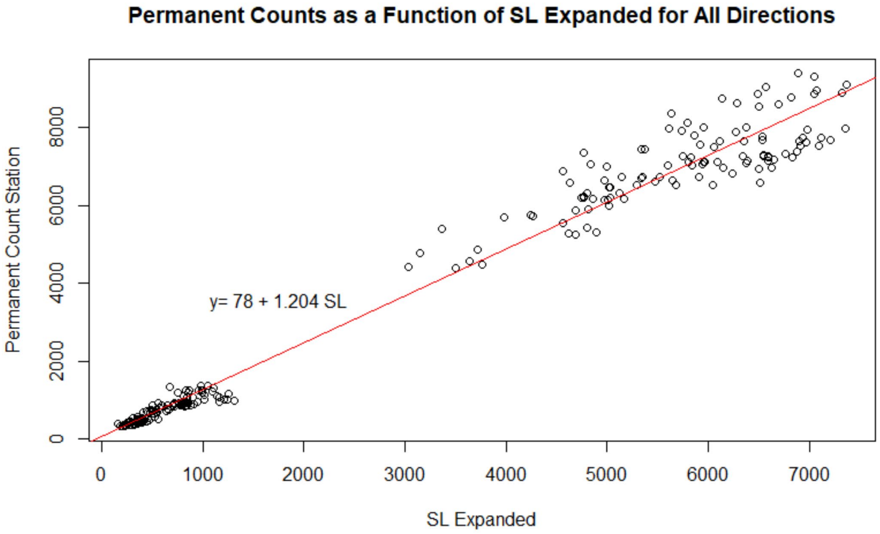
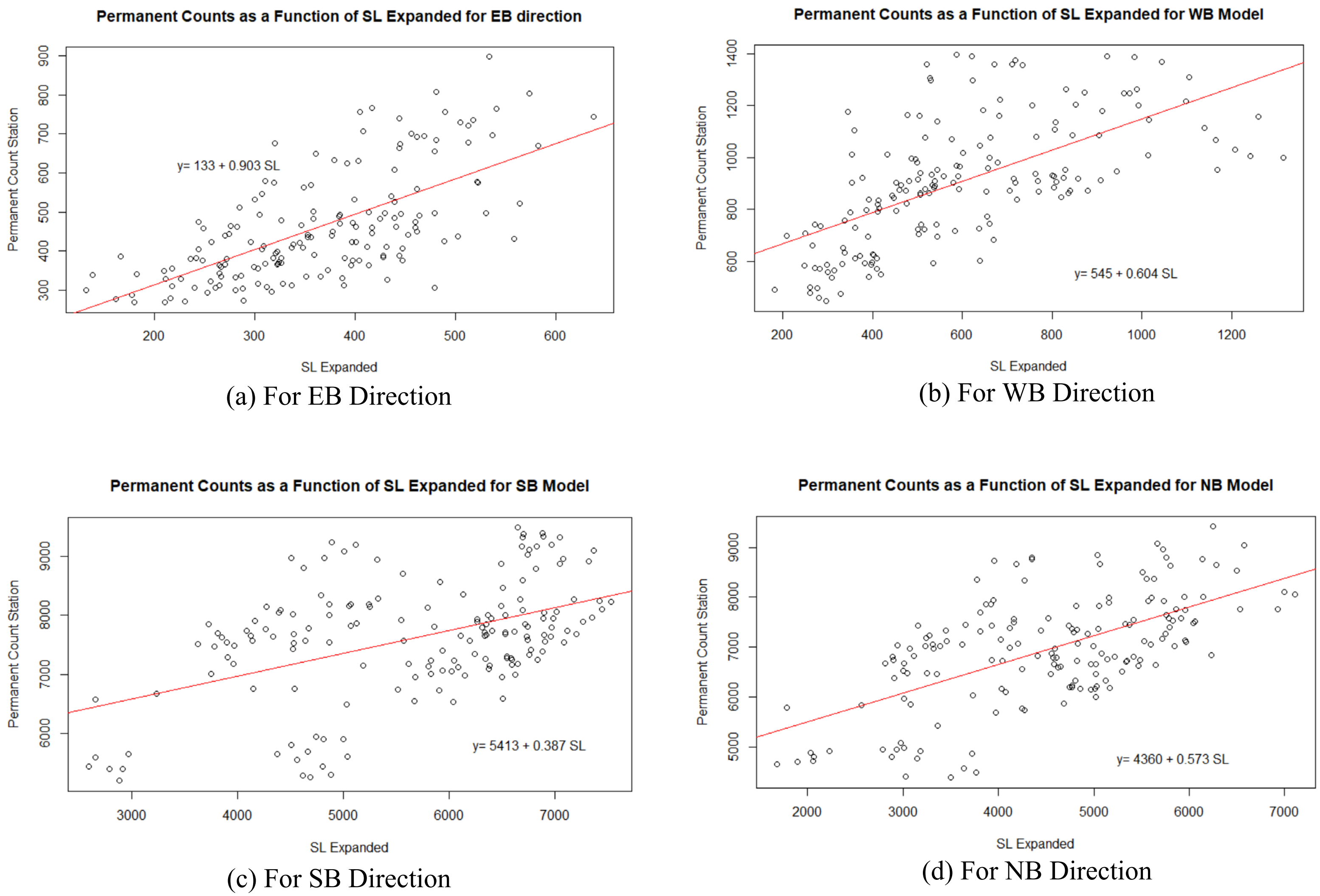


Figure 5-16 PCS Counts vs. SL Expanded fitted regression model for All Directions using hourly volumes (Scenario 2)

Table 5-10 Regression statistical inferences for PCS Counts as a function of SL Expanded volumes for Weekday average peak period

Direction	Regression Equation	R^2	p-value
EB	PCS = 133.0+ 0.903SL	42.8%	0.000
WB	PCS = 545.0 + 0.604SL	38.8%	0.000
SB	PCS = 5413.0 + 0.387SL	22.8%	0.000
NB	PCS = 4360.0 + 0.573SL	37.9%	0.000
All Directions	PCS = 370.0+ 1.298SL	90.7%	0.000



(a) For EB Direction

(b) For WB Direction

(c) For SB Direction

(d) For NB Direction

Figure 5-17 PCS Counts vs. SL Expanded fitted regression model using hourly volumes (Scenario 3)

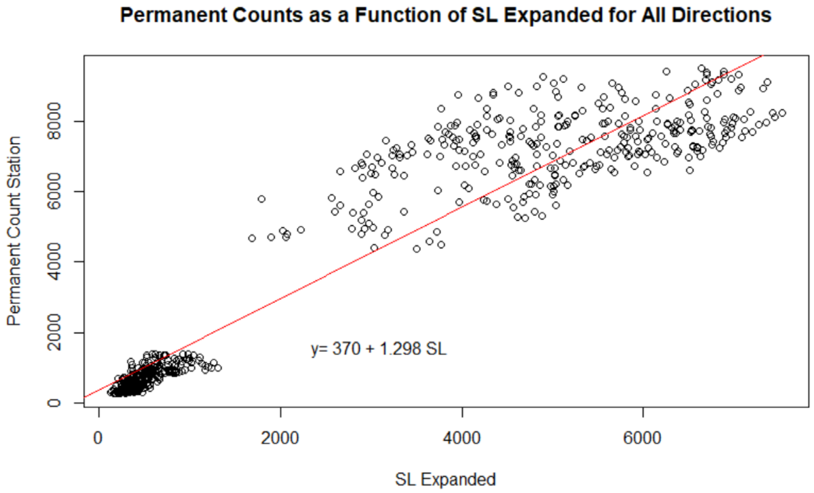


Figure 5-18 PCS vs SL Expanded fitted regression model for All Directions using hourly volumes (Scenario 3)

Based on the statistical results, the accuracy of the hourly volume estimation based on the regression model selected for the Tuesday Peak Hour (Scenario 1 of the SL Expanded models) was selected for comparison with the accuracy of the default SL Expanded data. The MAPE values of the models reported in Table 5-11 indicate how close the developed models are to the real-world data collected from the PCS. It is evident that the regression models developed from SL Expanded based on specific direction and all direction outperforms the default SL Expanded output. As indicated in Table 5-11, the deviation in the hourly volume for the SL Expanded volume ranges between 27% and 43% for the four directions in the testing dataset. Utilizing the direction-specific models drops these values to between 14% and 23%, while using the all-direction model dropped these values to between 13% and 25%.

Table 5-11 MAPE of different regression models for hourly volume estimation

Used Model	Measure	Flagler Bridge EB	Flagler Bridge WB	I-95 NB	I-95 SB
SL Expanded Hourly Aggregation	MAPE Training Data	23.65%	17.65%	21.86%	12.82%
	MAPE Testing Data	40.00%	43.00%	32.00%	27.00%
Direction-Specific Linear Regression based on SL Expanded (Tuesday Peak Hour)	MAPE: Training Data	9.52%	10.80%	6.08%	7.27%
	MAPE: Testing Data	22.00%	23.00%	8.00%	10.01%
All-Direction Linear Regression based on SL Expanded (Tuesday Peak Hour)	MAPE: Training Data	11.42%	18.49%	8.83%	8.88%
	MAPE: Testing Data	17.00%	25.00%	18.00%	13.00%

The MAPE values corresponding to the models using SL Expanded showed higher accuracy, which can be attributed to the fact that SL Expanded data undergoes machine learning processes that account for additional factors when estimating the volume. The expansion process also takes the PCS as reference, producing more realistic results at an hourly level.

5.7. Transferability of the Models Developed based on Combining SL and PCS Data

This section reports on testing the transferability of the volume estimation models developed based on combining SL and PCS data to other links in the network (other than the PCS links from which the data was collected to develop the model). If these models can provide accurate estimates for the volumes in the network, they can be used as an important source for estimating missing volume data in a network. As found earlier, regression models developed based on the hourly SL Expanded performed better compared to the model that uses the hourly SL Index. The transferability of the developed model to estimate the daily and hourly volumes based on PCS data and the SL data is tested for the intersection of Okeechobee Boulevard and Tamarind Avenue/Parker Avenue, which is different than the locations used in the development of the regression models. The benchmark data to assess the model's accuracy is in the form of hourly tube counts that are available at the investigated intersection for only one day, October 30, 2019, and are collected from the regional traffic management center of the FDOT. Accordingly, the SL data was retrieved for the same location, date, and time intervals, and are used to test the transferability. Unfortunately, data for more days are not available for this location. Thus, the testing results may be different if such data would have been available, allowing the use of the average volumes over multiple days in the testing phase.

To measure the accuracy of the transferability of the developed models, MAPE is computed between the estimated values and the ground truth data from the tube counts. In Table 5-12 to Table 5-15, the models based on Tuesday Peak Hour, Weekday Average Peak Hour, and Weekday Average Peak Period are referred to as Scenario 1, Scenario 2, and Scenario 3, respectively. The results in Table 5-12 and Table 5-13 show the results of estimating the volumes based on the models developed for each direction separately, whereas Table 5-14 and Table 5-15 include the results based on the All Directions model. Table 5-12 shows significantly low errors for the developed models, with the lowest errors resulting from the models developed based on the average of the Tuesday Peak Hour

(Scenario 1). This indicates that the regression models developed using the PCS and SL Expanded volumes for each direction separately exhibit very good transferability for various data aggregation levels, especially for the hourly aggregation. The reason behind the poorer transferability of daily aggregated data can be the difference in traffic conditions between the PCS locations (higher volume conditions) and the tube counts (lower volume conditions). The transferability for the SB and NB directions, which is not presented in this document, is very poor. These directions have lower volumes than the EB and WB directions.

In contrast, Table 5-13 shows higher errors for the hourly models, indicating that the models based on SL Index data do not provide good transferability compared to the models based on the SL Expanded. However, the models based on the SL Index resulted in lower errors for the daily volume estimation. Similar to the previous results, MAPE for the SB and NB directions is very high. As a result, the SB and NB directions were ignored for testing the transferability of the models.

Table 5-12 Testing Transferability of the direction-specific models developed based on PCS Counts and SL Expanded Volumes at the selected intersection

	MAPE	
Regression Model	**WB Approach**	**EB Approach**
SL Expanded Hourly Data – Scenario 1	3%	2%
SL Expanded Hourly Data – Scenario 2	3%	12%
SL Expanded Hourly Data – Scenario 3	9%	18%
SL Expanded Daily Data	25%	47%

Table 5-13 Testing Transferability of the direction-specific models developed based on PCS Counts and SL Index at the selected intersection

	MAPE	
Regression Model	**WB Approach**	**EB Approach**
SL Index Hourly Data – Scenario 1	11%	76%
SL Index Hourly Data – Scenario 2	8%	74%
SL Index Hourly Data – Scenario 3	6%	71%
SL Index Daily Data	9%	38%

The All Directions models developed using the SL Expanded volumes and SL Index resulted in higher MAPE than the models developed for each direction separately, as shown in Table 5-14 and Table 5-15. Similar to the results mentioned above, the models based on the SL Expanded volumes resulted in better transferability for the hourly volume estimation, while the SL Index produced better estimation for the daily volume estimation. It is evident that for the hourly volume estimation, estimates based on the direction-specific SL Expanded model exhibited very small errors for the WB and EB (3.0% and 2% errors, respectively). Similarly, for the daily volume estimation, the SL Index-based direction-specific model produced better results (9.0% for the WB and 38% for the EB). However, the All Directions model generated MAPE ranging from 14.0% to 22.0%. The estimates for the NB and SB exhibited large errors and thus they are not reported in this document.

Table 5-14 Testing Transferability of the model developed based on PCS Counts and SL Expanded at the selected intersection – All Directions

	MAPE	
Regression Model	**WB Approach**	**EB Approach**
SL Expanded Hourly Data – Scenario 1	22%	14%
SL Expanded Hourly Data – Scenario 2	26%	18%
SL Expanded Hourly Data – Scenario 3	37%	29%
SL Expanded Daily Data	38%	7%

Table 5-15 Testing Transferability of the model developed based on PCS Counts and SL Index at the selected intersection – All Directions

	MAPE	
Regression Model	**WB Approach**	**EB Approach**
SL Index Hourly Data – Scenario 1	28%	18%
SL Index Hourly Data – Scenario 2	32%	22%
SL Index Hourly Data – Scenario 3	44%	33%
SL Index Daily Data	16%	14%

5.8. Summary

The results presented in this chapter indicate that there are significant discrepancies between the AADT and MADT estimated based on the SL Expanded volume data and the PCS data used as ground truth data for the two locations. Large errors were also found when estimating the seasonal factors based on the SL data.

Considering the results above, this study investigated and refined the volume estimation based on SL data by developing regression models that relate the SL data to the ground truth PCS data. These models were then investigated for use to expand the SL data rather than using the SL Expanded data. It was possible in this study to develop well fitted regression models between the SL measures and the ground truth volumes with high R^2 values and acceptable significant levels. This was the case for both the daily and hourly volumes.

There are indications based on the results obtained in this study that the use of regression models can improve the link volume estimation for links with relatively high volumes. For example, when using the SL Expanded volumes, the MAPE for the testing data ranged between 10% and 16%, depending on the testing location. The volumes estimated using the direction-specific regression models have MAPE values of 10% to

13%. The volumes estimated using the All Directions regression model have MAPE values of 5% to 23%. The deviation in the hourly volume for the SL Expanded volume ranges between 27% and 43% for the four directions for the testing dataset. Utilizing the direction-specific models dropped these values to between 14% and 23% and using the All Directions model dropped these values to between 13% and 25%.

The transferability test of the volume estimation models developed based on PCS data to other locations was conducted based on a one-day data for one other location due to data limitation. Using data from multiple days from this location would have been better to make conclusions about the transferability. In general, the transferability test results indicate acceptable results for the directions with heavy volumes but high errors for the directions with lower volumes.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

This study proposes methods to use data from multiple sources to improve scenario-based analysis and MRM. The study develops a method to identify traffic scenarios and representative days considering different traffic patterns throughout the year. This research also proposes methods for integrating crowdsourced data into the OMDE process, which is required for MRM considering the modeled scenario. The study then checks the quality and transferability of the crowdsourced data for potential use in estimating segment-level volumes. The following subsections present the conclusions based on the results of this research and the recommendations for future work.

6.1. Summary and Conclusion

The selection of a representative day or a representative condition for the purpose of AMS and MRM to assess traffic performance requires clear methods and guidance, which are currently not available. Practitioners often adopt ad-hoc approaches to identify travel conditions to represent the real world and use traffic measures for these conditions such as traffic volume, speed, travel time, etc., as inputs to the model development and calibration process. This research studied the use of the K-means and GMM clustering methods with the aim of determining the variations in the temporal and spatial patterns based on key traffic measures Then, it compares the use of input variables aggregated for an entire facility and peak period versus using segregated input variables that are segregated in time and space. Four different segregation levels were examined: no segregation, spatial segregation, temporal segregation, and spatio-temporal segregation. In addition, the clustering was done with different numbers of clusters.

The assessment of the clustering results for a case study indicates that the K-means clustering algorithm with four clusters and spatiotemporal segregation level produced the best results. This conclusion is based on the assessment using the measures recommended in statistical measure literature (the t-SNE plot) and techniques based on traffic engineering literature. Utilizing no spatial segregations of the road segments, no temporal segregation of the peak period into intervals, and a lower number of clusters was less effective in clustering the data into distinctive patterns that account for the variations in traffic conditions along the roadway segments and within the peak period considering the day-to-day variations throughout the year. The study also showed that despite its theoretical advantage, the GMM clustering was less effective than the K-means clustering in identifying the traffic patterns in this study.

The results of this study clearly show that the use of an average day of the year or the peak season is not acceptable because it will not allow for an effective simulation model development and calibration. In addition to the fact that averaging volume and travel time data results in synthetic days that do not occur in the real world, such averaging results in diluted congestion levels. The analysis of the case study indicates that a large percentage of the days (the days in Clusters 1 and 4, which constitute about a third of the days) have more congestion levels than those of the averages. Thus, for example, the use of the averages for making highway designs may result in the under-design of the facilities.

This study demonstrated the values of measures based on traffic engineering literature, in addition to the commonly used statistical measures used to identify the quality of the clustering results. Measures identified in this study based on traffic engineering concepts including the Sum of Representative Day Distances, heat map and fundamental traffic diagram, were demonstrated to be critical in assessing the clustering result quality.

In order to achieve an accurate multi-resolution modeling (MRM) platform, a static traffic assignment-based ODME procedure combining crowdsourced data and an initial origin-destination (O-D) matrix from a regional demand forecasting model (SERPM 7.0) is utilized to improve the initial O-D matrix obtained from SERPM 7.0. After testing twelve different variations of the ODME processes and assessing the performance of the results, it was determined that Method 3(b) produced the best results. The method utilizes an initial ODME runs with an initial seed matrix from the SERPM 7.0 model to minimize the deviation from the initial O-D matrix and traffic counts. It then uses the resulting O-D matrix to estimate the trip generation from each zone but uses the SL index to estimate the trip distribution. Finally, a second ODME run is conducted to minimize the deviation from this resulting O-D matrix and the traffic counts. This method produced the best deviation from the counts and the crowdsourced data and slightly the worst deviation from the O-D matrix obtained.

An analysis was conducted to determine the quality and transferability of estimating segment-level data based on crowdsourced data. The results indicate that the default SL Expanded volumes without correction or calibration based on local data do not provide good estimates of the daily and hourly volumes. Large errors were also found when estimating the seasonal factors based on the SL data. Considering the results above, this study investigated and refined the volume estimation based on SL data by developing regression models that relate the SL data to the ground truth PCS data. These models were then investigated for use to expand the SL data rather than using the SL Expanded data. It was possible in this study to develop well fitted regression models between the SL measures and the ground truth volumes with high R^2 values and acceptable significant levels. This was the case for both the daily and hourly volumes.

There are indications based on the results obtained in this study that the use of regression models can improve the link volume estimation for links with relatively high volumes. However, further research is needed to investigate the results for additional case studies. The transferability test of the volume estimation models developed based on PCS data to other locations was conducted based on a one-day data for one other location due to data limitation. Using data from multiple days from this location would have been better to make conclusions about the transferability. In general, the transferability test results indicate acceptable results for the directions with heavy volumes, but high errors for the directions with lower volumes.

6.2. Research Contributions

Analysts or practitioners scope, develop, analyze, and calibrate simulation models to existing travel conditions and validate them to emulate real-life scenarios within the transportation network. Accuracy of AMS effectively depends on checking data and model quality. The multi-scenario and MRM methodology developed in this study using spatio-temporal data from detectors and crowdsourced platform paved the way for identifying travel patterns and improving the quality of the O-D matrix required for accurate multi-scenario analysis and MRM practices.

This study proved that using clustered data in modeling can have a significant impact on improving the analysis and thus the decisions made based on the analysis compared to existing practices. It will be up to the agencies to choose which representative day to model based on the clustering results, given that there are multiple identified representative days (one for each resulting cluster). The analyst can examine the number of days in each cluster and the congestion level in the representative day to determine which day to model.

The O-D matrices from the demand models only consider the results from demand surveys, which do not consider the network operations and capacity constraints. The results in this study indicate that the use of the regional demand model always overestimates the number of trips. This study also developed a unique method to integrate crowdsourced data with the existing O-D matrix from the regional planning model to estimate an accurate O-D matrix through the ODME process. As a result, the developed method provides agencies with accurate O-D matrices to conduct MRM. Additionally, the transferability and reliability of the crowdsourced data from a third-party vendor (i.e., Streetlight) explored in this study created an opportunity for agencies to utilize crowdsourced data as an accurate source of network-wide traffic information for AMS purposes.

6.3. Recommendations for Future Studies

A number of research topics can be recommended to extend the research of this study, as listed below:

1. The methodology to support scenario-based analysis and the associated representative days presented in this study can be expanded to include the operational conditions for eventful days (i.e., days with incident, work zone, weather events).
2. The methodology to support scenario-based analysis and the associated representative days be extended to arterial networks and multi-facility networks, including use in MRM analysis.
3. The ODME process developed by integrating crowdsourced data with a regional planning model can be used as the basis of an MRM framework. Further refinement based on integrating crowdsourced data from other sources, partial counts, and turn movement counts can be explored to develop an O-D matrix similar to real-world conditions.
4. The ODME procedure used in this study is based on static assignment. Further assessment is need for the advantage of using dynamic traffic assignment-based ODME.
5. Data accuracy and transferability of other crowdsourced data sources from other vendors can be checked using the methods developed in this study. Cross-validation of the results obtained based on data from different vendors will be extremely powerful in validating the quality of the resulting matrices.

6. Investigation is needed for the use of crowdsourced data to validate the utilized paths between the origins and destinations.
7. Further analysis is needed for additional case study networks with better ground truth data availability to confirm the quality and transferability of models developed to expand the measured mobile data to segment-level volumes using regression analysis or machine learning.

REFERENCES

1. Al Mamun, Abdullah, Raihanul B. Tanvir, Masrur Sobhan, Kalai Mathee, Giri Narasimhan, Gregory E. Holt, and Ananda M. Mondal. (2021). Multi-Run Concrete Autoencoder to Identify Prognostic lncRNAs for 12 Cancers. International Journal of Molecular Sciences 22, no. 21: 11919. https://doi.org/10.3390/ijms22211191.

2. Alibabai, H., & Mahmassani, H. S. (2008). Dynamic Origin-Destination Demand Estimation Using Turning Movement Counts. Transportation Research Record, 2085(1), 39-48.

3. Al-Kaisy, A., & Huda, K. T. (2022). Empirical Bayes Application on Low-Volume Roads: Oregon Case Study. Journal Of Safety Research, 80, 226-234.

4. Alvarez, P. (2012). A Methodology to Estimate Time Varying User Responses to Travel Time and Travel Time Reliability in a Road Pricing Environment. Doctoral Dissertation, Florida International University.

5. Antoniou, C., Ben-Akiva, M., & Koutsopoulos, H. N. (2004). Incorporating Automated Vehicle Identification Data Into Origin-Destination Estimation. Transportation Research Record, 1882(1), 37-44.

6. Army Modeling and Simulation Office. (2020). Modeling and Simulation Glossary (web page). https://www.ms.army.mil/library2/glossary.html

7. Asakura, Y., Hato, E., & Kashiwadani, M. (2000). Origin-Destination Matrices Estimation Model Using Automatic Vehicle Identification Data and its Application to The Han-Shin Expressway Network. Transportation, 27(4), 419-438.

8. Ashok, K. (1996). Estimation And Prediction of Time-Dependent Origin-Destination Flows. Doctoral dissertation, Massachusetts Institute of Technology.

9. Azimi, M., & Zhang, Y. (2010). Categorizing Freeway Flow Conditions by Using Clustering Methods. Transportation Research Record, 2173(1), 105-114.

10. Barceló, J. (Ed.). (2010). Fundamentals of Traffic Simulation (Vol. 145, p. 439). New York: Springer.

11. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

12. Brownlee, J. (2019). Probabilistic Model Selection with AIC, BIC, and MDL. Probability; Machine Learning Mastery: San Francisco, CA, USA.

13. Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC In Model Selection. Sociological Methods & Research, 33(2), 261-304.

14. Cantelmo, G., Cipriani, E., Gemma, A., & Nigro, M. (2014). An Adaptive Bi-Level Gradient Procedure for the Estimation Of Dynamic Traffic Demand. IEEE Transactions on Intelligent Transportation Systems, 15(3), 1348-1361.

15. Carrese, S., Cipriani, E., Mannini, L., & Nigro, M. (2017). Dynamic Demand Estimation and Prediction for Traffic Urban Networks Adopting New Data Sources. Transportation Research Part C: Emerging Technologies, 81, 83-98.

16. Cascetta, E., Inaudi, D., & Marquis, G. (1993). Dynamic Estimators of Origin-Destination Matrices Using Traffic Counts. Transportation Science, 27(4), 363-373.

17. Chang, G. L., & Wu, J. (1993). Recursive Estimation of Ime-Varying OD Flows From Traffic Counts in Freeway Corridors. In 72nd Annual Meeting of the Transportation Research Board, Washington, DC.

18. Chatzimparmpas, A., Martins, R. M., & Kerren, A. (2020). t-visne: Interactive Assessment And Interpretation of t-SNE Projections. IEEE Transactions On Visualization and Computer Graphics, 26(8), 2696-2714.

19. Chen, H., Yang, C., & Xu, X. (2017). Clustering Vehicle Temporal and Spatial Travel Behavior Using License Plate Recognition Data. Journal of Advanced Transportation, 2017.

20. Chen, M., & Chien, S. I. (2001). Dynamic Freeway Travel-Time Prediction With Probe Vehicle Data: Link Based Versus Path Based. Transportation Research Record, 1768(1), 157-161.

21. Corradino Group. (2013). Tidewater Region External Origin and Destination Study. Virginia Department of Transportation, Richmond.

22. De Palma, A., & Marchal, F. (2002). Real Cases Applications of the Fully Dynamic METROPOLIS Tool-Box: An Advocacy For Large-Scale Mesoscopic Transportation Systems. Networks and Spatial Economics, 2(4), 347-369.

23. Demissie, M. G., Phithakkitnukoon, S., Sukhvibul, T., Antunes, F., Gomes, R., & Bento, C. (2016). Inferring Passenger Travel Demand to Improve Urban Mobility in Developing Countries Using Cell Phone Data: A Case Study Of Senegal. IEEE Transactions on Intelligent Transportation Systems, 17(9), 2466-2478.

24. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

25. Dixon, M. P., & Rilett, L. R. (2000). Real-Time Origin-Destination Estimation Using Automatic Vehicle Identification Data. In Transportation Research Board 79th Annual Meeting.

26. Dobrota, N., Stevanovic, A., & Mitrovic, N. (2022). Modifying Signal Retiming Procedures and Policies by Utilizing High-Fidelity Modeling with Medium-Resolution Traffic Data. Transportation Research Record, 2676(3), 660-684.

27. Dowling, R., Skabardonis, A., Halkias, J., McHale, G., & Zammit, G. (2004). Guidelines For Calibration Of Microsimulation Models: Framework And Applications. Transportation Research Record, 1876(1), 1-9.

28. FHWA, (2013a). Analysis, Modeling, and Simulation (AMS) Testbed Framework for Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management (ATDM) Programs, Federal Highway Administration, U.S. Department of Transportation.

29. FHWA. (2013b). Analysis, Modeling, and Simulation (AMS) Testbed Requirements for Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management (ATDM) Programs, Federal Highway Administration, U.S. Department of Transportation.

30. Florian, M., M. Mahut, and N. Tremblay. (2001). A Hybrid Optimization Mesoscopic Simulation Dynamic Traffic Assignment Model. IEEE Intelligent Transportation Systems, 2001, Proceedings.

31. FSUTMSOnline. “Model Download: Statewide and Regional Models” Web Page. https://www.fsutmsonline.net/index.php?/model_pages/modD44/index/.

32. Georgia Department of Transportation. (2016). Existing Volume Development and Origin-Destination Data. Downtown Connector Study. http://www.dot.ga.gov/BuildSmart/Studies/Documents/DowntownConnector/DataReport.pdf.

33. Geroliminis, N., & Sun, J. (2011). Properties of a Well-Defined Macroscopic Fundamental Diagram for Urban Traffic. Transportation Research Part B: Methodological, 45(3), 605-617.

34. Granato, S. (2017). Various Uses for INRIX/Streetlight Data: Ohio Plus Border Area. Ohio Department of Transportation (ODOT).

35. Hadi, M., Ozen, H., Shabanian, Sh., and Xiao, Y. (2012). Use of Dynamic Traffic Assignment in FSTUMS in Support of Transportation Planning in Florida. Final Report,

Prepared for Florida Department of Transportation, by the Florida International University Lehman Center for Transportation Research, Miami, FL.

36. Hadi, M., Xiao, Y., Wang, T., Qom, S. F., Azizi, L., Iqbal, M. S., … & Massahi, A. (2016). Framework For Multi-Resolution Analyses of Advanced Traffic Management Strategies.

37. Hadi, M., Xiao, Y., Zhan, C., & Alvarez, P. (2012). Integrated Environment for Performance Measurements and Assessment of Intelligent Transportation Systems Operations. Prepared for Florida Department of Transportation, by the Florida International University Lehman Center for Transportation Research, Miami, FL.

38. Hadi, M., Zhou, X., & Hale, D. (2022). Multiresolution Modeling for Traffic Analysis: Guidebook (No. FHWA-HRT-22-055). United States. Federal Highway Administration.

39. Hadi. M, Ozen, H, Shabanian, S, Xiao, Y, Zhao, W, Ducca, F. (2012). Use of Dynamic Traffic Assignment in FSUTMS in Support of Transportation Planning in Florida. Technical report submitted to Florida Department of Transportation.

40. Hans, E., Chiabaut, N., & Leclercq, L. (2014). Clustering Approach For Assessing the Travel Time Variability of Arterials. Transportation Research Record, 2422(1), 42-49.

41. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, And Prediction. Springer Science & Business Media.

42. He, X., Cai, D., Shao, Y., Bao, H., & Han, J. (2010). Laplacian Regularized Gaussian Mixture Model for Data Clustering. IEEE Transactions on Knowledge and Data Engineering, 23(9), 1406-1418.

43. Hobbs, F. D. (2016). Traffic Planning and Engineering: Pergamon International Library of Science, Technology, Engineering, and Social Studies. Elsevier.

44. Huang, Z. (1998). Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery, 2(3), 283-304.

45. Jamali, A (2014). O–D Demand Estimation Base on Automatic Vehicle Identification Data. Master Thesis, Sharif University Technology.

46. Kikuchi, S., & Tanaka, M. (2000). Estimating an Origin-Destination Table Under Repeated Counts of In-Out Volumes At Highway Ramps: Use of Artificial Neural Networks. Transportation Research Record, 1739(1), 59-66.

47. Ku, W. C., Jagadeesh, G. R., Prakash, A., & Srikanthan, T. (2016). A Clustering-Based Approach for Data-Driven Imputation of Missing Traffic Data. In 2016 IEEE Forum on Integrated and Sustainable Transportation Systems (FISTS) (pp. 1-6). IEEE.

48. Lin, P. W., & Chang, G. L. (2006). Modeling Measurement Errors and Missing Initial Values in Freeway Dynamic Origin–Destination Estimation Systems. Transportation Research Part C: Emerging Technologies, 14(6), 384-402.

49. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S. (2013). Understanding And Enhancement of Internal Clustering Validation Measures. IEEE Transactions On Cybernetics, 43(3), 982-994.

50. Lowry, M. (2014). Spatial Interpolation of Traffic Counts Based On Origin–Destination Centrality. Journal of Transport Geography, 36, 98-105.

51. Marutho, D., Handaka, S. H., & Wijaya, E. (2018). The Determination of Cluster Number at K-Mean Using Elbow Method and Purity Evaluation on Headline News. In 2018 International Seminar on Application for Technology of Information and Communication (pp. 533-538). IEEE.

52. Morshed, S. A., Arafat, M., Mokhtarimousavi, S., Khan, S. S., & Amine, K. (2021). 8R Resilience Model: A Stakeholder-Centered Approach of Disaster Resilience for Transportation Infrastructure and Network. Transportation Engineering, 4, 100058.

53. Morshed, S. A., Ahmed, K. M., Amine, K., & Moinuddin, K. A. (2021). Trend Analysis of Large-Scale Twitter Data Based on Witnesses during a Hazardous Event: A Case Study on California Wildfire Evacuation. World Journal of Engineering and Technology, 9(2), 229-239.

54. Mussone, L., & Matteucci, M. (2013). OD Matrices Network Estimation from Link Counts by Neural Networks. Journal of Transportation Systems Engineering and Information Technology, 13(4), 84-92.

55. Na, S., Xumin, L., & Yong, G. (2010). Research on K-Means Clustering Algorithm: An Improved K-Means Clustering Algorithm. In 2010 Third International Symposium on Intelligent Information Technology and Security Informatics (pp. 63-67). IEEE.

56. Nath, R. P. D., Lee, H. J., Chowdhury, N. K., & Chang, J. W. (2010). Modified K-Means Clustering for Travel Time Prediction Based on Historical Traffic Data. In International conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 511-521). Springer, Berlin, Heidelberg.

57. Netek, R., Pour, T., & Slezakova, R. (2018). Implementation of Heat Maps in Geographical Information System–Exploratory Study on Traffic Accident Data. Open Geosciences, 10(1), 367-384.

58. Ni, M., He, Q., & Gao, J. (2016). Forecasting the Subway Passenger Flow Under Event Occurrences with Social Media. IEEE Transactions on Intelligent Transportation Systems, 18(6), 1623-1632.

59. Park, B. B. (2002). Hybrid Neuro-Fuzzy Application in Short-Term Freeway Traffic Volume Forecasting. Transportation Research Record, 1802(1), 190-196.

60. Parry, K., & Hazelton, M. L. (2012). Estimation of Origin–Destination Matrices From Link Counts and Sporadic Routing Data. Transportation Research Part B: Methodological, 46(1), 175-188.

61. Pereira, F. C., Rodrigues, F., Polisciuc, E., & Ben-Akiva, M. (2015). Why So Many People? Explaining Nonhabitual Transport Overcrowding with Internet Data. IEEE Transactions on Intelligent Transportation Systems, 16(3), 1370-1379.

62. Platzer, A. (2013). Visualization of SNPs with t-SNE. PloS one, 8(2), e56883.

63. PTV Group. (2019). PTV Vissim 11 User Manual. Karlsruhe, Germany: PTV AG.

64. PTV Group. (2019). PTV Visum 2020 - Manual. Karlsruhe, Germany: PTV AG.

65. PTV Vision. (2013). VISUM 14 User Manual.

66. Rodrigues, J. G., Pereira, J. P., & Aguiar, A. (2017). Impact of Crowdsourced Data Quality on Travel Pattern Estimation. In Proceedings of the First ACM Workshop on Mobile Crowdsensing Systems and Applications (pp. 38-43).

67. Roll, J. (2019). Evaluating Streetlight Estimates of Annual Average Daily Traffic in Oregon (No. OR-RD-19-11).

68. Sanchez, Luis, Luis Muñoz, Jose Antonio Galache, Pablo Sotres, Juan R. Santana, Veronica Gutierrez, Rajiv Ramdhany et al. (2014). SmartSantander: IoT Experimentation Over a Smart City Testbed. Computer Networks 61, 217-238.

69. Shao, Yang, et al. (2020). Semi-analytical Solutions to the Lighthill-Whitham-Richards Equation with Time-Switched Triangular Diagrams: Application to Variable Speed Limit Traffic Control. IEEE Transactions on Automation Science and Engineering.

70. Shi, H., Yao, Q., Guo, Q., Li, Y., Zhang, L., Ye, J., ... & Liu, Y. (2020). Predicting Origin-Destination Flow via Multi-Perspective Graph Convolutional Network. In 2020 IEEE 36th International Conference on Data Engineering (ICDE) (pp. 1818-1821). IEEE.

71. Spiegelman, C., Park, E. S., & Rilett, L. R. (2010). Transportation Statistics and Microsimulation. CRC Press.

72. StreetLight Data. (2019). StreetLight Data's AADT 2018 Methodology and Validation White Paper. https://www.streetlightdata.com.

73. StreetLight Insight 2020 Whitepaper Version 1. (2020) https://www.streetlightdata.com/whitepapers/

74. Tostes, A. I. J., de LP Duarte-Figueiredo, F., Assunção, R., Salles, J., & Loureiro, A. A. (2013). From Data to Knowledge: City-Wide Traffic Flows Analysis and Prediction Using Bing Maps. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing (pp. 1-8).

75. Turner, S., & Koeneman, P. (2017). Using Mobile Device Samples to Estimate Traffic Volumes (No. MN/RC 2017-49). Minnesota. Dept. of Transportation. Research Services & Library.

76. Turner, S., Martin, M., Griffin, G., Le, M., Das, S., Wang, R., ... & Li, X. (2020). Exploring Crowdsourced Monitoring Data for Safety.

77. Turner, S., Tsapakis, I., & Koeneman, P. (2020). Evaluation of StreetLight Data's traffic count estimates from mobile device data (No. MN 2020-30). Minnesota. Dept. of Transportation. Office of Policy Analysis, Research & Innovation.

78. Van der Maaten, L., & Hinton, G. (2008). Visualizing Data Using t-SNE. Journal of Machine Learning Research, 9(11).

79. Van Der Zijpp, N. J. (1997). Dynamic Origin-Destination Matrix Estimation from Traffic Counts and Automated Vehicle Identification Data. Transportation Research Record, 1607(1), 87-94.

80. Vasudevan, M., & Wunderlich, K. (2013). Analysis, Modeling, and Simulation (AMS) Testbed Framework for Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management (ATDM) Programs (No. FHWA-JPO-13-095). United States. Department of Transportation. Intelligent Transportation Systems Joint Program Office.

81. Wang, H., & Song, M. (2011). Ckmeans. 1d. dp: Optimal K-Means Clustering In One Dimension by Dynamic Programming. The R journal, 3(2), 29.

82. Wei, C. H., & Lee, Y. (2007). Development of Freeway Travel Time Forecasting Models by Integrating Different Sources of Traffic Data. IEEE Transactions on Vehicular Technology, 56(6), 3682-3694.

83. Wu, C. H., Ho, J. M., & Lee, D. T. (2004). Travel-Time Prediction with Support Vector Regression. IEEE Transactions on Intelligent Transportation Systems, 5(4), 276-281.

84. Wunderlich, K. E., Vasudevan, M., & Wang, P. (2019). TAT Volume III: Guidelines for Applying Traffic Microsimulation Modeling Software 2019 Update to the 2004 Version (No. FHWA-HOP-18-036). United States. Federal Highway Administration.

85. Xia, J., & Chen, M. (2007). Defining Traffic Flow Phases Using Intelligent Transportation System-Generated Data. Journal of Intelligent Transportation Systems, vol. 11, no. 1, pp. 15–24.

86. Yang, H., Cetin, M., & Ma, Q. (2020). Guidelines for Using Streetlight Data for Planning Tasks (No. FHWA/VTRC 20-R23). Virginia Transportation Research Council (VTRC).

87. Zhang, K., Sun, D., Shen, S., & Zhu, Y. (2017). Analyzing Spatiotemporal Congestion Pattern on Urban Roads Based on Taxi GPS Data. Journal of Transport and Land Use, 10(1), 675-694.

88. Zhang, Z., Li, M., Lin, X., & Wang, Y. (2020). Network-Wide Traffic Flow Estimation with Insufficient Volume Detection and Crowdsourced Data. Transportation Research Part C: Emerging Technologies, 121, 102870.

89. Zhao, S., Guo, Y., Sheng, Q., & Shyr, Y. (2014). Advanced Heat Map And Clustering Analysis Using Heatmap3. Biomed Research International, 2014.

90. Zhou, X., & Mahmassani, H. S. (2006). Dynamic Origin-Destination Demand Estimation Using Automatic Vehicle Identification Data. IEEE Transactions on Intelligent Transportation Systems, 7(1), 105-114.

91. Zhou, X., Hadi, M., & Hale, D. K. (2021). Multiresolution Modeling for Traffic Analysis: State-of-Practice and Gap Analysis Report (No. FHWA-HRT-21-082). United States. Federal Highway Administration.

92. Zhou, X., Qin, X., & Mahmassani, H. S. (2003). Dynamic Origin-Destination Demand Estimation with Multiday Link Traffic Counts for Planning Applications. Transportation Research Record, 1831(1), 30-38.

93. Zhu, F., & Li, L. (2010). An Optimized Video-Based Traffic Congestion Monitoring System. In 2010 Third International Conference on Knowledge Discovery and Data Mining (pp. 150-153). IEEE.

VITA

SYED AHNAF MORSHED

EDUCATION AND AWARDS

2011-2014	B.Sc., Civil Engineering Islamic University of Technology, Dhaka, Bangladesh
2017-2018	M.Sc., Civil Engineering University of Louisiana, Lafayette, Louisiana
2018-2021	Doctoral Candidate and Graduate Assisant Civil Engineering – Transportation Florida International University, Miami, Florida
2021-2022	Doctoral Candidate and Dissertation Year Fellow (DYF) Civil Engineering – Transportation Florida International University, Miami, Florida

Outstanding Leadership Award, President of Institute of Transportation Engineers (ITE) Student Chapter, Florida International University, 2022

FLPRITE District Outstanding Student Chapter Award, Florida Puerto Rico ITE District, 2022

Lifesavers Traffic Safety Scholar, Chicago, 2022

Dissertation Year Fellowship, Florida International University, Spring and Summer 2022

People's Choice Award, STRIDE Student's Poster Showcase Competition, 100th TRBAM, 2020

Graduate & Professional Student Committee (GPSC) Conference Funding Award, University Graduate School, Florida International University, 2020 and 2022

Annual Book Scholarship, Gold Coast ITE, 2020

Outstanding Contribution Award, Assistant Logistics Chair, ITE Student Leadership Summit, Florida International University, 2020

Best Project, CENNOVATION, Bangladesh, 2014

Winner of PTAK Prize, Certified Supply Chain Analyst (CSCA) by ISCEA, Bangladesh Chapter, 2013

2021-2022	President, FIU ITE Student Chapter
2020-2021	Vice President, FIU ITE Student Chapter
2019-2020	Treasurer, FIU WTS Student Chapter
2019-2020	Vice President, FIU Bangladesh Student Organization
2017-2018	Secretary, International Student's Council, UL
2018-Present	American Society of Civil Engineers (ASCE), Member
2019-Present	Institute of Transportation Engineers (ITE), Member

PUBLICATIONS AND PRESENTATIONS

Morshed, Syed Ahnaf, Mohammed Hadi, Virginia P. Sisiopiku. "A Novel Multi-Agent Based Simulation Study on the Extension of Metrorail in Miami Beach Region". Presented at the 7th Annual UTC Conference for the Southeastern Region, March 2022.

Mamun, Md Mahmud Hasan, Kamar Amine, Mohammed Hadi, Syed Ahnaf Morshed and Thomas Hill, "Comparison of the Use of Fixed Targets and Varying Targets in the Calibration of Traffic Simulation Models". Presented at the 101st Annual Meeting of the Transportation Research Board, January 2022

Sisiopiku, Virginia P., Syed Ahnaf Morshed, Sahila Sarjana, and Mohammed Hadi. "Transportation Users' Attitudes and Choices of Ride-Hailing Services in Two Cities with Different Attributes." Journal of Transportation Technologies 11, no. 2 (2021): 196-212.

Morshed, Syed Ahnaf, Arkabrata Sinha, Qian Zhang, and Jovan Tatar. "Hygrothermal conditioning of wet-layup CFRP-concrete adhesive joints modified with silane coupling agent and core-shell rubber nanoparticles." Construction and Building Materials 227 (2019): 116531.