11-30-2020

Methodologies to Quantify Transit Performance Metrics at the System-Level Using High-Resolution GPS, Stop-Level, and GTFS Archived Transit Data

Travis Bradley Glick
Portland State University

Recommended Citation

Glick, Travis Bradley, "Methodologies to Quantify Transit Performance Metrics at the System-Level Using High-Resolution GPS, Stop-Level, and GTFS Archived Transit Data" (2020). *Dissertations and Theses.* Paper 5623.
https://doi.org/10.15760/etd.7495

Methodologies to Quantify Transit Performance Metrics at the System-Level

Using High-Resolution GPS, Stop-Level, and GTFS Archived Transit Data

by

Travis Bradley Glick

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Civil and Environmental Engineering

Dissertation Committee:
Miguel A. Figliozzi, Chair
Avinash Unnikrishnan
Christopher M. Monsere
Robert L. Fountain

Portland State University
2020

ABSTRACT

Performance metrics have typically focused at two main scales: a microscopic scale that focuses on specific locations, time-periods, and trips; and, a macroscopic scale that averages metrics over longer times, entire routes, and networks. When applied to entire transit systems, microscopic methodologies often have computational limitations while macroscopic methodologies ascribe artificial uniformity to non-uniform analysis areas. These limitations highlight the need for a middle approach.

This dissertation presents a mesoscopic analysis based around timepoint-segments, which are a novel application of an existing system for many transit agencies. In the United States, fix-route transit is typically defined by a small subset of bus stops along each route, called timepoints. For this research, routes are divided into a consecutive group of bus stops with one timepoint at the center. Each timepoint-segment includes all data collected in that segment during one hour of operation.

The utilized data sources are widespread and generally available to transit agencies. A methodology for merging and cleaning the data sources is proposed that: first, identifies broken data collection system to flag missing and inaccurate data; second, defines parameters of probability distributions, representative of specific locations, times, and routes, using sufficient statistics; and third, replaces flagged values with a random, but probabilistically representative value. The merged and stochastically cleaned data is aggregated by timepoint-segment to reduce subsequent computational requirements, yet maintains high granularly for statistical analysis after aggregation.

The results of linear and non-linear regressions for service durations, at and between bus stops, are presented and discussed. Independent variables were chosen based on previous published literature, but also included several updated classes of variables to provide comparisons for stop types, traffic signals, vehicle interactions, and time-of-day. The coefficients and performance of aggregated models are compared to previously published methods. The results show that factors identified at the microscopic scale (e.g. passenger movements, bus interactions at stops, travel times, travel speeds, unplanned stops, bus bunching, etc.), can be examined in aggregate without lost utility and without the heavy computation burden required to process large microscopic datasets, while also capturing double the variability in the data.

Visuals for congestion and headway performance, based on the aggregated datasets, are designed to examine transit performance along a route, between routes, and for specific segments. These visuals are a potentially useful tool for evaluating performance along routes and for identifying areas that may require a closer examination. Additionally, the methods are not computationally intensive and may be easily customized to examine specific locations, times, or feature sets.

The methodologies for data cleaning, regression modeling, and performance visuals, provide a foundation for how timepoint-segments may prove useful to researchers and agencies. The aggregated analysis reduces variability caused by singular atypical events, but still preserves enough detail for a robust statistical analysis. Overall, this approach improves realism, which is beneficial for evaluating the key trade-offs ridership, service, accessibility, and costs. Mesoscopic performance measures may help to understand relationship between key factors influencing transit operations, evaluate uncertainty, examine variations in service, determine points sensitive to disruption, quantify congestion costs for users and agencies, and compare travel patterns between different routes, days of the week, and peak versus off-peak travel.

Keywords

Transit, Performance Metrics, Network-Level, Aggregated Analysis, Mesoscopic

DEDICATION

To my wife, Lucille Glick.

ACKNOWLEDGEMENTS

This dissertation is the culmination of a decade of work at Portland State University. I had the privilege of following a continuous path from undergraduate to today due to the support, financially and emotionally, of more people and organizations than I could ever have imagined. I am truly grateful and deeply humbled by the opportunities that I have been given.

First, I must recognize and thank Dr. Miguel Figliozzi, my advisor and committee member: your guidance made this dissertation possible. We began working together six years ago and I am thankful for your advice, feedback, deadlines, reminders, and financial support. Above all, I am thankful for your ongoing patience as I learned and grew, both professionally and personally.

To members of the dissertation committee: Dr. Robert Fountain, Dr. Chris Monsere, and Dr. Avinash Unnikrishnan. This past year has been unorthodox and I am grateful to you each of you for agreeing to work with me. Thank you for continuing to do so; and, for navigating this new landscape of remote work and shifting schedules.

The datasets used for this research were graciously provided by TriMet staff members: Steve Calls and Miles Crumbly. I am very thankful to Miles for our regular meetings and data handoffs, as well has his answers to my countless questions.

To all those have I have worked with at Portland State on homework, studying, papers, projects, reports, and everything in-between, thank you. I must also thank the staffs of the Civil and Environmental Engineering Department and Computer Action Team. It is their efforts that keep the school running and the amount of help they have provided to me personally is without measure.

Financially, support has come from many sources over the years. I received funding over many years from the National Institute of Transportation and Communities (NITC), Transportation Research and Education Center (TREC), the Dwight David Eisenhower Transportation Fellowship Program, the Freight Mobility Research Institute (FMRI), and Portland State University scholarships and fellowships. Thank you, all.

Finally, to my wife and our family, thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 — OVERVIEW

1.1. Introduction

Public transit routes comprise a network that serves multiple, and often conflicting, objectives: maximize ridership, provide fast and reliable travel times, increase accessibility for disadvantaged individuals and communities, and reduce costs. The realization of these objectives requires both a baseline understanding of the factors affecting each objective and, perhaps more importantly, tools that can help policy makers evaluate the tradeoffs between the objectives.

The creation of performance metrics has therefore been a primary goal of ongoing transit research. Such metrics impact how evaluations are performed, the planning process, and future decisions, which subsequently impact transit reliability, travel times, travel speeds, operating costs, and system efficiency (Levinson, 1983). New tools for data collection and analysis have allowed for better and more informed decision making; yet, the ability to examine more aspects of the transportation system with higher detail results in a trade-off between the level of detail and the scope of each analysis. The development of performance measures is largely shaped by the data available for analysis and by the financial, logistical and computational complexities of its implementation.

1.2. General Background

The Federal Highway Administration recognizes transit benefits to include "reductions in highway congestion, air and noise pollution, energy consumption, and automobile accidents" (Nason & Williams, 2019). Unfortunately, transit ridership has declined in recent years for more than 85% of the United States Metropolitan Statistical Areas (MSA) (Siddiqui, 2018). Declining ridership is not the focus of this dissertation; rather, it serves as a backdrop for why new methodologies for performance metrics are important. This section provides context for some of the current pressures generally facing transit systems as well as examples specific to Portland, OR.

1.2.1. Transit Agencies

Transit agencies are often large organizations that typically change practices slowly and deliberately. New federal regulations requiring "performance data to inform decision-making and outcomes" provided two years for agencies to develop their plans while also providing trainings and ongoing resources (FTA, Office of Planning & Environment, 2016). As such, new methodologies must be clearly explained, be easy to implement, and use data and data-collection systems already available to agencies. While the adoption process for new systems is relatively slow, the decision for when and what to upgrade is largely based on what has proved useful to other agencies. For example, global position systems (GPS) technology onboard buses has opened new research objectives and useful methodologies have already been employed by agencies (Stoll, et al., 2016). While initially uncommon, an increasing number of agencies have invested in GPS data collection.

More generally, systems to collect and archive transit data are widespread and critical components of transit design and policy making. Tried-and-true data collection systems make up the core of transit data collection; operators, planners, and app makers apply well-established methodologies daily (Noch, 2019). However, when archived data is used, older data formats dominate the discourse and current practice. While newer, high-resolution, and more useful systems exist, they are not widespread and few practice-ready methodologies exist.

Agencies are further limited by monetary constraints. Agencies are dependent on public money and often serve populations least able to provide compensation. Within TriMet during 2019, passenger fares accounted for just 18% of total revenues (TriMet, 2019). Additionally, 55% of trips in 2019 were made by low-income riders (i.e. riders eligible for reduced fares) (TriMet, 2019).

Ancillary Policies

A policy does not need to come from transit agencies or be directed at transit to affect it. New state laws, such as OR House Bill 2001, have changed the zoning rules within urban areas to allow for multifamily homes in zones previously restricted to single family (80th Oregon Legislative Assembly, 2019). Portland's building codes were subsequently updated to meet the requirements of this new law and have further restricted garages and parking. For example, homes less than 22 ft. wide are no longer permitted garage entrances on their front and some units of multiplex homes must be built without garages (Bureau of Planning and Sustainability, 2020).

Transit agencies require tools to examine operations and evaluate potential changes. In the case of the new OR law, parking has been reduced and lack of parking is a known positive contributor to transit ridership (Cotugno, et al., 2008). Mode share tradeoffs have been studied in the past and models for ridership elasticity, when available, are highly useful; however, transit and travel time elasticities are not readily available for all cities.

1.2.2. Population Trends

Populations are trending towards urban areas in most parts of the world (United Nations, 2019). Over the past ten years in the United States, the population has increased by about 22.0 million people; during that same period, the urban population increased by about 24.7 million people (Worldometers.info, 2020). The United Nations predicts that 90% of the United States population will be urban by 2050, up from 83% in 2020 (United Nations, 2019). Yet, a rising urban population does not necessarily mean rising urban population densities.

In 2017, a study by Güneralp, et al. found that urban densities are likely to decline as an effect of urban sprawl (Güneralp, et al., 2017) despite population increases. Given the relationship between urban density and transit usage, a trend towards decreasing densities may potentially have a negative effect on transit usage (Shyr, et al., 2017). While the United States may experience decreasing urban densities overall, some cities or some areas within cities are likely to see increased densities. For example, the Oregon portion of the Portland Metropolitan Statistical Area (MSA) has an urban growth boundary preventing external sprawl; however, Vancouver, WA, part of the same MSA, does not have this restriction.

1.2.3. Transit Ridership

Within Portland, OR, approximately 12.0% (±0.9%) of workers over the age of 16 use public transit to commute to work. Within Portland's city limits, transit share is more than 2.4 times the national average (United States Census Bureau, 2020). However, transit systems often cross city boundaries (TriMet, 2020); by considering transit commuter trips within the urbanized parts Portland's metropolitan statistical area, transit's mode share falls by nearly half to 7.0% (±0.5%). Transit ridership is often higher within areas of a high density. For example, the urbanized parts of New York's metropolitan statistical area show about one-third (i.e. 32.5% (±0.3%)), of commuter trips using public transit (United States Census Bureau, 2020).

1.3. Motivation

Populations and population densities have a complex relationship with transit. These complexities are further complicated by internal and external policies, which must balance economic, social, and political forces. Maintaining transit coverage of the urban area will require more stations and longer transit lines as sprawl continues. Increasing the number of stops or the travel distances can potentially increase travel times, travel time uncertainties, and, subsequently, costs to both users and agencies. Combined, these factors reduce overall service attractiveness to users. Yet, the highway systems of many cities already operate at their maximum capacity during peak periods; as such, population increases, without increases in transit ridership, will result in rising congestion. The challenges outlined above, while not comprehensive, provide insight into existing system pressures and exemplify the need for cost effective and easily implemented tools for analyzing and improving transit systems.

Current methodologies typically examine performance at two main scales: a microscopic scale examining specific points, segments, and trips; and, a macroscopic scale that focuses on performance averaged over large time periods, whole routes, or entire networks. The methods to understand travel times, travel speeds, passenger service times, bus interactions, and other microscopic factors often have computational limitations when applied to networks. Conversely, macroscopic analysis of entire routes and transit systems are good at understanding high-level trends, but also ascribe an artificial uniformity to non-uniform service times and areas. Both analysis levels are useful to planning processes; yet, their limitations highlight the need for an alternative approach.

1.4. Contribution

Given the current pressures (e.g. social, political, economic, environmental, etc.) surround transit planning. It is timely to inform transit policy utilizing novel and advanced data analysis methods that can take advantage of available datasets.

Transit systems in the United States are typically a collection of fixed routes with scheduled service times for defined stop locations between an origin and a destination. Service schedules are both defined and maintained using a subset of these stops call timepoints, which are spaced out along the route. To maintain on-time performance, vehicles that arrive early at timepoints are instructed to wait until the scheduled departure time. Timepoints are a critical component of transit networks and will be used to define the unique segments used in this mesoscopic analysis. Routes are broken down into segments centered around timepoint stops, called timepoint-segments (TPS). The number of timepoint-segments per route is defined by the number of timepoints along that route and at its originating and terminal stops (typically five-eight).

For this research, data is aggregated in one-hour increments within each TPS. Four data sources, commonly available to transit agencies and researchers, are used; but, the size of the combined data set (120 GB before any analysis) necessitates methods that keep the analysis at a manageable level. As such, the data is sufficiently aggregated to reduce computational requirements; yet, preserves enough granularity to allow for a robust statistical analysis on a complete year of transit data. This timepoint aggregation represents more consistent service times and areas. Furthermore, the factors identified at the microscopic scale can be examined for trends in aggregate, which reduces variability caused by atypical and non-representative service events.

Performance measures at this mesoscopic scale may be used to understand segments individually or in context of its routes, other routes, or the entire network. This approach improves realism, which is beneficial for evaluating the key trade-offs ridership, service, accessibility, and costs. Mesoscopic performance measures may help to understand relationship between key factors influencing transit operations, evaluate uncertainty, examine variations in service, determine points sensitive to disruption, quantify congestion costs for users and agencies, and compare travel patterns between different routes, days of the week, and peak versus off-peak travel.

1.4.1. Structure of Dissertation

The body of the dissertation is divided into chapters. Figure 1-1 is a flowchart for the general structure of the dissertation that shows how each chapter relates to three key ideas: the research problem, performance metrics, and available datasets. These ideas are interrelated. Historically, the available data sets influence the types of performance measures that can be created, which then influence which data sets are prioritized by agencies and researchers. That interconnectivity is at the heart of this research.

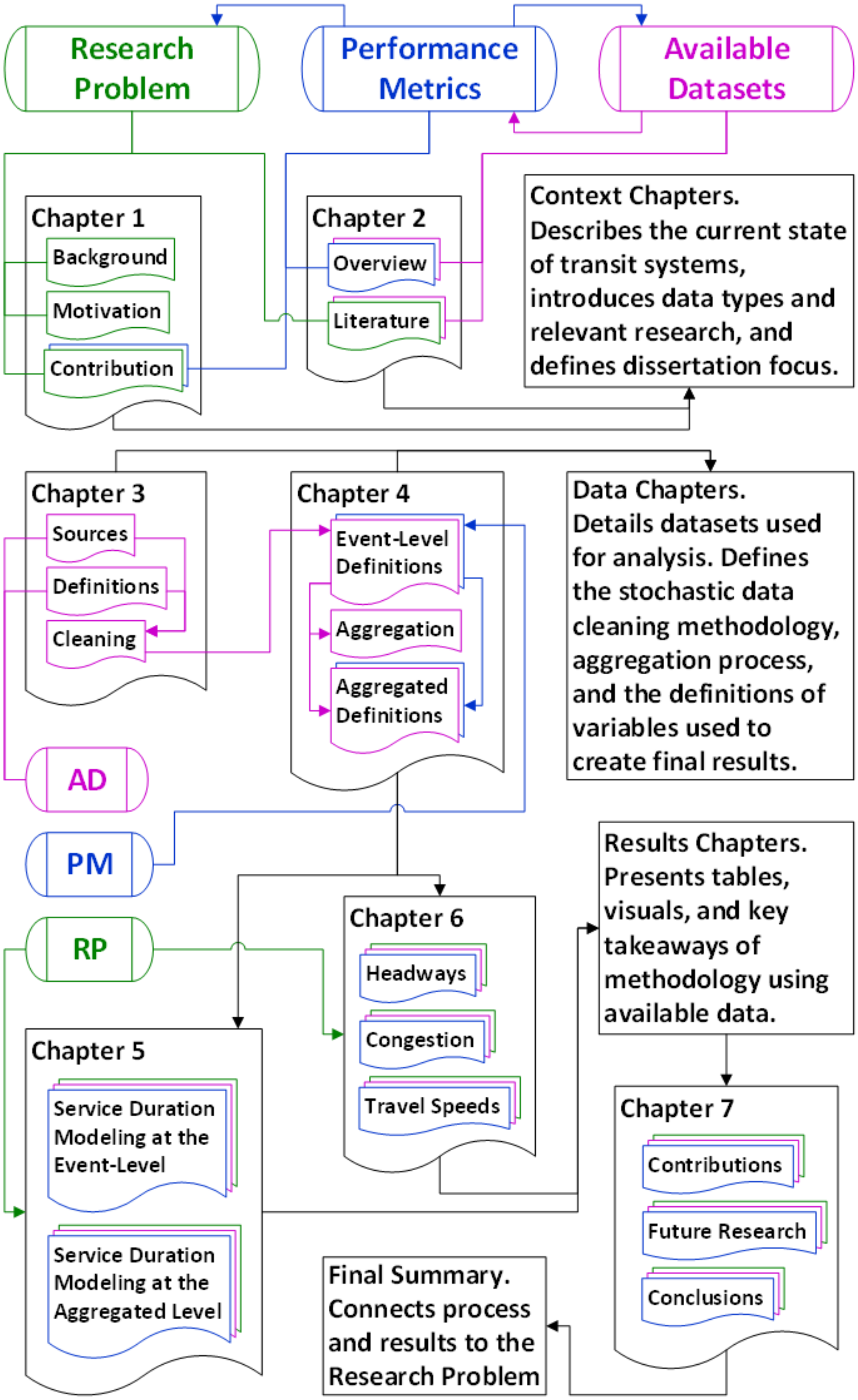
Research Problem
Performance Metrics
Available Datasets
Chapter 1
Background
Motivation
Contribution
Chapter 2
Overview
Literature
Context Chapters. Describes the current state of transit systems, introduces data types and relevant research, and defines dissertation focus.
Chapter 3
Sources
Definitions
Cleaning
Chapter 4
Event-Level Definitions
Aggregation
Aggregated Definitions
Data Chapters. Details datasets used for analysis. Defines the stochastic data cleaning methodology, aggregation process, and the definitions of variables used to create final results.
AD
PM
RP
Chapter 6
Headways
Congestion
Travel Speeds
Results Chapters. Presents tables, visuals, and key takeaways of methodology using available data.
Chapter 5
Service Duration Modeling at the Event-Level
Service Duration Modeling at the Aggregated Level
Chapter 7
Contributions
Future Research
Conclusions
Final Summary. Connects process and results to the Research Problem

Figure 1-1 — Flowchart for the general structure of dissertation. Colors indicate relationship to key topics: the research problem (green), performance measures (blue) and available data sets (purple).

While each chapter has a distinct focus, Figure 1-1 also implies three functional pairs to first six chapters. For the first pair, Chapter 1 defines the motivation and goals for this research and Chapter 2 describes the current body of literature related to those goals. Taken together, these chapters establish context for why this research is useful and how it relates to current transit systems and ongoing research. Next, Chapter 3 and Chapter 4 focus on the data, where they collectively establish the datasets, notation, and variable definitions. Separately, Chapter 3 may be more generally applicable than Chapter 4. The former introduces the datasets then outlines a stochastic cleaning methodology that may be used with other datasets. The latter is more specific to this research; it establishes the methodology for timepoint-segment aggregation and defines specific variables needed for the two results chapters. For the final pair, Chapter 5 focuses on service duration modeling and Chapter 6 examines headways, travel speeds and congestion. Together, they establish the potential usefulness of a mesoscopic as both chapters use the same aggregated data sets, applied differently, to produce distinct but complimentary results. For the body of the dissertation, Chapter 7 is not a paired chapter. It presents a final discussion of contributions and conclusions, which tie together key ideas.

Lastly, an appendix follows the main body. Appendix A defines the notation, Appendix B includes tables of variable definitions, and Appendix C contains additional tables and figures that are supplementary to those provided in the body of the dissertation.

Regarding the COVID-19 Pandemic

As indicated by publicly released data by Transit, transportation have seen substantial declines in ridership internationally (Transit, 2020). Within the United States, new regulations at the state and local levels (e.g. maximum occupancy limits, cleaning requirements every four hours, etc.) (Kate Brown, 2020) have further reduced efficiencies and increased costs. Many transit agencies have responded by reducing service to meet decreased demand (Metro, 2020) (TriMet, 2020).

This research was written during, but does not include data collected relevant to the Global Coronavirus Pandemic of 2020. The projections and estimates presented within this dissertation are based on and reflect the more typical operations pre-COVID-19. However, the methodologies may be later applied to data collected during 2020.

CHAPTER 2 — LITERATURE REVIEW

2.1. Introduction

New technologies and data availability have changed how the public, researchers, and agencies understand transit. For the agencies, the design and subsequent usability of performance measures are a central tool to evaluate their systems. This chapter will: first, introduce the primary sources of data used by agencies and researchers for transit evaluation and planning; second, introduce literature around performance measures related to this dissertation.

2.2. Transit Data

Historically, performance measures required data be collected manually (Ma & Wang, 2014). Manual data sets are high cost, difficult to collect, and limited in scope. Surveys, for example, are often biased towards literate passengers, longer trips, and seated passengers (Simon & Furth, 1985).

2.2.1. Archived AVL/APC

Modern collection systems and analysis methods have opened alternative research avenues. For example, automatic passenger counters (APC) and automatic vehicle location (AVL) data are part of a collection of technologies used for intelligent transportation systems (ITS). Such technologies and subsequent methodologies, have been shown to improve safety, operations, and planning for transit (Noch, 2019).

Stop Event Data

Stop Event Data (SED) is collected at bus stops whether or not a bus stops to service passengers. It includes operational information including, but not limited to: arrival times, departure times, scheduled stop times, door open durations, average speed between stops, and passenger movements. SED is widespread across transit agencies and often includes records from automatic counting systems for the number of passenger entering (i.e. boarding, ONS), exiting (i.e. alightings, $OFFS$), wheelchair lift usage ($LIFT$), and estimated passenger loads ($LOAD$).

SED has been a historical staple for research and analysis of route-level performance metrics; unfortunately, the use of SED is structurally limited, as the data only allows for averages between bus stops. As such, performance metrics near signalized intersections, on congested segments, or between distant bus stops lack spatial accuracy. While it may be possible to determine that a problem is occurring between two stops with a high degree of accuracy, the specific location of the problem remains uncertain without additional data sources. SED has also provided the means for research into air quality at bus stops (Moore, et al., 2012), sidewalks at intersections (Slavin & Figliozzi, 2011), and sidewalks at mid-block locations (Moore, et al., 2014).

Stop Disturbance Data

Often supplemental to SED, stop disturbance data (SDD) records information at locations where bus speeds fall to zero. Each record in the data set includes the time and duration of the stop, door activity, and stop types. Stop types are useful to understanding transit performance. For example, timepoints are used to correct for schedule discrepancies when drivers are running ahead. Other stop types include unscheduled stops at or between bus stops, or denote when buses pass a scheduled stop.

SDD provides insights into travel behaviors obscured by the structure of SED. While SDD does not include passenger movements, it helps reduce the need for estimation between stops; yet it is still limited. While SDD captions non-motion, it cannot differentiate between scenarios (to cover same distance) where an individual bus that traveled at 15 mph (24.1 kph) for two minutes then 6 mph (9.7 kph) for one minute from another bus that traveled at 12 mph (19.3 kph) for three minutes; for that, additional information is required.

High-Resolution Data

High-resolution data (HRD) collects GPS coordinates and timestamps at set intervals from onboard buses. TriMet augmented its archived data sets with HRD, at five-second resolution, in 2013. HRD helps alleviate some limitations of other AVL/APC data systems and provides a means to better examine bus behaviors between scheduled and unscheduled stop events.

2.2.2. General Transit Specification Feed

The General Transit Feed Specification (GTFS) is a standard format for the publication of transit data by transit agencies. The General Transit Feed Specification Reference (aka The Static Transit Reference), a public reference document, defines term definitions, field types, dataset files, file requirements, and field definitions that comprise a GTFS dataset (Google Developers, 2019). GTFS data has undergone many updates as new data becomes more widespread; to improve back-compatibility, organizations host current versions and archive revisions (MobilityData, 2019). In addition, some agencies have datafiles or data fields that are unique to their own operations. Such additions may be documented by agencies as unofficial or proposed GTFS data elements (TriMet Developer Resources, 2019).

Agencies produce their own GTFS datasets and often host current versions; unfortunately, these versions are usually limited to scheduled data for current operations. TriMet, for example, updates their GTFS datasets at least once per month. To promote data accessibly, other organizations archive current and past versions for most transit agencies (MobilityData, 2019).

2.2.3. Other Sources

Transit data is not always collected by transit agencies. Smartphones and other Bluetooth enabled devices allow for alternative collection methods. Often, agencies may buy data collection from private firms. Researches have also used proprietary data, such as roadside radar and Bluetooth. For example, radar data has been used to confirm that when buses are between stops, travel speeds remain close to that of general traffic (Stoll, 2016).

2.3. Performance Measures

The tools created by researchers for agencies vary in scope; some apply to single points while others apply generally to the transit system. Speeds, travel times, and congestion have all been of particular interest; additionally, these measures may be focused on transit or be used to gain understanding of general traffic conditions.

The Transit Capacity and Quality of Service Manual describes a range of potential factors that are related to service reliability. Factors from within the transit system (such as the age and quality of the vehicles, schedule, driver experience, route length, and control strategies) are influenced by and related to external factors (such as weather, signalized intersections, commuter patterns, demand, construction, and demographics) (Kittelson & Associates, et. al. , 2013). A primary goal of ongoing research has been to quantify each of these factors by itself and in relation to each other.

2.3.1. Buses as Probes

Early research efforts provided evidence that buses are subject to the same type of long-duration delays at automobiles, but the reverse is not always true. For example, buses will delay at specified timepoint bus stops when they are ahead of schedule (Hall & Vyas, 2000) (Cathey & Dailey, 2002). The data provided by TriMet has been used extensively to study the non-transit performance on major arterials in Portland (Bertini & Tantiyanugulchai, 2004) (Berkow, et al., 2008).

2.3.2. Service Times and Reliability

SED has been combined with data from loop detectors and traffic signal patterns to understand travel times and service reliability (Skabardonis & Geroliminis, 2005). Different studies have used this data to examine the point-segment level, the stop-to-stop segment level, and the route level. (Hall & Vyas, 2000) (Bertini & El-Geneidy, 2003) (Chakroborty & Kikuchi, 2004). The influence of traffic signals on bus operations has also led to research on the performance of the adaptive traffic signal system (SCATS) (Slavin, et al., 2012) and transit signal priority (TSP) (Albright & Figliozzi, 2013). The addition of detailed signal timing data allowed for Feng, *et al.* (2014) (2015) to successfully estimate the impacts of traffic volumes and intersections on transit travel times.

The addition of HRD has created opportunities to visualizes high resolution bus trajectories between stops, identify lower performance segments and signal queuing, and categorize speed breakdowns (Glick, et al., 2015). Without integrating HRD with other sources, GPS data can reduce reliance interpolation and improve methods where buses are used as proxies. Expanding on HRD applications, the same research group applied GPS data to multi-stop segments. The resulting space-time diagrams can show locations of slow speeds or high congestion (Stoll, et al., 2016). The steady rate of GPS data collection allows for heatmaps that can show clusters of GPS data points. As a first step into applying HRD, the heat maps showed locations of bus stops, intersections, and crosswalks that would have been obscured by SED. While this research provided a means to identify locations of delay, it did not provide a method for identifying the specific cause or quantifying the effect.

Improving on these results, HRD data was aggregated by location and time, which allowed for performance metrics examining percentiles and confidence intervals of travel times and travel speeds. That study also provided a methodology for removing bus stop influence when using buses as proxies (Glick & Figliozzi, 2017). The analysis more accurately represented vehicles by creating performance measures that could overcome the traditional issues of using buses to study traffic: at bus stops, buses stop to service passengers while other vehicles do not.

For each trip, quantifying transit travel-times requires breaking down trips into their service-times at bus stops and travel-times between stops. Between stops, HRD has been used effectively to create practice-ready methodologies that expand what is capable using more traditional data sets. At bus stops, time spent serving passengers, commonly known as dwell time, is a primary and known contributor to transit travel-time and travel time variability (Transit Cooperative Research Program, 2013).

2.3.3. Door Open Duration

Many studies have focused on understanding door open duration ($DWELL$), both as a stand-alone issue and in the context of travel times. Many different contributing factors have been identified by ongoing research. An obvious factor is passenger movements. Passengers entering the bus and leaving the bus have different impacts on $DWELL$ (Bertini & El-Geneidy, 2004) and their effect is also non-linear (El-Geneidy & Vijayakumar, 2011). Other research found that door choice does not have a significant effect on the magnitude of the passenger movement coefficients. (González, et al., 2012).

Other independent variables influencing $DWELL$ include, but are not limited to, payment methods (e.g. cash versus credit cards) (Milkovits, 2008) (Tirachini, 2013), bus models (e.g. low-floor versus high-floor buses or rigid versus articulated buses (Sun, et al., 2014), day of the week, time of day, passenger loads (Dueker, et al., 2004), standing passengers (Li, et al., 2012), and the location of a bus stops (Glick & Figliozzi, 2017).

For $DWELL$ prediction, SED and video (Fricker, 2011) have been the primary source. Given the limitations of both data types, previous studies are subject to the inherent limitations of the data. For example, prediction methods, based on SED, suffer from low performance at scheduled timepoints, transfer locations, and stops near intersections or traffic signals (Dueker, et al., 2004). Some of these issues have been resolved by integrating SED and HRD data sources. Some benefits of adding GPS have been discussed in previous research (Glick, et al., 2015), but question remain about different modeling approaches and the addition of new data sets.

Many models of previous studies dropped locations known to reduce model effectiveness, such as stops surrounding signalized intersections. HRD allows for the creation of new variables that may indicate if a bus stopped due to a traffic signal or congestion between service stops. For models of individual bus stops near signalized intersections, these new variables improved predictive power, the adjusted R-squared, to an average of 0.40 from an average of 0.15 and reduced the need to excluded specific stop locations in pooled models of multiple stops (Glick & Figliozzi, 2017).

2.3.4. Bus Interactions

Another research avenue of ongoing study is "bus bunching." When buses from overlapping service group together, travel time and other service instabilities occur. Bus bunching, when buses are from the same route, has been identified as a contributing factor to longer waiting times, uneven bus loading, overcrowding, and an overall reduction in service capacity (Daganzo, 2009) (Bartholdi III & Eisenstein, 2012) (Delgado, et al., 2012). Overlapping service from different routes is also an area of ongoing research. As it relates to bus bunching, overlapping service was initially shown to minimal effects on bus bunching (Diab, et al., 2016). However, additional research into overlapping service has found travel time instabilities and significant effects on service durations at bus stops.

Research into $DWELL$ and bus interactions between buses has mostly ignored the impact of bus interactions of separate transit routes on service durations. Preliminary research efforts (Glick & Figliozzi, 2019) used a limited sample to define 2-bus interaction types and quantify their effects on dwell times. Following research included all stops within the TriMet network and considered additional interaction when more than two buses have overlapping stop service. Results indicated that service times increase as the number of buses servicing the same stop increases. For overlapping routes, there is a probability distribution and time penalty associated to all buses. When multiples routes service the same stop, it is not possible to control the order of vehicle arrivals. Overlapping routes create more variability in service times at bus stops and therefore may contribute significantly to bus bunching as a result. In addition, the mean number of passenger boardings and mean $DWELL$ are substantially higher when there are bus interactions.

2.3.5. Systems Level Modeling

The trajectory and service characteristics of individual trips are difficult to predict; anomalies in expected operations of one bus can influence the operations of other buses and factors compound. The methods to understand travel times, travel speeds, service durations, bus interactions, and other microscopic factors are important, but often scale poorly when applied to larger segments or entire transit systems.

Performance measures at the system level often look at large-scale trends. This macroscopic approach can examine systemwide trends over time, but often cannot quantify how individual routes or buses contribute to these trends. TriMet, like many other transit agencies, provides tables for some of these macroscopic trends (TriMet, 2019), but also provides a breakdown by individual routes (TriMet, 2019). While useful to understanding general variability of a transit network, macroscopic route-level analyses do not provide information about overlapping service and generally obscure the high variability in demands, costs, and performance along single routes.

Network Modeling

Another area of transit research is network modeling. While outside the scope of this research, the typical computational requirements of network modeling provide context for the mesoscopic approach. Network model formulations are often limited by their computation times. Most research efforts make tradeoffs between detail and usability through assumptions that simplify their processes. For example, assumptions of constant vehicle frequencies for specific routes (Mandl, 1980), idealized passenger boarding times at bus stops (Palma & Lindsey, 2001), fixed demand along each route (Lee & Vuchic, 2005), or simplified networks without overlapping routes (Yan, et al., 2013)) somewhat reduce computational requirements but simultaneously limit model realism.

2.3.6. Cost Estimation

The types of performance measures used by transit agencies are primarily focused on aspects of the transit system within their control. These measures are important to improve service, but also to quantify how cost is distributed across transit systems. Costs may be borne by users, agencies, or both.

Transit users consider direct costs, such as fares, but also the indirect costs associated with waiting time at bus stops, transfer times, and in-vehicle travel times. Each of these factors has a theoretical cost associated with the elapsed time. For these users, the benefit of trips lies in the destinations, not the trip itself. For agencies, the trips account for a majority of costs and benefits. Agency revenues come from user trips in the form of fares and government subsidies. Other revenues include advertising, grants, and bonds. Agency costs are mostly direct expenditures that include administration and facilities but are primarily operational. Operational performance is influenced by internal policy, but also external factors, such as roadway geometry and traffic congestion.

Congestion

Traffic congestion reduces travel speeds, which increases costs associated with service times. Reducing congestion can have a positive impact on transit, traditional (i.e. not bus) drivers, and other users of the roadway. Benefits are evidenced by reductions in time costs, noise, pollutants, and the number of potential conflicts with bikes and pedestrians. The past research into performance measures, cited through Chapter 2, are directly related to congestion, but are not direct measures of the additional costs that are caused directly by congestion.

Furth and Halawani identified this gap in research and proposed a methodology to estimate costs resulting from traffic congestion at the route-level (Furth & Halawani, 2018). That research, while useful to understand the separate sources of user and agency costs, suffers from some of the same problems as other route-level analysis; specifically, routes are mostly non-homogenous and route-level analyses obscure key variations.

2.4. Conclusion

The historical data sources and previous research establish a foundation for future research; the methods used throughout this dissertation are guided by the results from those works. For example, regression modeling can proceed with foreknowledge of some expected results and without testing the full range of available independent variables because their significance or contribution has already been thoroughly tested. The datasets for this research are specific to Portland, OR. Therefore, the literature focused on TriMet data (from Portland) is also useful to focus examples or test cases. In particular, Route 9, which has been well studied, will be used as a test case for the timepoint-segment analysis to check validity of results and establish a baseline for performance.

CHAPTER 3 — DATA

3.1. Introduction

The primary objectives of Chapter 3 are to: one, introduce data sources and variables that are key to the cleaning methodology; two, detail that cleaning methodology by which broken passenger counters and outliers are identified; and three, explain how problematic data was stochastically corrected.

Please refer to Appendix A for a full explanation of the set notation (Wikipedia contributors, 2020), which is a non-typical variant of set-builder notation (Wikipedia contributors, 2020) (ProofWiki contributors, 2020), which relies heavily on indexed families (Wikipedia contributors, 2020), indexing sets (ProofWiki contributors, 2020), indexing functions (ProofWiki contributors, 2020), and predicated logic (ProofWiki contributors, 2020). Throughout Chapter 3, VAR (i.e. an example variable), will be used to introduced new ideas a notation.

Definition 3-1 — VAR [u] is the *Example Variable* with defined [u] units. VAR will be used as a placeholder to explain concepts and to introduce new notation or functions.

The notation, outlined in Appendix A, and many variables, introduced in Chapter 3, are a foundational part of Chapter 4, which uses established notation to build on ideas and define new variables. A consistent notation will be especially useful when defining aggregated variables at the timepoint-segment level. Indexes and variables are summarized in tables in Appendix B.

3.2. Sources

This study relied on two main types of data: first, archived Automatic Vehicle Location and Automatic Passenger Counter (AVL/APC) data; and second, General Transit Feed Specific (GTFS) data. The TriMet provided AVL/APC data, which was provided upon request, included Stop Event Data (SED), Stop Disturbance Data (SDD), and High-Resolution Data (HRD) sets that each had the same buses, routes, and dates and times. A new format of HRD, called breadcrumb data (BCD), was included beginning July 2018. BCD provides all values included with HRD, but adds additional identifying information. To keep consistent sources across months, BCD was not directly used.

3.2.1. Transit Maps

This data was collected for the Oregon portion of Portland Metropolitan Area. Figure 3-1 and Figure 3-2 (on the next two pages), show the extent of the transit system on the same scale: first, as an overlay on the real street map (TriMet, 2020); second, as TriMet's stylized map (TriMet, 2020). Full versions of both maps may be access online. TriMet service consists of light rail (MAX), high-frequency (Figure 3-5) and low-frequency bus lines, the WES commuter line and the Portland Streetcar. This research focuses solely on data collected for bus lines.

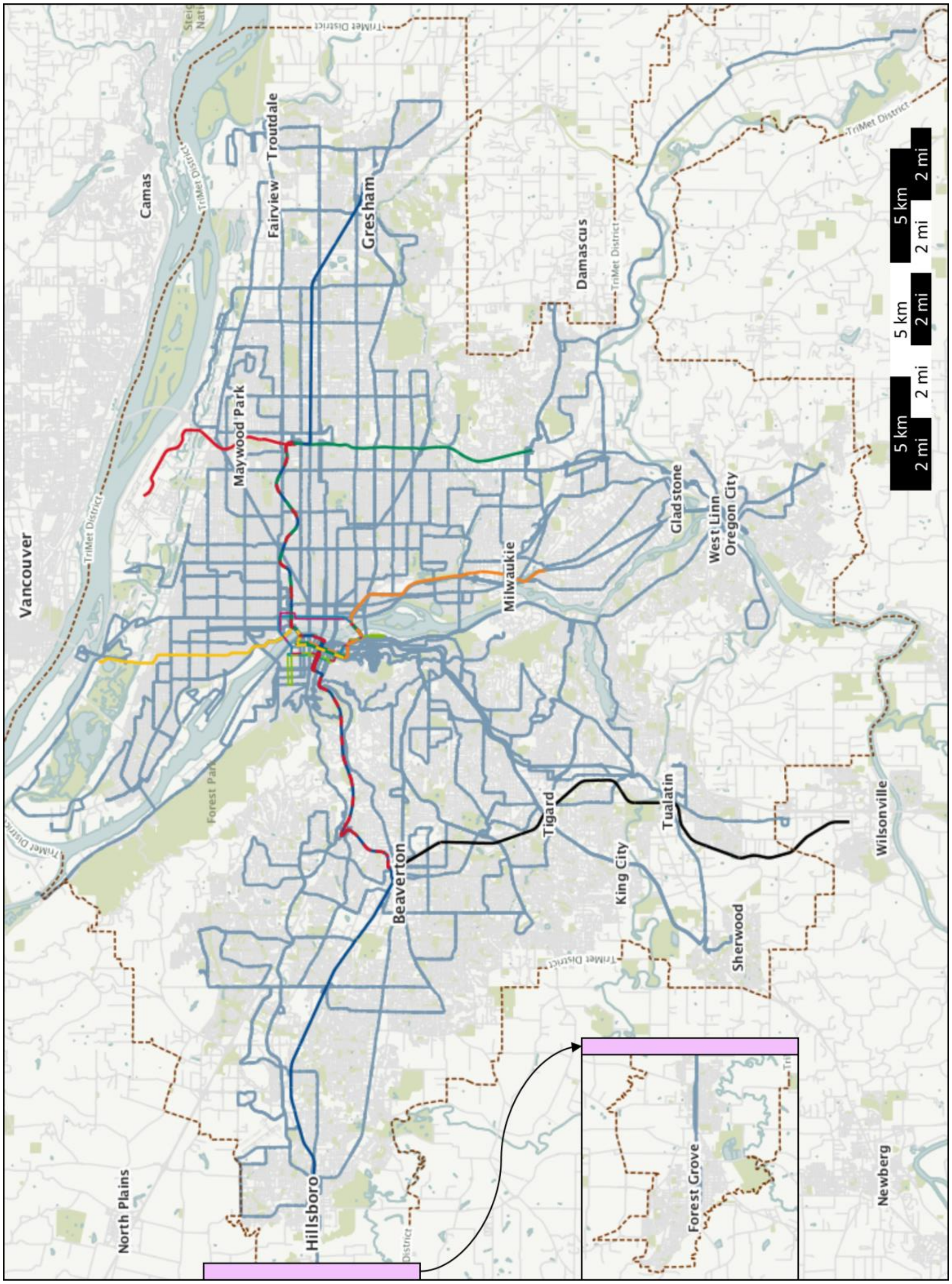

Figure 3-1 — TriMet transit ap (rotated). Visit https://ride.trimet.org/ for an interactive map of the TriMet transit system.

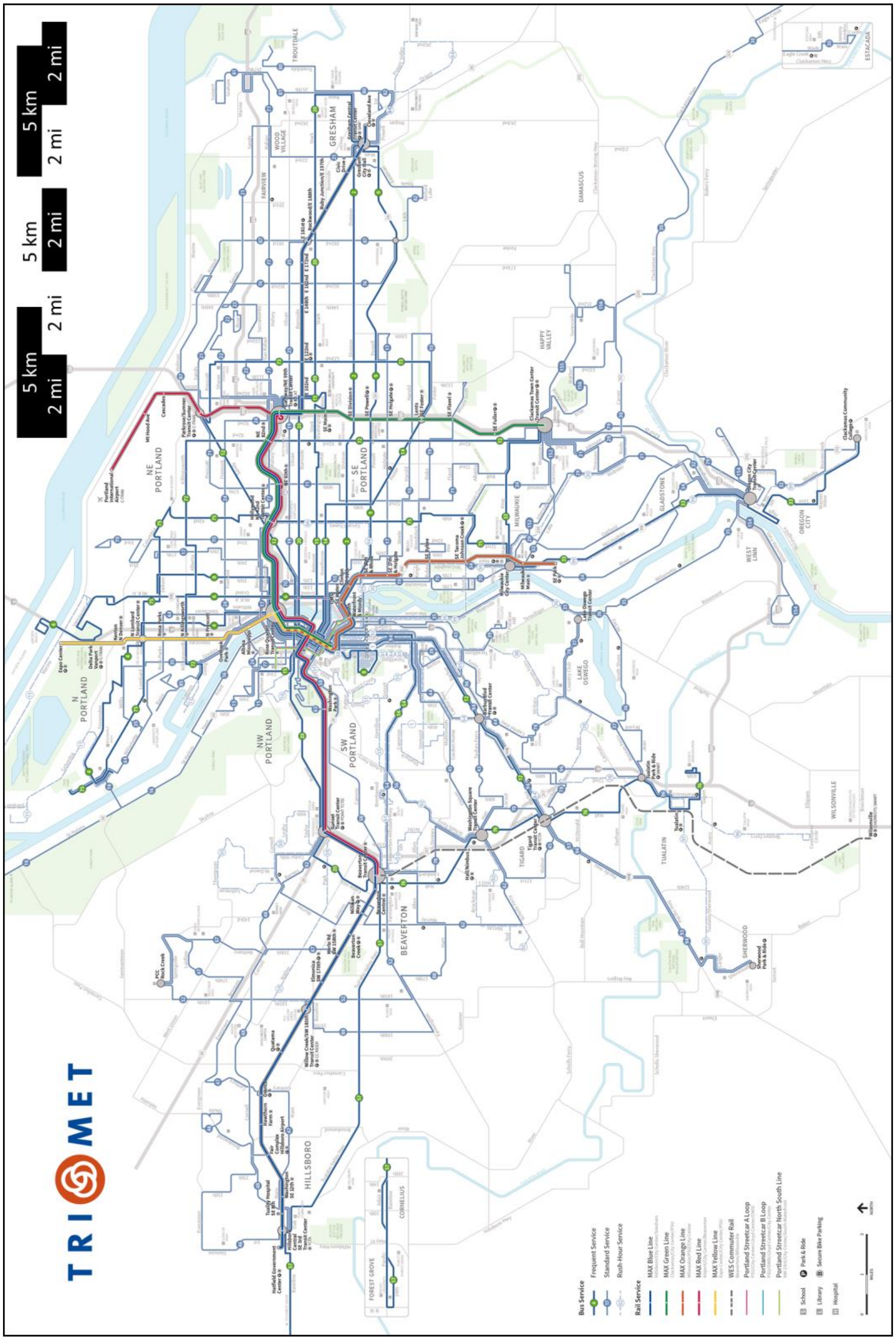

Figure 3-2 — TriMet stylized transit map (rotated) with labels for MAX and bus routes. Visit https://trimet.org/maps/img/trimetsystem.png for a full-size version.

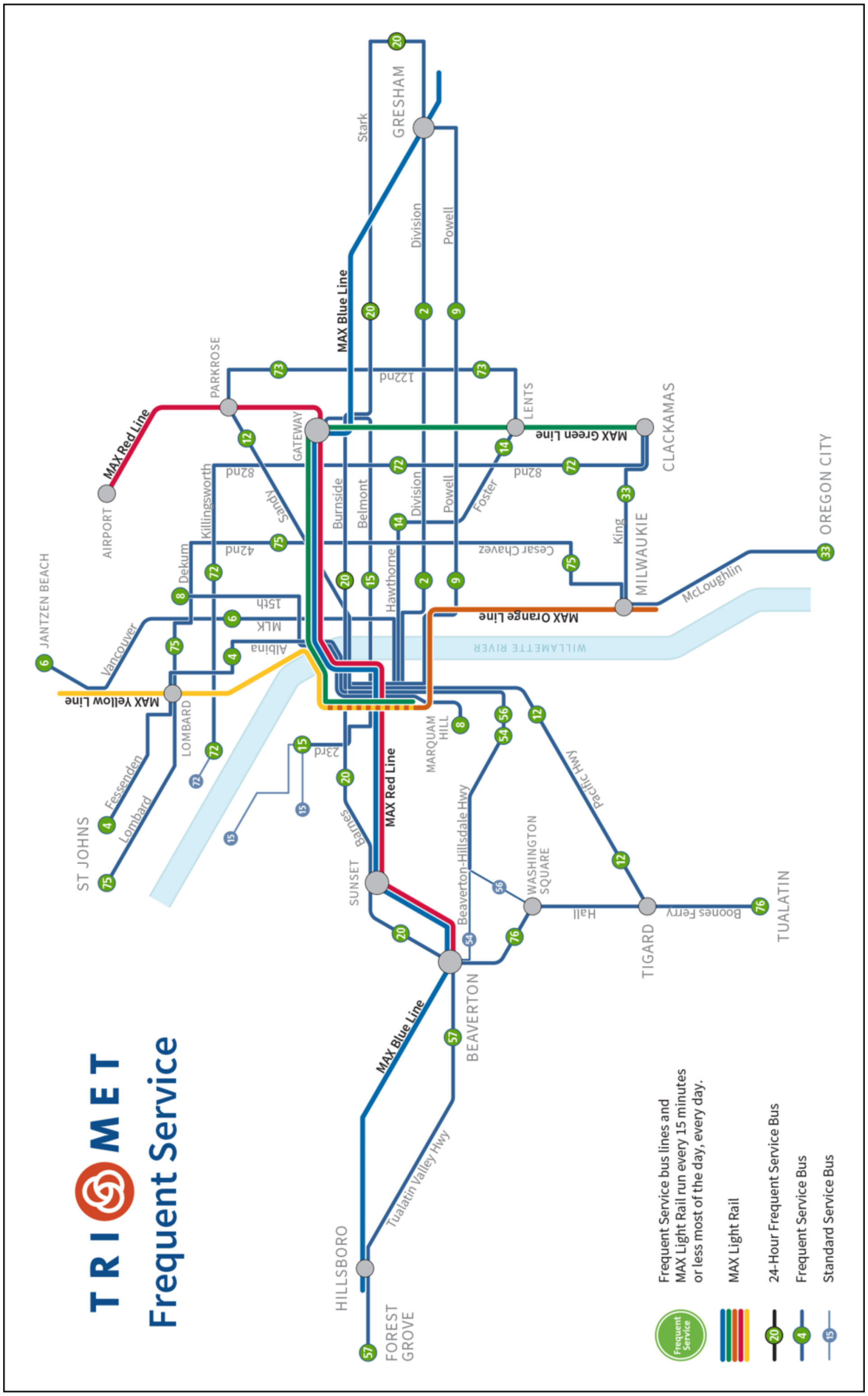

Figure 3-3 — TriMet stylized transit map of high frequency routes (rotated). Visit https://trimet.org/maps/img/frequentservice.png for a full-size version.

3.2.2. Provided Datasets

Unfortunately, data from June 2018 could not be provided and HRD data was unavailable for December 2017; thus, these months were excluded from analysis. Additionally, the first half of September 2017 was excluded due to missing data. The sizes of the raw data files (as provided) are given in Table 3-1.

Table 3-1 — File sizes, in GB, of original AVL/APC data as file.csv (comma separated values) files provided by TriMet.

		SED	**SDD**	**HRD**	**BCD**
2017	Sep	1.498	1.130	3.259	*-NA-*
	Oct	1.584	1.199	5.923	*-NA-*
	Nov	1.518	1.150	5.712	*-NA-*
2018	Jan	1.572	1.192	5.117	*-NA-*
	Feb	1.425	1.078	5.378	*-NA-*
	Mar	1.634	1.203	6.107	*-NA-*
	Apr	1.573	1.156	5.876	*-NA-*
	May	1.643	1.208	6.133	*-NA-*
	Jul	1.589	1.173	5.925	10.629
	Aug	1.650	1.210	6.231	10.990
	Sep	1.554	1.137	5.911	10.507
	Oct	1.699	1.238	6.512	11.488
	Nov	1.606	1.179	6.144	10.892
AVL/APC Total		**20.546**	**15.254**	**74.229**	**54.505**

For GTFS data, TriMet typically updates their GTFS datasets once or twice a month. To promote data accessibility, other organizations archive current and past versions for most transit agencies (OpenMobilityData, 2019), including TriMet. A total of 65 archived versions of TriMet's GTFS datasets were required to cover the time period used for this analysis. Each version does not include a fully unique data set, as many fields remain constant. Given the standardization across these datasets, all 65 were able to be merged into a single GTFS dataset with unique entries that span the full analysis period. The combined file is 9.945 GB.

For march 2018, the raw SED (Table 3-2) included 26 columns with 10.4 million rows. As provided, there are some issues with directly utilization of the entries. The first example is the SERVICE_DATE, which was not provided in a format that could be understood natively within R-studio. Another issue lies with the column headers. The first two columns of Table 3-2 and Table 3-3 for SDD are meant to give the same data; however, they do not match.

Table 3-2 — Example data table for March 2018 Stop Event Data. The total number of columns and rows in the raw data are shown.

	Columns Numbers and Names				
	1	2	⋯	25	26
Row Numbers	SERVICE_DATE	VEHICLE_ NUMBER		TRAIN_ MILAGE	PATTERN_ DISTANCE
1	02MAR2018:00:00:00	3521	⋯	34.47	0
2	02MAR2018:00:00:00	3521		34.57	0
3	02MAR2018:00:00:00	3521		34.59	535
⋮	⋮		⋱	⋮	
10,409,430	12MAR2018:00:00:00	2650	⋯	49.36	88993
10,409,431	12MAR2018:00:00:00	2650		50.82	96407
10,409,432	12MAR2018:00:00:00	2650		51.65	101047

Table 3-3 — Example data table for March 2018 Stop Disturbance Data. The total number of columns and rows in the raw data are shown.

	Columns Numbers and Names				
	1	2	⋯	25	26
Row Numbers	OPD_DATE	VEHICLE_ID		POINT_ ACTION	PLAN_ STATUS
1	10MAR2018:00:00:00	3245	⋯	D	P
2	10MAR2018:00:00:00	3245		HO	UP
3	10MAR2018:00:00:00	3245		D	P
⋮	⋮		⋱	⋮	
11,782,154	18MAR2018:00:00:00	3329	⋯	D	P
11,782,155	18MAR2018:00:00:00	3329		D	P
11,782,156	18MAR2018:00:00:00	3329		H	U

The high-resolution data (Table 3-4) similar had different column headers than the previous two data types. As such, all files required a detailed check and pre-processing to ensure that the headers matched across the archived data.

Table 3-4 — Example data table for March 2018 High Resolution Data. The total number of columns and rows in the raw data are shown.

	Columns Numbers and Names				
	1	2	⋯	9	10
Row Numbers	OPD_DATE	EVENT_ NO_TRIP		GPS_ LONGITUDE	GPS_ LATTITUDE
1	02MAR2018:00:00:00	1001217891	⋯	-122.7302	45.50902
2	02MAR2018:00:00:00	1001217891		-122.7301	45.50905
3	02MAR2018:00:00:00	1001217891		-122.7299	45.50909
⋮	⋮		⋱	⋮	
72,460,676	24MAR2018:00:00:00	1017240652	⋯	-122.6307	45.55747
72,460,677	24MAR2018:00:00:00	1017240652		-122.6307	45.55794
72,460,678	24MAR2018:00:00:00	1017240652		-122.6307	45.55891

Additionally, the identifying information was not the same across files and required an external dataset (i.e. the GTFS data) to cross reference data entries. Yet, the GTFS data sets also required some notable processing before they could be used in a meaningful way. Figure 3-4 shows an overview of the files included in the GTFS datasets archives: first, as downloaded; second, after merging. The number of rows, number of columns, and the size of the files is listed. On the following page, Figure 3-5 provides and overview of the processing steps required for the archived AVL/APC datasets.

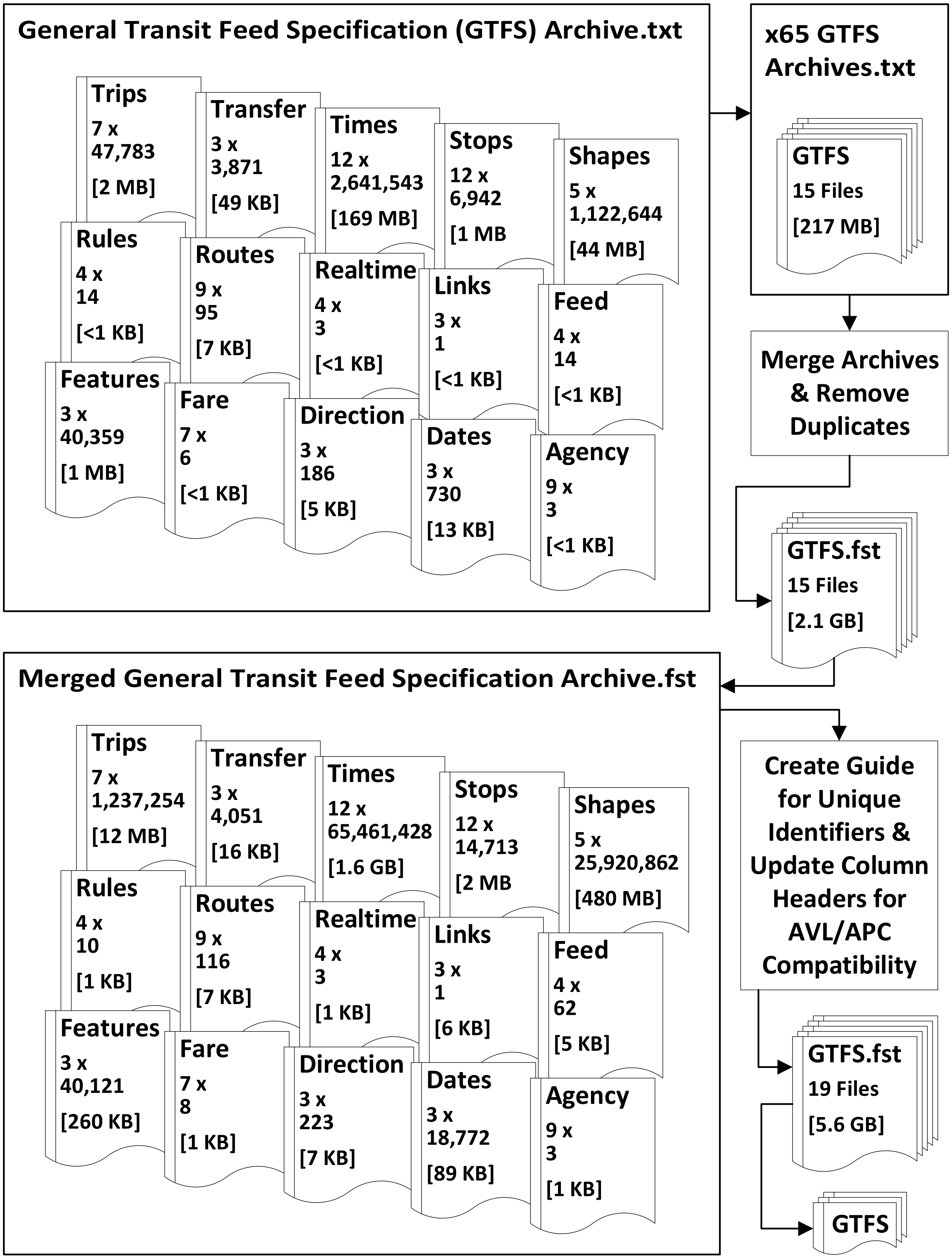
General Transit Feed Specification (GTFS) Archive.txt
Trips 7 x 47,783 [2 MB]
Transfer 3 x 3,871 [49 KB]
Times 12 x 2,641,543 [169 MB]
Stops 12 x 6,942 [1 MB
Shapes 5 x 1,122,644 [44 MB]
Rules 4 x 14 [<1 KB]
Routes 9 x 95 [7 KB]
Realtime 4 x 3 [<1 KB]
Links 3 x 1 [<1 KB]
Feed 4 x 14 [<1 KB]
Features 3 x 40,359 [1 MB]
Fare 7 x 6 [<1 KB]
Direction 3 x 186 [5 KB]
Dates 3 x 730 [13 KB]
Agency 9 x 3 [<1 KB]
x65 GTFS Archives.txt
GTFS 15 Files [217 MB]
Merge Archives & Remove Duplicates
GTFS.fst 15 Files [2.1 GB]
Merged General Transit Feed Specification Archive.fst
Trips 7 x 1,237,254 [12 MB]
Transfer 3 x 4,051 [16 KB]
Times 12 x 65,461,428 [1.6 GB]
Stops 12 x 14,713 [2 MB
Shapes 5 x 25,920,862 [480 MB]
Rules 4 x 10 [1 KB]
Routes 9 x 116 [7 KB]
Realtime 4 x 3 [1 KB]
Links 3 x 1 [6 KB]
Feed 4 x 62 [5 KB]
Features 3 x 40,121 [260 KB]
Fare 7 x 8 [1 KB]
Direction 3 x 223 [7 KB]
Dates 3 x 18,772 [89 KB]
Agency 9 x 3 [1 KB]
Create Guide for Unique Identifiers & Update Column Headers for AVL/APC Compatibility
GTFS.fst 19 Files [5.6 GB]
GTFS

Figure 3-4 — Overview of GTFS archives and processing before applying datasets to process in Figure 3-5.

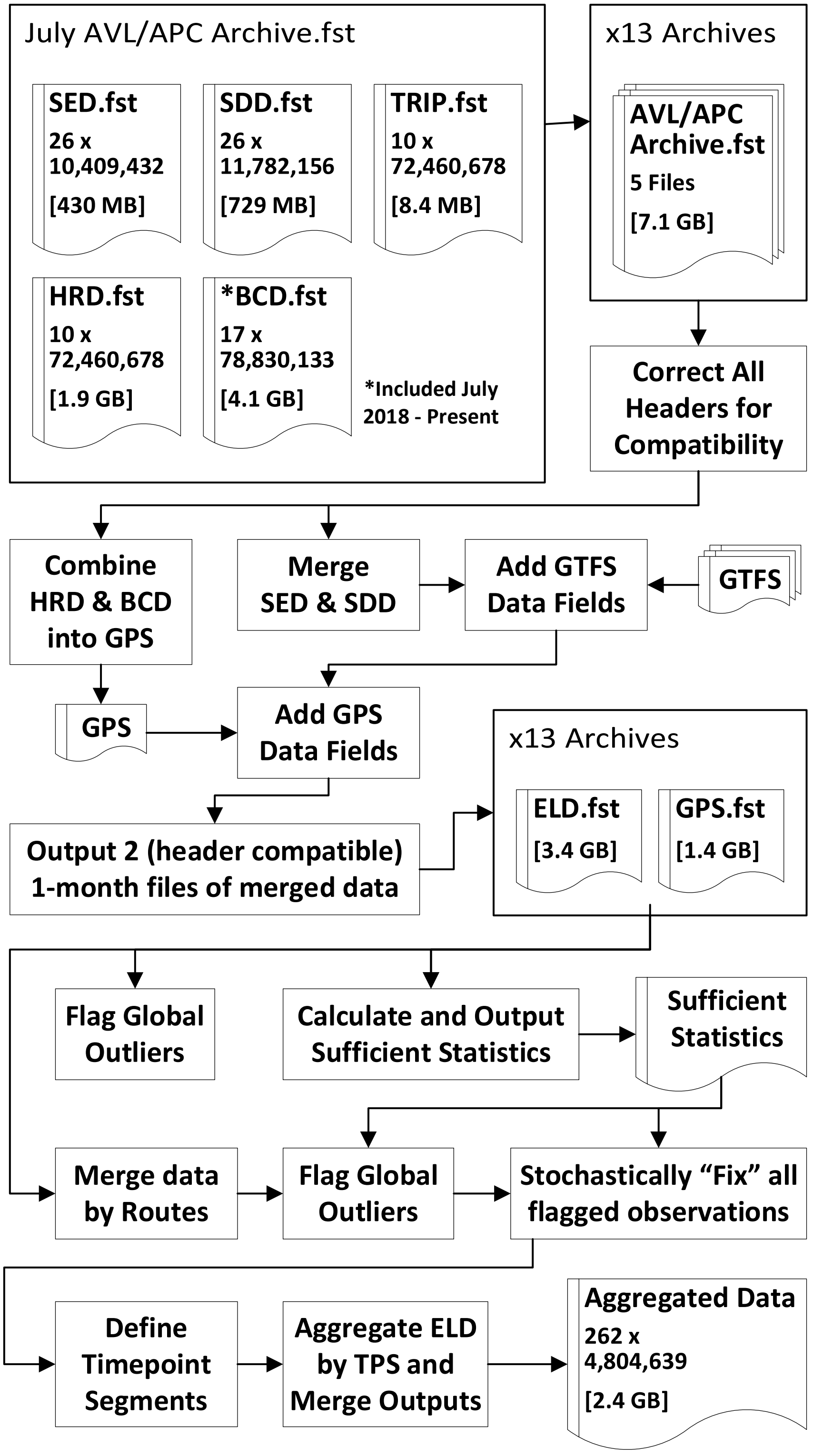

Figure 3-5 — Overview of archived AVL/APC archives and flowchart for data processing leading to aggregated datasets.

3.2.3. Dates and Clocks

All of Greater Portland falls within the Pacific Time Zone. The calendar date ($DATE$) and the Pacific Local Time (PLT) are defined by Pacific Standard Time (PST, UTC-08:00) or Pacific Daylight Time (PDT, UTC-07:00) depending on observation of standard time or daylight-saving time, respectively. Clocks are transitioned forward to 03:00 PDT on the second Sunday in March at 02:00 PST. Clocks transition back to 01:00 PST on the first Sunday in November at 02:00 PDT (Wikipedia contributors, 2020).

Definition 3-2 — $DATE$ is an actual *Calendar Date* as defined by Pacific Local Time (PLT). The index $\mathcal{d}_0$ is defined as a subset of I that includes all observations VAR_i that occurred on a unique $DATE_i$. The family of all $\mathcal{d}_0$ is contained in $\mathbb{d}_0$.

$DATE_i$ is not part of the original dataset. Instead, a service (SVC) date (${}_{SVC}DATE_i$) is recorded for every row of ELD. ${}_{SVC}DATE_i$ is not defined by $DATE_i$; instead, it is defined by TriMet's service schedule, which begins at 04:00 PLT and ends at 04:00 PLT the following morning.

Definition 3-3 — ${}_{SVC}DATE$ is a *Service Date* defined by the TriMet service schedule for specific routes and lines. ${}_{SVC}DATE$ is recorded in the dataset. The index $\mathcal{d}$ is defined as a subset of I that includes all observations VAR_i that occurred on a unique ${}_{SVC}DATE_i$. The family of all $\mathcal{d}$ is contained in $\mathbb{d}$.

(3.2.1)
$$\{VAR_i\}_{i\in \mathcal{d}} = \{VAR_i : \Phi_{\mathcal{d}}(PLT_i)\}_{i\in I} \ ,$$

Given that: $\Phi_{\mathcal{d}}(\mathrm{PLT}_i) = \begin{Bmatrix} \text{True,} & \text{if } 04{:}\,00_{\mathcal{d}_0} \le \mathrm{PLT}_i < 04{:}\,00_{\mathcal{d}_0+1} \\ \text{False,} & \text{otherwise} \end{Bmatrix}.$

The $04{:}\,00_{\mathcal{d}_0}$ to $04{:}\,00_{\mathcal{d}_0+1}$ service schedule is used for all routes, except for routes with 24-hour service. TriMet implemented its first all-day service lines (routes 20 and 57) at the start of September 2018. The service schedule for these two routes begins and ends two hours earlier at 02:00 PLT. As 24-hour service routes do not have a distinct first or last trip, service dates will be assumed to follow the schedule of index $\mathcal{d}$ for all trips.

3.2.4. Service Times

Transit agencies define and maintain service schedules (SKD) and service times at bus stop locations called timepoints. Each bus stop has a bus catchment area (Figure 3-6), also known as a bus-bay, that typically extends about 15m (50ft) before and about 10m (35ft) after the bus stop. If the number of stops or distance between timepoints is large, pseudo-timepoints are sometimes added by agencies to improve interpolation.

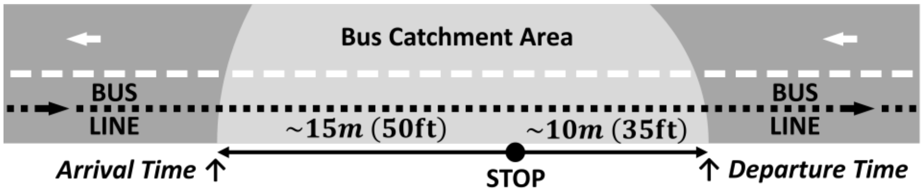

Figure 3-6 — Bus catchment area (bus-bay) for typical TriMet bus stops.

At all bus stops, scheduled time $({}^{t}SKD)$ is part of the dataset; at timepoints and pseudo-timepoints, the ${}^{t}SKD$ is the same time as is found on the schedule and or all other locations, ${}^{t}SKD$ is interpolated from upstream and downstream timepoints. ${}^{t}SKD$ is recorded as an integer number of seconds-after-midnight ($\mathcal{M}$sec). Due to the discrepancies between $DATE_i$ and ${}_{SVC}DATE_i$ at different times of day, the units, $\mathcal{M}$sec, have values larger than 86,400, the number of seconds (sec) in a typical 24-hour day. $\mathcal{M}$sec have minimum values of 14,400 and maximums of 100,799.

Definition 3-4 — ${}^{t}SKD$ [$\mathcal{M}$sec] is an officially scheduled departure time at a bus stop in TriMet's network.

Three types of event variables (i.e. ${}^{E}VAR$) are recorded within the ELD that are related to the bus-bay. Service events (${}^{E}SVC$) and disturbance events (${}^{E}DSTB$) occur when vehicles stop within or outside the bus-bay, respectively. Thru events (${}^{E}THRU$) occur when vehicles do not stop within a bus-bay that is part of their service schedule. ${}^{E}SVC$ and ${}^{E}THRU$ have associated ${}^{t}SKD,$ while ${}^{E}DSTB$ do not. Also, ${}^{E}THRU$ cannot occur outside a bus-bay or at bus stops that are not part of their regular service.

Definition 3-5 — ${}^{E}SVC$ [$\mathbb{B}$] is a binary event where a vehicle stops within a bus-bay that is part of secluded service.

Definition 3-6 — ${}^{E}DSTB$ [$\mathbb{B}$] is a binary event where a vehicle stops outside of a bus-bay that is part of secluded service.

Definition 3-7 — ${}^{E}THRU$ [$\mathbb{B}$] is a binary event where a vehicle does not stop within a bus-bay that is part of secluded service.

Arrival times (${}^{t}ARR$) and departure times (${}^{t}DEP$) are recorded within two data fields, also using units of $\mathcal{M}$sec. However, the definitions of ${}^{t}ARR$ and ${}^{t}DEP$ are dependent on the event type. For ${}^{E}THRU_i = 1$, ${}^{t}ARR_i$ is equal to ${}^{t}DEP_i$ and associated service durations and passenger movements should be zero.

Definition 3-8 — ${}^{t}ARR$ [$\mathcal{M}$sec] is a vehicle arrival time defined for:

- Service (${}^{E}SVC$) ~ Observed time that a vehicle enters a bus-bay.
- Disturbance (${}^{E}DTSB$) ~ Observed time a vehicle stops moving for more than five-seconds outside a bus-bay.
- Thru (${}^{E}THRU$) ~ Observed time that a vehicle passes a bus stop.

Definition 3-9 — ${}^{t}DEP$ [$\mathcal{M}$sec] is a vehicle departure time defined as:

- Service $({}^{E}SVC)$ ~ Observed time that a vehicle exits a bus-bay.
- Disturbance $({}^{E}DTSB)$ ~ Observed time a vehicle stops moving for more than five-seconds outside a bus-bay.
- Thru $({}^{E}THRU)$ ~ Observed time that a vehicle passes a bus stop.

Bus-Bay Service Durations

Within the bus-bay, agencies and researchers use both the bus-bay stop duration $({}^{T}BAY)$, which is difference between departure and arrival times $({}^{t}DEP - {}^{t}ARR)$ and the door open duration $({}^{T}DWL)$. Both ${}^{T}DWL$ and ${}^{T}BAY$ are recorded as an integer number of seconds. A superscript T will be used to indicate service durations.

Definition 3-10 — ${}^{T}DWL$ [Sec] is a *Door Open Duration* at bus stops and is recorded in integer seconds defined by the total time vehicle doors are open at a bus stop.

Definition 3-11 — ${}^{T}BAY$ [Sec] is a *Bus-Bay (Stop) Duration* and is recorded in integer seconds defined by the difference between arrival time and departure time.

$$\forall {}^{T}BAY_i \in \{{}^{T}BAY_i\}_{i \in J}, ({}^{T}BAY_i = {}^{t}DEP_i - {}^{t}ARR_i) \tag{3.2.2}$$

3.3. Merging and Cleaning

This research relies heavily on the *R* programming language in RStudio interface (RStudio Team, 2019) to clean and process the data. In additional native *R* functions, several library packages are used.

Reading and writing files using functions native to *R* takes a prohibitively long amount of time, but can be improved using the libraries “data.table” (Dowle, et al., 2019) and “fst” (Klik, et al., 2019). Combined, these additions to the *R* programming language allow for reductions (for this research) in required computation times, external storage, and active RAM of more than 99%, 60%, and 35%, respectively. These benefits allow for more data to be examined simultaneously while reducing downtime during active data analysis.

The library “lubridate” improves date and time functions (Grolemund & Wickham, 2011). Libraries “sp” (Pebesma & Bivand, 2005) and “rgdal” (Bivand, et al., 2019) provide the means to analyze spatial data and convert between GPS coordinates and the Oregon State-Plane North coordinate system used by TriMet. The library “zoo” provides efficient functions for data interpolation (Zeileis & Grothendieck, 2005). Finally, library “relaimpo” and library “car” are companion packages for linear regression modeling; “relaimpo” is used to calculate variable contributions (Grömping, 2006) and “car” is used to calculate variance inflation factors (Fox & Weisberg, 2018).

Combining a year of HRD, SED, SDD, and GTFS data results in about 120 GB of data before the addition of new data fields. As such, the data was primarily processed in one-month groups. If multiple months are needed simultaneously for a specific step, the read functions of the library “fst” allows for selective reading of data files, such that only the relevant rows or columns may be loaded into RAM. Data from multiple months becomes simultaneously accessible using this approach, but results must be carefully parsed between the source files. The merged data set includes two compatible files for each analysis period: the first file includes a row for each stop event; the second includes a trajectory, as GPS coordinates and timestamps, for every trip in the first data set.

3.3.1. Broken Passenger Counters

APCs were first installed on TriMet vehicles in 1981 as part of TriMet's early adoption of ITS (PB Farradyne Inc.; Battelle, 2001). All vehicles, which may be identified using their unique identification number (VEH), have APCs that record while in service.

Definition 3-12 — VEH is a *Vehicle Identification Number* that is unique to each bus or train in TriMet's network. The index $\mathscr{v}$ is defined as a subset of I that includes all observations VAR_i that were recorded on each unique VEH_i. The family of all $\mathscr{v}$ is contained in $\mathbb{v}$.

One or more APCs are located at each door of TriMet's buses and trains and use a combination of infrared and passive scanning technology to detect motion and body heat (Rose, 2009). The information collected by the APCs is: first, processed to differentiate between passengers entering (ONS) versus existing ($OFFS$) and to correct for systematic undercounting (Strathman, et al., 2005) using algorithms, which are periodically validated using surveys collected manually onboard vehicles (TriMet, 2020); and second, recorded into SED archives along with door open duration at each bus stop.

Definition 3-13 — ONS [Pax] is a number of passengers (Pax) *Boarding* (i.e. entering) a vehicle at a bus stop. Passengers enter through the front door only.

Definition 3-14 — $OFFS$ [Pax] is a number of passengers *Alighting* (i.e. exiting) a vehicle at a bus stop. Passengers exit through the front and back doors.

Definition 3-15 — $LOAD$ [Pax] is the *Estimated Passenger Load* onboard a vehicle at a given location.

Unfortunately, SED records are created even if APCs are malfunctioning. Faulty equipment is not usually identified until pre-scheduled bus maintenance and can therefore remain in service for long periods without detection. As such, the first step in data cleaning is to identify problematic passenger data caused by broken or malfunctioning APCs.

Flagging Data

When identifying and flagging problematic data, a check $(\check{\ })$ is added above the associated index or index set. VAR_i remains relevant to other sets and is therefore not relabeled as $VAR_{\check{\imath}}$, nor will the notation, $VAR_{\check{\imath}}$, be used to identify a flagged observation. Instead a single flagged observation is identified using notation from equation (3.3.1) and a set of flagged observations are identified using notation from equation (3.3.2).

Definition 3-16 — $\check{\imath}$ and $\check{\mathscr{s}}$ are a flagged index and a flagged index set. A flag for a given index set will always include the same objects as the unflagged index set.

(3.3.1) $$VAR_{i=\check{\imath}} = \{VAR_i\}_{i=\check{\imath}} = \{VAR_i : i = \check{\imath}\}$$

(3.3.2) $$\{VAR_i\}_{i \in \check{\mathscr{s}}} = \{VAR_i : i \in \check{\mathscr{s}}\}$$

Records (i.e. ONS_i, $OFFS_i$, and ${}^{T}DWL_i$) are initially flagged based on based on totals for one vehicle on one service day (i.e. $\forall i, i \in (\mathscr{v} \cap \mathscr{d})$). To simplify notation, we can use the fact that all possible pairs of $\mathscr{v} \in \mathbb{v}$ and $\mathscr{d} \in \mathbb{d}$ may be described using a cartesian product of the two families (e.g. $A \times C = \{(a, c), (a, d), (b, c), (b, d)\}$ is the cartesian product of two sets if $A = \{a, b\}$ and $C = \{c, d\}$). $\mathscr{v}_d$ is defined as a unique intersection of a vehicle index $\mathscr{v}$ and date index $\mathscr{d}$ according to equation (3.3.3).

(3.3.3) $$\forall \mathscr{v}_d \in \mathbb{v}_d, (\mathscr{v}_d = \bigcap \mathscr{v}'_d : \mathscr{v}'_d \in \mathbb{v}'_d)\ ,$$

Given that: $\mathbb{v}'_d = \mathbb{v} \times \mathbb{d} = \big\{(\mathscr{v}, \mathscr{d}) : \big((\mathscr{v} \in \mathbb{v}) \wedge (\mathscr{d} \in \mathbb{d})\big)\big\}$.

Using summation notation, ${}^{T}DWL_{\mathcal{v}_d}$, $ONS_{\mathcal{v}_d}$, and $OFFS_{\mathcal{v}_d}$ are total door open duration, total boardings, and total alightings for observations $\forall i \in \mathcal{v}_d$. If $\Phi_{\widecheck{\mathcal{v}_d}}(\cdots)$ (i.e. the predicate) from equation (3.3.4) is true, all $i \in \mathcal{v}_d$ are also part of a flagged index set $\widecheck{\mathcal{v}_d}$; if the predicate is false, then $i \notin \widecheck{\mathcal{v}_d}$ and $\widecheck{\mathcal{v}_d}$ is an empty set.

$$(3.3.4) \qquad \forall i \in \mathcal{v}_d, \left(i \in \widecheck{\mathcal{v}_d} : \Phi_{\widecheck{\mathcal{v}_d}}\left({}^{T}DWL_{\mathcal{v}_d}, ONS_{\mathcal{v}_d}, \mathrm{OFFS}_{\mathcal{v}_d}\right)\right),$$

Given that: $\Phi_{\widecheck{\mathcal{v}_d}}\left({}^{T}DWL_{\mathcal{v}_d}, ONS_{\mathcal{v}_d}, \mathrm{OFFS}_{\mathcal{v}_d}\right) = \begin{cases} \text{True}, & \text{if } \phi_A \vee (\phi_B | \neg\phi_C) \\ \text{False}, & \text{otherwise} \end{cases}$, *where* ϕ_A, ϕ_B, and ϕ_C are defined as:

$$\left\{\begin{array}{l} \phi_A \coloneqq \left({}^{T}DWL_{\mathcal{v}_d} > 0\right) \wedge \left(\min\left[ONS_{\mathcal{v}_d}, OFFS_{\mathcal{v}_d}\right] = 0\right) \\ \phi_B \coloneqq \left(2 \times \left|\mathrm{ONS}_{\mathcal{v}_d} - \mathrm{OFFS}_{\mathcal{v}_d}\right|\right) / \left(\mathrm{ONS}_{\mathcal{v}_d} + \mathrm{OFFS}_{\mathcal{v}_d}\right) > 0.15 \\ \phi_C \coloneqq ONS_{\mathcal{v}_d} + OFFS_{\mathcal{v}_d} = 0 \end{array}\right\}.$$

For ϕ_B to be true in all cases, ϕ_C must be false to prevent division by zero. ϕ_A captures cases where passenger counters do not record some type of movement, but the bus is stopping to serve passengers. ϕ_B captures large discrepancies between the number of boarding and alighting passengers. Vehicles are commonly flagged for multiple consecutive service days; an average of 11.0% and 2.6% of vehicles were flagged on any given day using ϕ_A and $\left(\phi_B \middle| \neg(\phi_C \wedge \phi_A)\right)$, respectively.

All i contained in any $\widecheck{\mathcal{v}_d}$ are also contained in $\check{\mathcal{f}}$, which is defined as: $\check{\mathcal{f}} = \bigcup \widecheck{\mathcal{v}_d}, \forall \widecheck{\mathcal{v}_d} \in \widecheck{\mathbb{v}_d}$. Additionally, $\check{\mathcal{f}}$ is defined to contain i corresponding to many first and last stops. While buses are expected to stop at each of these locations, data is commonly missing. All $VAR_i \in \left\{VAR_i : i \in \check{\mathcal{f}}\right\}$ are excluded from subsequent cleaning calculations.

$LOAD_i$ is dependent on the passenger counters and is therefore also suspect when the counters are presumed malfunctioning. ${}^{T}BAY_i$ is related to ${}^{T}DWL_i$, but was not flagged using the same methodology. $\{{}^{t}ARR_i : i \in I\}$ and $\{{}^{t}DEP_i : i \in I\}$ were mostly not missing even when passenger counters were malfunctioning. Furthermore, ${}^{T}BAY_i$ is necessary to determining which observations will need new or corrected values and which should remain as zeros.

3.3.2. Outliers

Outliers, in the context of this analysis, are data points that are non-representative of typical bus operations. Two types of outliers will be identified: first, global outliers, which are not location specific, primarily capture the most obvious atypical operations at a system level; second, local outliers, which are tailored to a specific locations, times, and routes, and capture behaviors that are atypical of a specific location. Identifying and removing global outliers is a necessary first step. If not removed from subsequent calculations, the probability distribution, calculated for local outliers, can often be non-representative of the real-world operations.

3.3.3. Global Outliers

First, global outliers were identified based on $P_p(\cdots)$, a percentile function calculated for all real, non-zero values and all real, non-zero values within a given service hour (HR).

Definition 3-17 — HR [Integer] is a *Service Hour* defined as the rounded down hour of PLT and is recorded as an integer value between 0 and 23. The index $\hbar$ is defined as a subset of I that includes all observations VAR_i that were recorded during a unique HR_i. The family of all $\hbar$ is contained in $\mathbb{h}$.

Definition 3-18 — $P_p(\cdots)$ is a function to calculate the continuous sample percentile of an input set, where p is a decimal percent between 0 and 1.

$$(3.3.5) \qquad P_p(\{VAR_i\}_{i\in\mathscr{s}}) = (1-\gamma)VAR_{(k)} + (\gamma)VAR_{(k+1)} \ ,$$

Where: $VAR_{(k)}$ and $VAR_{(k+1)}$ are the (k)th and $(k+1)$th order statistic of the ordered input set; $k = \lfloor (p)\|\mathscr{s}\| + (1-p) \rfloor$ *and* $\gamma = (p)\|\mathscr{s}\| + (1-p) - k$; *and, given that* $\lfloor x \rfloor$ is the floor funtion (i.e. rounding x down to nearest integer value), *and* $\|\mathscr{s}\|$ is defined as the number of elements in $\mathscr{s}$.

For ONS_i, $OFFS_i$, and ${}^{T}DWL_i$; first, the 99.9th percentiles were calculated for $\{VAR_i : i \in (J \backslash \breve{\mathscr{f}})\}$, which contains all real, non-zero values that are not part of a flagged set; second, the 99.99th percentiles were calculated for real, non-zero, and non-flagged values with each $\mathscr{h}$, $\{VAR_i : i \in (J \cap \mathscr{h}) \backslash \breve{\mathscr{f}}\}$. The maximum of these two percentiles (for each variable) was use as a cutoff within that service hour, $\mathscr{h}$. Cutoffs were also calculated for ${}^{T}BAY_i$ using $\{{}^{T}BAY_i : i \in J\}$ and $\{{}^{T}BAY_i : i \in (J \cap \mathscr{h})\}$.

Definition 3-19 — ${}^{MAX}VAR_{\mathscr{h}}$ is a cutoff for $VAR_i \in \{VAR_i\}_{i\in\mathscr{h}}$ and is used to identify global outliers. It has the same the units as VAR_i. The broken APC flag, $\breve{\mathscr{f}}$, is used for ONS_i, $OFFS_i$, and ${}^{T}DWL_i$, but not for ${}^{T}BAY_i$.

$$(3.3.6) \qquad {}^{MAX}VAR_{\mathscr{h}} = \max \begin{bmatrix} P_{0.999}\left(\{VAR_i\}_{i\in(J\backslash\breve{\mathscr{f}})}\right), \\ P_{0.9999}\left(\{VAR_i\}_{i\in(J\cap\mathscr{h})\backslash\breve{\mathscr{f}}}\right) \end{bmatrix}$$

Up to this point, the flagged index sets have been the same for ONS_i, $OFFS_i$, and ${}^{T}DWL_i$. For the global outliers, flagged values are specific to each VAR_i and therefore added to separate flagged index sets. The sets, $\breve{\mathscr{g}}'_1$, $\breve{\mathscr{g}}'_2$, $\breve{\mathscr{g}}'_3$, and $\breve{\mathscr{g}}'_4$ contain any i for which its corresponding conditional statement is true in equation (3.3.7). The indexes, {1,2,3,4} correspond to ONS_i, $OFFS_i$, ${}^{T}DWL_i$, and ${}^{T}BAY$, respectively.

$$\forall i \in I, \left(i \in \begin{Bmatrix} \check{\mathscr{g}}'_1 : ONS_i > {}^{MAX}ONS_{\mathscr{h} \ni i} \\ \check{\mathscr{g}}'_2 : OFFS_i > {}^{MAX}OFFS_{\mathscr{h} \ni i} \\ \check{\mathscr{g}}'_3 : {}^{T}DWL_i > \min \begin{bmatrix} {}^{T}BAY_i, \\ {}^{MAX}DWL_{\mathscr{h} \ni i} \end{bmatrix} \\ \check{\mathscr{g}}'_4 : {}^{T}BAY_i > {}^{MAX}BAY_{\mathscr{h} \ni i} \end{Bmatrix}\right) \tag{3.3.7}$$

As an example, in May 2018, 467 entries of ONS_i were flagged out of a possible 3,625,575 (i.e. 1 entry per 7,764); the 99.9th percentile for $\left\{ONS_i : i \in \left(J \backslash \check{\mathscr{b}}_1\right)\right\}$ was 19, which raised the cutoff for trips before 5:00 AM and after midnight. The maximum ${}^{MAX}ONS_{\mathscr{h}}$ was 33 between 3:00 PM and 4:00 PM.

3.3.4. Localized Outliers

The identification of broken APCs and global outliers was performed month by month, out of necessity, due to data size and computational limitations. Once identified, sufficient statistics could be calculated for the non-flagged observations and used to define parameters of probability distributions. These distribution parameters are based observations from all months and are specific to a unique combination of bus routes (RTE, $\mathscr{r}$), route directions (DIR, $\mathscr{r}_d$), bus stop locations (LOC, ℓ), service hour (HR, $\mathscr{h}$) for weekdays ($\mathscr{w} = 0$) or weekends ($\mathscr{w} = 1$).

Definition 3-20 — RTE is a *Route Identification Number* for TriMet's network. It is unique to each transit route, but not to the direction of travel. The index $\mathscr{r}$ is defined as a subset of I that includes all observations VAR_i that were recorded for a unique RTE_i. The family of all $\mathscr{r}$ is contained in $\mathbb{r}$.

Definition 3-21 — DIR is a *Direction of Travel* for TriMet routes. 1 is typically inbound to the Portland city center. The index $\mathscr{r}_d$ is defined as partitions of $\mathscr{r}$ and includes all observations VAR_i that were recorded for a unique RTE_i and DIR_i. The family of all $\mathscr{r}_d$ is contained in $\mathbb{r}_d$.

Definition 3-22 — LOC is a *Location Identification Number* for TriMet's bus stops. The index ℓ is defined as a subset of I that includes all observations VAR_i that were recorded at a unique LOC_i. The family of all ℓ is contained in $\mathbb{l}$.

Definition 3-23 — DAY is a *Day-of-the-Week* for which an observation was recorded. The index $\mathscr{w}$ is defined as a subset of I that includes all observations VAR_i that were recorded on weekdays (i.e. Monday through Fridays) or weekends (i.e. Saturday and Sunday). The family of both $\mathscr{w}$ is contained in $\mathbb{w}$.

Each unique analysis zone, indexed by $\mathscr{z}$, is defined as one unique intersection of ℓ, $\mathscr{r}_d$, $\mathscr{h}$, and $\mathscr{w}$. Each $\{VAR_i : i \in (\mathscr{z} = x)\}$ is assumed to have characteristic behaviors that may be defined independently of any $\{VAR_i : i \in (\mathscr{z} \neq x)\}$.

$$\forall \mathscr{z} \in \mathbb{z}, (\mathscr{z} = \bigcap \mathscr{z}' : \mathscr{z}' \in \mathbb{z}') \, , \tag{3.3.8}$$

Where: $\mathbb{z}' = \mathbb{l} \times \mathbb{r}_\mathrm{d} \times \mathbb{h} \times \mathbb{w}$

$$= \left\{(\ell, \mathscr{r}_\mathrm{d}, \mathscr{h}, \mathscr{w}) : \left((\ell \in \mathbb{l}) \wedge (\mathscr{r}_\mathrm{d} \in \mathbb{r}_\mathrm{d}) \wedge (\mathscr{h} \in \mathbb{h}) \wedge (\mathscr{w} \in \mathbb{w})\right)\right\}.$$

Once a parameter for a probability distribution has been defined for a given variable, distributions of a sample maximum, based on order statistics, are used to calculate cutoff values that may identify (i.e. "flag") local outliers.

Discrete Distributions

If a discrete random variable, X, has known cumulative distribution function (CDF), $F_X(x)$, then the theoretical maximum value, $X_{(n)}$, from an ordered sample, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, also has a known CDF, $F_{X_{(n)}}(x)$, and is defined in equation (3.3.9) (Casella & Berger, 2002).

$$F_{X_{(n)}}(x) = P\left(X_{(n)} \leq x\right) = \left(F_X(x)\right)^n \tag{3.3.9}$$

Using $F_{X_{(n)}=VAR_{(n)}}(x = {}^{MAX}VAR_{\mathscr{z}}) = 0.95$, the value, ${}^{MAX}VAR_{\mathscr{z}}$, is defined such that there is a 95% probability that the maximum value of a sample of size n will be smaller than ${}^{MAX}VAR_{\mathscr{z}}$. $\{ONS_i : i \in \mathscr{z}\}$ are assumed to follow a *Poisson* $({}^{ONS}\lambda_{\mathscr{z}})$ distribution, where ${}^{ONS}\lambda_{\mathscr{z}}$ is the mean number of passengers that board each stopping bus within zone $\mathscr{z}$. However, calculating ${}^{ONS}\lambda_{\mathscr{z}}$ requires excluding flagged values. As such, $\mathscr{z}'$ will be defined as the set difference between $\mathscr{z}$ and the intersection of relevant flags, $\check{\mathscr{f}}$, $\check{\mathscr{g}}'_1$, and $\check{\mathscr{g}}'_3$. For ${}^{ONS}\lambda_{\mathscr{z}}$, including $\check{\mathscr{g}}'_2$ (i.e. the flag for $OFFS_i$) would remove data unnecessarily.

(3.3.10) $$ {}^{ONS}\lambda_{\mathscr{z}} = \left(ONS_{\mathscr{z}^{\{1\}}}\right) / \left(\textstyle\sum_{i \in \mathscr{z}^{\{1\}}} \left[\mathbf{1}_{\{DWL_i > 0\}}\right]\right) ,$$

Where: $\mathscr{z}^{\{1\}} = \mathscr{z} \backslash \left(\check{\mathscr{f}} \cap \check{\mathscr{g}}'_1 \cap \check{\mathscr{g}}'_3\right)$; *and, given that:* $\mathbf{1}_{\{\dots\}}$ is an indicator function that is equal to 1 if $\{\cdots\}$ is true, or 0 if $\{\cdots\}$ is false.

${}^{ONS}\lambda_{\mathscr{z}}$ applies when buses stop to serve passengers, thus excluding zeros when buses do not stop. As an example, Figure 3-7 shows a histogram of $n = 10{,}000$ random values from a *Poisson*$(\lambda = 3)$ distribution, the PDF of the maximum, $X_{(n)}$, and the cutoff value, x, used for data cleaning.

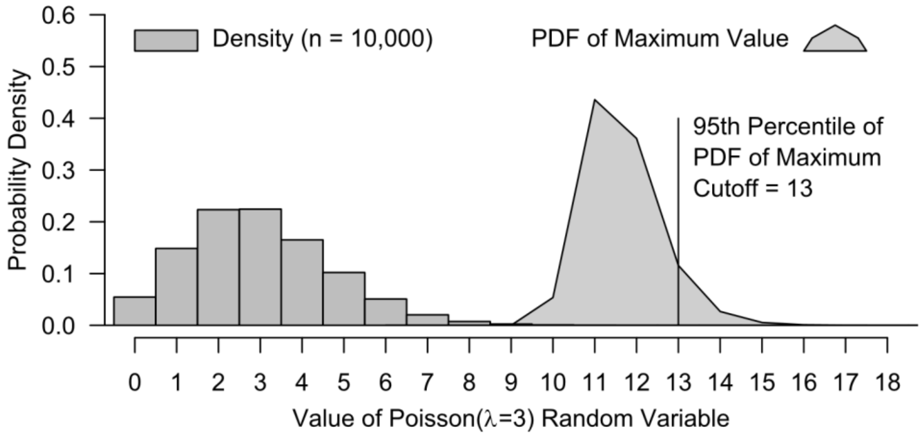

Figure 3-7 — Histogram of *Poisson*$(\lambda = 3)$ random variables and calculated probability density function (PDF) of sample maximum.

Within a given analysis zone, passenger alightings (i.e. $\{OFFS_i : i \in z\}$) may be assumed to follow a *Binomial* $(n, {}^{OFFS}p_z)$ distribution where there is a ${}^{OFFS}p_z$ probability any one passenger will exit a bus at a stop, given n passengers (i.e. the current $LOAD_i$). Equation (3.3.11) was used to define the expected probability.

(3.3.11) $${}^{OFFS}p_z = \left(OFFS_{z^{\{2\}}}\right) / \left(\sum_{i \in z^{\{2\}}} \left[(LOAD_i) \mathbf{1}_{\{DWELL_i > 0\}} \right]\right),$$

Where: $z^{\{2\}} = z \backslash \left(\check{\mathscr{f}} \cap \check{\mathscr{g}}'_2 \cap \check{\mathscr{g}}'_3\right)$.

The formulation of ${}^{OFFS}p_z$ in equation (3.3.11) is important for "fixing" flagged data, but somewhat problematic to use as a cutoff due to errors from ONS_i or $OFFS_i$ upstream of a given stop. A useful assumption is that *Poisson* $({}^{OFFS}\lambda_z)$ can be a reasonable approximation for *Binomial* $(n, {}^{OFFS}p_z)$, given a large enough n and small enough ${}^{OFFS}p_z$ (Casella & Berger, 2002). Using this approximation, a cutoff for $\{OFFS_i : i \in z\}$ was defined using the same procedure as for $\{ONS_i : i \in z\}$, such that ${}^{OFFS}\lambda_z$ is defined in equation (3.3.12).

(3.3.12) $${}^{OFFS}\lambda_z = \left(OFFS_{z^{\{2\}}}\right) / \left(\sum_{i \in z^{\{2\}}} \left[\mathbf{1}_{\{DWELL_i > 0\}} \right]\right),$$

Where: $z^{\{2\}} = z \backslash \left(\check{\mathscr{f}} \cap \check{\mathscr{g}}'_2 \cap \check{\mathscr{g}}'_3\right)$.

Continuous Distributions

For continuous distributions, if $f_Y(y)$ and $F_Y(y)$ (i.e. the PDF and CDF, respectively) are known a continuous random variable Y, then $f_{Y_{(n)}}(y)$ (i.e. the PDF of the maximum value, $Y_{(n)}$) is also known. $f_{Y_{(n)}}(y)$ may be used to define the CDF of the maximum and a cutoff value for a given variable.

(3.3.13)
$$f_{Y_{(n)}}(y) = P\big(Y_{(n)} = y\big) = nf_Y(y)\big(F_Y(y)\big)^{n-1}$$

Like with discrete distributions, $F_{Y_{(n)}=VAR_{(n)}}(y = {}^{MAX}VAR_z) = 0.95$ is defined such there is a 95% probability that the maximum value of a sample of size n will be smaller than ${}^{MAX}VAR_z$.

Both ${}^{T}DWL$ and ${}^{T}BAY$ are provided as an integer within the data, but may be reasonably assumed to follow a continuous *Lognormal* (μ_z, σ_z^2) distribution (Glick & Figliozzi, 2017). As such, a jitter, based on a continuous uniform distribution, is added to create a continuous distribution of values using equation (3.3.14) for ${}^{T}\widetilde{B}AY_i$ followed by equation (3.3.15) for ${}^{T}\widetilde{D}WL_i$. Important features of equations (3.3.14) and (3.3.15) are: non-real values are not included; zeros are unchanged; ${}^{T}\widetilde{V}AR_i$ is greater than one for all ${}^{T}\widetilde{V}AR_i \in \{{}^{T}\widetilde{V}AR_i : i \in J\}$; and ${}^{T}\widetilde{D}WL_i < {}^{T}\widetilde{B}AY_i$. This formulation allows for the natural logarithm to be taken for all non-zero values without generating zeros or negative values.

(3.3.14)
$$\forall {}^{T}\widetilde{B}AY_i \in \left\{{}^{T}\widetilde{B}AY_i\right\}_{i\in J}, \left({}^{T}\widetilde{B}AY_i = {}^{T}BAY_i + \begin{Bmatrix} U_{(0,0.5)}, & \text{if } {}^{T}BAY_i = 1 \\ U_{(-0.5,0.5)}, & \text{if } {}^{T}BAY_i > 1 \end{Bmatrix}\right),$$

Given that: $U_{(a,b)} \sim Uniform(a, b)$.

(3.3.15)
$$\forall {}^{T}\widetilde{D}WL_i \in \left\{{}^{T}\widetilde{D}WL_i\right\}_{i\in J}, \left({}^{T}\widetilde{D}WL_i = {}^{T}DWL_i + \begin{Bmatrix} U_{(0,b_1)}, & \text{if } {}^{T}BAY_i = {}^{T}DWL_i = 1 \\ U_{(0,0.5)}, & \text{if } {}^{T}BAY_i > {}^{T}DWL_i = 1 \\ U_{(-0.5,b_2)}, & \text{if } {}^{T}BAY_i = {}^{T}DWL_i > 1 \\ U_{(-0.5,0.5)}, & \text{if } {}^{T}BAY_i > {}^{T}DWL_i > 1 \end{Bmatrix}\right),$$

Given that: $U_{(a,b)} \sim Uniform(a, b)$, $b_1 = {}^{T}\widetilde{B}AY_i - 1$, *and* $b_2 = {}^{T}\widetilde{B}AY_i - {}^{T}BAY_i$.

When calculating the sufficient statistics, flagged values also needed to be excluded and these index sets slightly different between ${}^{T}\widetilde{D}WL_i$ and ${}^{T}\widetilde{B}AY_i$. All indices used for a ${}^{T}\widetilde{V}AR_i$ remain the same as indices for ${}^{T}VAR_i$; therefore, the flagged values are not changed. For a Lognormal distribution, the mean and variance of the distribution may be estimated according to equation (3.3.16).

$$\text{(3.3.16)} \qquad \begin{aligned} \text{Mean} &\coloneqq \exp[{}^{VAR}\mu_{\mathscr{z}} + (0.5)\,{}^{VAR}\sigma_{\mathscr{z}}^2] \\ \text{Variance} &\coloneqq (\exp[{}^{VAR}\sigma_{\mathscr{z}}^2] - 1)(\exp[(2)\,{}^{VAR}\mu_{\mathscr{z}} + {}^{VAR}\sigma_{\mathscr{z}}^2]) \end{aligned}\ ,$$

Where: ${}^{VAR}\mu_{\mathscr{z}}$ and ${}^{VAR}\sigma_{\mathscr{z}}^2$ are defined as:

$$\begin{Bmatrix} {}^{VAR}\mu_{\mathscr{z}} \coloneqq \left(\sum_{i\in\mathscr{z}'}\left[\ln[{}^{T}\widetilde{V}AR_i]\right]\right)/(\sum_{i\in\mathscr{z}'}[\mathbf{1}_{\{{}^{T}\widetilde{V}AR_i>0\}}]) \\ {}^{VAR}\sigma_{\mathscr{z}}^2 \coloneqq (\exp[{}^{VAR}\sigma_{\mathscr{z}}^2] - 1)(\exp[(2)\,{}^{VAR}\mu_{\mathscr{z}} + {}^{VAR}\sigma_{\mathscr{z}}^2]) \end{Bmatrix},$$

$$\textit{And where: } \mathscr{z}' = \begin{Bmatrix} (\mathscr{z} \cap J)\backslash\left(\breve{\mathscr{f}} \cap \breve{\mathscr{g}}_3'\right) & \text{for } {}^{T}\widetilde{V}AR_i \coloneqq {}^{T}\widetilde{D}WL_i \\ (\mathscr{z} \cap J)\backslash\breve{\mathscr{g}}_4' & \text{for } {}^{T}\widetilde{V}AR_i \coloneqq {}^{T}\widetilde{B}AY_i \end{Bmatrix}.$$

The formulation in of a lognormal distribution allows for the sufficient statistics of this distribution to be the number of observations (i.e. $\sum_{i\in\mathscr{z}'}[\mathbf{1}_{\{{}^{T}\widetilde{V}AR_i>0\}}]$), the sum of the natural log (i.e. $\sum_{i\in\mathscr{z}'}\left[\ln[{}^{T}\widetilde{V}AR_i]\right]$), and the sum of the natural log squared (i.e. $\sum_{i\in\mathscr{z}'}\left[\ln^2[{}^{T}\widetilde{V}AR_i]\right]$). With this three values, parameters of the lognormal distribution may be estimated for each analysis zone.

After estimating the parameters of each distribution and calculating the local cutoff value, local outliers could be identified. The sets, $\breve{\mathscr{g}}_1''$, $\breve{\mathscr{g}}_2''$, $\breve{\mathscr{g}}_3''$, and $\breve{\mathscr{g}}_4''$ contain any i for which its corresponding conditional statement is true in equation (3.3.17).

$$(3.3.17) \qquad \forall i \in J, (i \in \left\{\begin{matrix} \breve{\mathscr{g}}_1'' : ONS_i > {}^{MAX}ONS_{z\ni i} \\ \breve{\mathscr{g}}_2'' : OFFS_i > {}^{MAX}OFFS_{z\ni i} \\ \breve{\mathscr{g}}_3'' : {}^{T}\widetilde{D}WL_i > {}^{MAX}DWL_{z\ni i} \\ \breve{\mathscr{g}}_4'' : {}^{T}\widetilde{B}AY_i > {}^{MAX}BAY_{z\ni i} \end{matrix}\right\})$$

For the entire data set, excluding broken passenger counters, 0.103%, 0.075%, and 0.177% were flagged for $\{ONS_i : i \in J\}$, $\{OFFS_i : i \in J\}$, and $\{{}^{T}\widetilde{D}WL_i : i \in J\}$, respectively. There exists overlap between flagged values, such that the total is 0.287% of non-zero events were flagged at bus stops. For both discrete and continuous distribution, the procedure for flagging maximum values does not guarantee any values will be removed and allows the distributions to be customized to the demands of specific locations, routes, times, and days. Considering broken passenger counters, a total of 15.13% of bus stop service events were flagged due to either or both of ONS_i and $OFFS_i$.

Moving forward, $\breve{\mathscr{g}}_{\{1,2,3,4\}}$ are defined to contain the union of their respective $\breve{\mathscr{g}}'_{\{1,2,3,4\}}$ and $\breve{\mathscr{g}}''_{\{1,2,3,4\}}$. Thus, $\breve{\mathscr{g}}_{\{1,2,3,4\}}$ are sets of global and local outliers. Also, the complete set of flags for ONS_i, $OFFS_i$, ${}^{T}\widetilde{D}WL_i$, and ${}^{T}\widetilde{B}AY_i$ will be contained in the set $\breve{\mathscr{f}}_{\{1,2,3\}} = \breve{\mathscr{f}} \cup \breve{\mathscr{g}}_{\{1,2,3\}}$ and $\breve{\mathscr{f}}_4 = \breve{\mathscr{g}}_4$, respectively.

3.3.5. "Fixing" Flagged Data

When the set of observations contains corrected values, a hat "^" is added to the variable. $\widehat{V}AR_i$ is assumed to have the same properties as VAR_i, but $\{\widehat{V}AR_i, i \in I\} \neq \{VAR_i, i \in I\}$ because flagged values have been replaced. Replacement values are generated randomly, but subject to previously calculated global and local maximums and distribution minimums.

Bus-Bay Service Durations

Corrections begin $\forall {}^{T}\widetilde{B}AY_i \in \left\{ {}^{T}\widetilde{B}AY_i : i \in (I_0 \cup J) \right\}$ in equation (3.3.18), where the output, ${}^{T}\widehat{B}AY_i$, is used in equation (3.3.19) for ${}^{T}\widehat{D}WL_i$.

(3.3.18)
$$
{}^{T}\widehat{B}AY_i = \begin{Bmatrix} {}^{T}\widetilde{B}AY_i, & \text{if } i \notin \breve{\not{f}}_4 \\ \max\begin{bmatrix} {}^{Min}B_i, \\ \min[L_B, {}^{Max}B_i] \end{bmatrix}, & \text{if } i \in \breve{\not{f}}_4 \end{Bmatrix},
$$

Given that: $L_B \sim Lognormal({}^{BAY}\mu_z, {}^{BAY}\sigma_z^2) : z \ni i$; *and, where:* ${}^{Min}B_i$ and ${}^{Max}B_i$ are defined as:

$$
\begin{Bmatrix} {}^{Min}B_i = \begin{Bmatrix} {}^{T}\widetilde{D}WL_i + U_{({}^{T}\widetilde{D}WL_i - {}^{T}DWL_i, 0.5)}, & \text{if } \left(i \notin \breve{\not{f}}_3\right) \wedge \left({}^{T}\widetilde{D}WL_i > 1\right) \\ U_{(1,1.5)}, & \text{otherwise} \end{Bmatrix} \\ {}^{Max}B_i = \min[{}^{MAX}BAY_{\hbar \ni i}, {}^{MAX}BAY_{z \ni i}] + U_{(-0.5,0)} \end{Bmatrix}.
$$

(3.3.19)
$$
{}^{T}\widehat{D}WL_i = \begin{Bmatrix} 0 & \text{if } {}^{T}\widehat{B}AY_i = 0 \\ {}^{T}\widetilde{D}WL_i, & \text{if } \left(i \notin \breve{\not{f}}_3\right) \wedge \left({}^{T}\widehat{B}AY_i > 0\right) \\ \max\begin{bmatrix} {}^{Min}D_i, \\ \min[L_D, {}^{Max}D_i] \end{bmatrix}, & \text{if } \left(i \in \breve{\not{f}}_3\right) \wedge \left({}^{T}\widehat{B}AY_i > 0\right) \end{Bmatrix},
$$

Given that: $L_D \sim Lognormal({}^{DWL}\mu_z, {}^{DWL}\sigma_z^2) : z \ni i$; *and, where:*

$$
\begin{Bmatrix} {}^{Min}D_i = U_{(1, \min[1.5, {}^{T}\widehat{B}AY_i])} \\ {}^{Max}D_i = \min\left[{}^{MAX}DWL_{\hbar \ni i}, {}^{MAX}DWL_{z \ni i}, {}^{T}\widehat{B}AY_i\right] + U_{(-0.5,0)} \end{Bmatrix}.
$$

The limits of the uniform distributions in both equations ensure that ${}^{T}\widehat{B}AY_i$ remains greater than ${}^{T}\widehat{D}WL_i$ in cases where neither value is equal to zero.

Passenger Movements

For $ONS_i \in \left\{ONS_i : i \in \breve{\not{f}}_1\right\}$ and $OFFS_i \in \left\{OFFS_i : i \in \breve{\not{f}}_2\right\}$, new values are generated randomly assuming *Poisson*$({}^{ONS}\lambda_z)$ and *Binomial*$(n, {}^{OFFS}p_z)$ distributions, respectively. The new values need to be created for each unique in sequence for stop events in order to calculate $LOAD_i$, which is needed for n in the Binomial distribution.

Definition 3-24 — $TRIP$ is a *Trip Identification Number* that is unique to one vehicle, for one day, for one complete route and direction.

- The index $\mathscr{a}$ is defined as a subset of I that includes all observations VAR_i that were recorded on each unique $TRIP_i$. The family of all $\mathscr{a}$ is contained in $\mathbb{a}$.
- To index a single trip, the index a will be used to represent all ordered events, such that $a \in \mathscr{a} = \{a_1, a_2, \ldots, a_n\}$ and $(i \leftarrow\!\shortmid a)$ is a function mapping index i from index a.
- The index $\dot{\mathscr{a}}$ is defined as a subset of $\mathscr{a}$ that includes all observations VAR_i that were recorded at a scheduled bus-stop locations. The family of all $\dot{\mathscr{a}}$ is contained in $\dot{\mathbb{a}}$.
- To index the scheduled stops for a single trip, the index $\dot{a}$ will be used to represented all ordered event at scheduled locations, such that $\dot{a} \in \dot{\mathscr{a}} = \{\dot{a}_1, \dot{a}_2, \ldots, \dot{a}_n\} \subseteq \mathscr{a}$ and $(i \leftarrow\!\shortmid \dot{a})$ and $(a \leftarrow\!\shortmid \dot{a})$ are functions mapping indexes i and a, respectively from index $\dot{a}$.

Each unique $\dot{\mathscr{a}} \in \dot{\mathbb{a}}$ contains a complete chronological sequence of bus-stop events. To generate new "fixed" values, equations (3.3.20), (3.3.21), and (3.3.22) are run in sequence for $(i \leftarrow\!\shortmid \dot{a}) = \dot{a}_1$, followed by each equation for $\dot{a} = \dot{a}_2$, repeated through $\dot{a} = \dot{a}_n$. In equation (3.3.22), checks are needed to make sure that the number of $OFFS_{\dot{a}}$, for non-flagged data, is not greater than the estimated passenger load from the previous stop. This process is repeated for each $\dot{\mathscr{a}} \in \dot{\mathbb{a}}$.

$$\text{(3.3.20)} \qquad \hat{O}NS_{i \leftarrow\!\shortmid \dot{a}} = \left\{\begin{matrix} ONS_{\dot{a}}, & \text{if } \left(i \notin \breve{\not{f}}_1\right) \wedge \phi_1 \\ \min[P_1, {}^{MAX}ONS_{\dot{a}}], & \text{if } \left(i \in \breve{\not{f}}_1\right) \wedge \phi_1 \\ 0, & \text{if } \neg\phi_1 \end{matrix}\right\},$$

Where: $P_1 \sim$ *Poisson*$({}^{ONS}\lambda_{\mathscr{z}}) : \mathscr{z} \ni (i \leftarrow\!\shortmid \dot{a})$; $\phi_1 \coloneqq (\dot{a} \neq \dot{a}_n) \wedge \left({}^{T}\widehat{B}AY_{\dot{a}} > 0\right)$; *and,* ${}^{MAX}ONS_{\dot{a}} = \min\left[{}^{MAX}ONS_{\hbar \ni (i \leftarrow\!\shortmid \dot{a})}, {}^{MAX}ONS_{\mathscr{z} \ni (i \leftarrow\!\shortmid \dot{a})}\right]$.

$$(3.3.21)\qquad \hat{L}OAD_{i\mapsfrom\dot{a}} = \begin{Bmatrix} LOAD_0 + \hat{O}NS_{\dot{a}}\,, & \text{if } \dot{a} = \dot{a}_1 \\ \hat{L}OAD_{\dot{a}-1} + \hat{O}NS_{\dot{a}} - \hat{O}FFS_{\dot{a}}\,, & \text{if } \dot{a}_1 < \dot{a} < \dot{a}_n \\ \hat{L}OAD_{\dot{a}-1} - \hat{O}FFS_{\dot{a}}\,, & \text{if } \dot{a} = \dot{a}_n \end{Bmatrix},$$

Where: $LOAD_0$ is the reamining passenger load the previous transit trip from the same vehicle on the same day.

Not all trips can begin with passengers. As such, when $LOAD_0$ is calculated is specific to individual trip patterns, not specific routes.

(3.3.22)

$$\hat{O}FFS_{i\mapsfrom\dot{a}} = \begin{Bmatrix} 0, & \text{if } \neg\phi_2 \\ OFFS_{\dot{a}}, & \text{if } \left(i \notin \breve{\mathscr{f}}_2\right) \wedge \left(OFFS_{\dot{a}} \leq \hat{L}OAD_{\dot{a}-1}\right) \wedge \phi_2 \\ \hat{L}OAD_{\dot{a}-1}, & \text{if } \left(i \notin \breve{\mathscr{f}}_2\right) \wedge \left(OFFS_{\dot{a}} > \hat{L}OAD_{\dot{a}-1}\right) \wedge \phi_2 \\ \min\begin{bmatrix} B_1, \\ {}^{MAX}OFFS_{\dot{a}} \end{bmatrix}, & \text{if } \left(i \in \breve{\mathscr{f}}_2\right) \wedge \phi_2 \end{Bmatrix},$$

Where: $B_1 \sim$ *Binomial*$\left(\hat{L}OAD_{\dot{a}-1}, {}^{OFFS}p_{\mathscr{z}}\right) : \mathscr{z} \ni (i \mapsfrom \dot{a})$; $\phi_2 \coloneqq (\dot{a} \neq \dot{a}_1) \wedge \left({}^{T}\widehat{B}AY_{\dot{a}} > 0\right)$; *and*, ${}^{MAX}OFFS_{\dot{a}} = \min\left[{}^{MAX}OFFS_{\hslash\ni(i\mapsfrom\dot{a})}, {}^{MAX}OFFS_{\mathscr{z}\ni(i\mapsfrom\dot{a})}\right]$.

The goal of this data cleaning methodology is to use as much of the data as possible while not artificially inflating or deflating the mean or variance by using data that is non-representative of typical bus operations or data collected through faulty equipment.

Flagged Data Statistics

Table 3-5 shows the mean and variances for original and corrected values. For this data cleaning, the means of *Original* Non-Flagged data and *Corrected* All Data have a statistically significant, non-zero difference. Given the number of data points, such a result is not unexpected.

Table 3-5 — Mean and variances of original and corrected data.

		Num. Obs.	**Mean**	**Variance**
${}^{T}DWL$	$\left\{{}^{T}\widetilde{D}WL_i : i \in \breve{\mathcal{I}}_3\right\}$	82,484	427.0	194628.4
	$\left\{{}^{T}\widetilde{D}WL_i : i \notin \breve{\mathcal{I}}_3\right\}$	45,753,800	14.0	450.2
	$\left\{{}^{T}\widetilde{D}WL_i : i \in I\right\}$	45,836,290	14.7	1105.9
	$\left\{{}^{T}\widehat{D}WL_i : i \in I\right\}$	45,836,290	15.1	762.3
ONS	$\left\{ONS_i : i \in \breve{\mathcal{I}}_1\right\}$	6,010,377	1.14	29.46
	$\left\{ONS_i : i \notin \breve{\mathcal{I}}_1\right\}$	39,825,910	1.22	3.61
	$\{ONS_i : i \in I\}$	45,836,290	1.21	7.00
	$\left\{\hat{O}NS_i : i \in I\right\}$	45,836,290	1.24	3.48
$OFFS$	$\left\{OFFS_i : i \in \breve{\mathcal{I}}_2\right\}$	6,006,803	1.05	16.64
	$\left\{OFFS_i : i \notin \breve{\mathcal{I}}_2\right\}$	39,829,480	1.23	3.50
	$\{OFFS_i : i \in I\}$	45,836,290	1.21	5.23
	$\left\{\hat{O}FFS_i : i \in I\right\}$	45,836,290	1.24	3.37

As this data set is used for the aggregated analysis, each hour is aggregated separately. If a random subset of n data points are examined, such that n equals average number of data points within one hour, the null hypothesis, that the true difference in the means is zero, is failed to be rejected in 93% of trials.

3.4. Conclusion

The data sets provided and produced by TriMet are highly detailed, but cumbersome to work with. Even a single month of data requires significant time to force compatibility between the files by changing headers, and converting text-fields into usable values. A data-dictionary is required, which is not necessary part of the provided archives. For this research the dictionary was produced by a TriMet employee at request, but still required external sources, such as the GTFS datasets, to connect the files. Yet after these step, the files may be merged into a single much more comprehensive archive an any one file could provide. Unfortunately, the large number of errors in the data need to be addressed to prevent large percentages of the data from being excluded.

A primary object of the data cleaning was not to change the data unnecessarily and the methodology outlined in this chapter are intended to create a working dataset where errors are identified narrowly and corrected stochastically. By not using overly broad definitions for outliers, point-specific outliers could be captured, even if those points were not atypical for a different location. Similarly, all corrections were random and based on the data similar to the point being corrected. Stochastically corrected data will not be "real" data; but it is based on "real" data and has the key benefit of not requiring the direct exclusion of problematic data. This non-exclusion is key to the aggregation, which requires all datapoints be represented to prevent underestimates.

CHAPTER 4 — MODEL FORMULATION

4.1. Introduction

The primary objectives of Chapter 4 are to: first, establish the variables used for service duration modeling in Chapter 5 and headways, congestion, and speed analysis in Chapter 6; and second, provide context for those variables in terms of distributions and other key statistics.

4.2. Event-Level Dependent Variables

Following the merging and cleaning of SED, SDD, HRD, and GTFS data sets, the resulting data set contains details about the events at and between bus stops. This data set is will be called the Event Level Data (ELD) for this research. These variables will include modifiers as left-superscripts and left-subscripts as defined in Definition A-6.

4.2.1. Service Durations

Service duration variables (${}^{T}VAR$) are the amount of time vehicles spend at bus stops and traveling between bus stops. These durations may be divided between: stopping events within bus-bays (${}^{E}SVC$) and outside of bus-bays (${}^{E}DSTB$); and, the moving duration between stopping events. ${}^{T}\widehat{D}WL_i$ and ${}^{T}\widehat{B}AY_i$ serve as dependent variables for regression models predicting service durations within bus-bays. Table 4-1 provides details for $\left\{{}^{T}\widehat{D}WL_i : i \in J\right\}$ and $\left\{{}^{T}\widehat{B}AY_i : i \in J\right\}$ and histograms for each variable and their logarithms are shown in Figure 4-1 for all non-zero stopping events. As was introduced by equation (A.3), the index set J is a partition of I, is dependent on the variable in the brackets, and captures all real, non-zero values for that variable only.

Table 4-1 — Mean, variance and percentiles for non-zero door open durations, $\{{}^{T}\hat{D}WL_i : i \in J\}$, and non-zero bus-bay stop durations, $\{{}^{T}\hat{B}AY_i : i \in J\}$.

			Percentiles								
Variable	**Mean**	**Var**	*1st*	*5th*	*15th*	*25th*	*50th*	*75th*	*85th*	*95th*	*99th*
$\{{}^{T}\hat{D}WL_i\}_{i\in J}$	15.1	762.3	2.6	3.8	5.0	6.0	8.8	15.0	21.2	43.8	109.1
$\{{}^{T}\hat{B}AY_i\}_{i\in J}$	32.4	2139.4	7.9	10.8	13.4	15.1	20.5	35.8	50.6	81.2	171.3

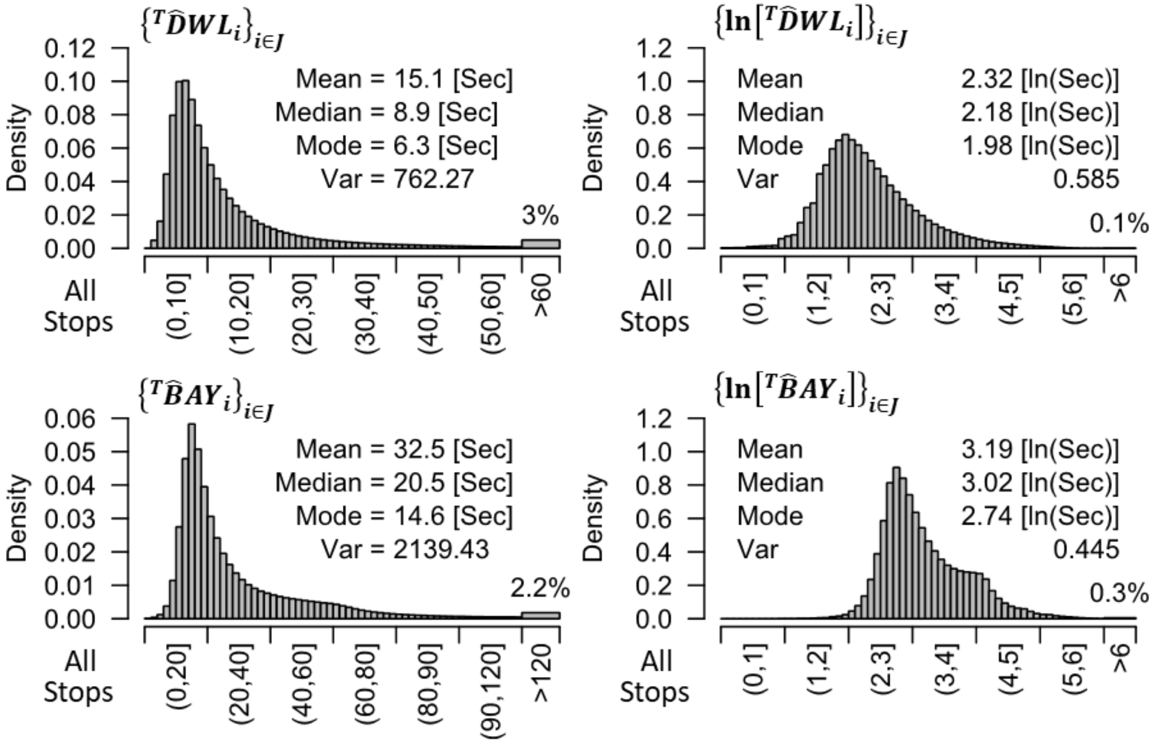

Figure 4-1 — Histograms of non-zero door-open durations, $\{{}^{T}\hat{D}WL_i : i \in J\}$ (Top-Left) and $\{\ln[{}^{T}\hat{D}WL_i] : i \in J\}$ (Top-Right), and non-zero bus-bay durations $\{{}^{T}\hat{B}AY_i : i \in J\}$ (Bottom-Left) and $\{\ln[{}^{T}\hat{B}AY_i] : i \in J\}$ (Bottom-Right).

Within the cleaned data, ${}^{T}\hat{D}WL_i$ is strictly less than ${}^{T}\hat{B}AY_i$ for non-zero values and observations are heavily skewed toward shorter service times. After taking the natural logarithm, the skew is less extreme, but still a notable part of the distributions. While the distribution for the entire network is useful, the distributions are location and time dependent. Less than 0.74%, 0.254%, 0.098%, and 0.044% of ${}^{T}\widehat{D}WL_i$ are greater than two three, four, and five minutes, respectively. Previous models for predicting ${}^{T}\widehat{D}WL$ have commonly excluded longer durations, as they do not represent typical operations and reduce model effectiveness for the vast majority of observations. This research will include ${}^{T}\widehat{D}WL_i$ and ${}^{T}\widehat{B}AY_i$ that are than five-minutes (4.2.1) in event-level regression models, which accounts for more than 99.9% of bus-bay events.

(4.2.1) $$\left\{{}^{T}\hat{V}AR_i : \left({}^{T}\hat{V}AR_i < 300\right) \wedge (i \in J)\right\}$$

Inter-stop Service

For a single trip with stops $\dot{a} = \{\dot{a}_1, \dot{a}_2, \dots, \dot{a}_n\}$, there are two important measures between a given bus-bay location, $\dot{a}$, and the previous bus-bay location, $\dot{a} - 1$: first, the amount of time a vehicle is moving; and second, the amount of time a vehicle is stopped. At bus stop $\{\dot{a} : \dot{a} \neq \dot{a}_1\}$, the number of ${}^{E}DSTB_i$ that occurred since leaving stop $\dot{a} - 1$ is denoted ${}^{E}DSTB_{\dot{a}}$, which is an integer value contained in $\mathbb{N}_0$. The total time of those disturbance stops is denoted ${}^{T}DSTB_{\dot{a}}$ and is defined as the sum of the differences between departure times and arrival times at those unplanned stopped. If no disturbances occurred, then ${}^{E}DSTB_{\dot{a}} = 0$ and ${}^{T}DSTB_{\dot{a}} = 0$.

Definition 4-1 — ${}^{T}DSTB_{\dot{a}}$ [Sec] is the *Disturbance Duration* of unscheduled stops between bus stop locations $(\dot{a} - 1)$ and $\dot{a}$.

(4.2.2) $${}^{T}DSTB_{\dot{a}} = \sum_{a=\left(a \hookleftarrow (\dot{a}-1)\right)+1}^{(a \hookleftarrow \dot{a})-1} \left[{}^{t}DEP_{\dot{a}} - {}^{t}ARR_{\dot{a}}\right]$$

${}^{T}DSTB_{\dot{a}}$ is needed to calculate the moving duration (i.e. ${}^{T}MOVE_{\dot{a}}$) between locations $(\dot{a}-1)$ and $\dot{a}$. ${}^{T}MOVE_{\dot{a}}$ is difference between the arrival time at stop $\dot{a}$ (i.e. ${}^{t}ARR_{\dot{a}}$) and the departure time at stop $(\dot{a}-1)$ (i.e. ${}^{t}DEP_{\dot{a}-1}$) minus the disturbance duration between the stop (i.e. ${}^{T}DSTB_{\dot{a}}$). As with ${}^{T}BAY_{i}$ and ${}^{T}DWL_{i}$ in the previous chapter, ${}^{T}DSTB_{i\hookleftarrow\dot{a}}$ and ${}^{T}MOVE_{i\hookleftarrow\dot{a}}$ are given a random jitter to make the variables continuous resulting in ${}^{T}\widetilde{D}STB_{\dot{a}}$ and ${}^{T}\widetilde{M}OVE_{\dot{a}}$.

Definition 4-2 — ${}^{T}MOVE_{\dot{a}}$ [Sec] is the *Moving Duration* between bus stop locations $(\dot{a}-1)$ and $\dot{a}$ and excludes the disturbance duration.

(4.2.3)

$$\forall {}^{T}MOVE_{i\hookleftarrow\dot{a}} \in \{{}^{T}MOVE_{\dot{a}}\}_{\dot{a}\in\dot{\mathbb{a}}}, ({}^{T}MOVE_{\dot{a}} = ({}^{t}ARR_{\dot{a}} - {}^{t}DEP_{\dot{a}-1}) - {}^{T}DSTB_{\dot{a}})$$

Using the bus-bay stop duration, moving duration, and the disturbance duration, the total travel duration for an individual trip, a, can be calculated. $\left\{{}^{T}\widehat{D}WL_{\dot{a}} : \dot{a} \in \dot{a}\right\}$ is not needed for total travel time because ${}^{T}\widehat{D}WL_{\dot{a}} < {}^{T}\widehat{B}AY_{\dot{a}}$, by definition. ${}^{T}\widetilde{T}RVL_{a}$ is not used directly for most calculations, but provides a means to check data cleaning and estimates.

Definition 4-3 — ${}^{T}\widetilde{T}RVL_{a}$ [Sec] is the *Total Travel Duration* from when a vehicle begins servicing passengers as its first stop and stops serving passengers at its last stop.

(4.2.4)

$$\forall {}^{T}\widetilde{T}RVL_{a} \in \left\{{}^{T}\widetilde{T}RVL_{a}\right\}_{a\in\mathbb{a}}, \left({}^{T}\widetilde{T}RVL_{a} = \sum_{\dot{a}=\dot{a}_1}^{\dot{a}_n}\left[{}^{T}\widehat{B}AY_{\dot{a}} + {}^{T}\widetilde{D}STB_{\dot{a}} + {}^{T}\widetilde{M}OVE_{\dot{a}}\right]\right)$$

Lastly, two additional binary variables are used in the regression models: $FREQ$ is defined to be 1 for high-frequency routes and 0 for low-frequency routes; and $WDAY$ is defined to be 1 for weekdays and 0 for weekends.

4.3. Event-Level Independent Variables

4.3.1. Location Variables

As mentioned above, a factor influencing service times is the location of the scheduled bus stop. Location variables (i.e. ${}^{L}VAR$) may refer generally to an area of a city (e.g. urban vs rural), describe features that are specific to a stop (e.g. shelters), or indicate proximity to features that are not part of the stop (e.g. intersections).

Definition 4-4 — ${}^{L}VAR$ [$\mathbb{B}$] is a *location-type* variable with binary units. ${}^{L}VAR_i$ is equal to 1 if true, 0 otherwise.

Timepoints, which were previously discussed, are a key location variable denoted ${}^{L}TP$. Times have unique behaviors, compared to other locations, and are a definitional part of the timepoint-segments (TPS) used for aggregation.

Definition 4-5 — ${}^{L}TP$ [$\mathbb{B}$] is a binary *Timepoint* variable.

Intersections

TriMet categories stops into four location types (i.e. Nearside (${}^{L}NEAR$), Farside (${}^{L}FAR$), Opposite (${}^{L}OPP$), and ${}^{L}AT$) based on stop placement relative to intersections. Figure 4-2 through Figure 4-5 show stop simplified versions of intersection and stop placement configurations.

Definition 4-6 — ${}^{L}NEAR$, ${}^{L}FAR$, ${}^{L}OPP$, and ${}^{L}AT$ [$\mathbb{B}$] are binary *stop-placements* location variables that indicate proximity to *nearside* (Figure 4-2), *farside* (Figure 4-3), *opposite* (Figure 4-4), and *at* (Figure 4-5) intersections, respectively.

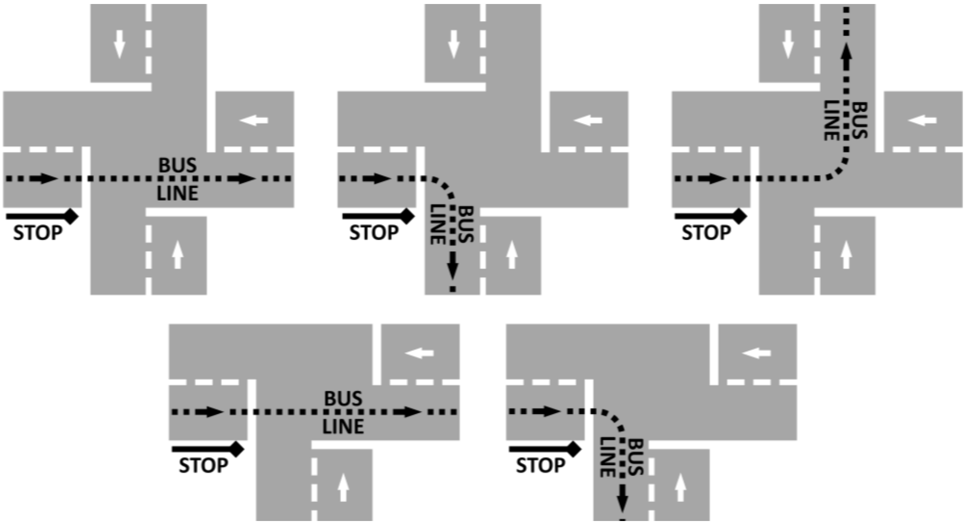

Figure 4-2 — Nearside $(^{L}NEAR)$ bus-stop placements relative to intersections.

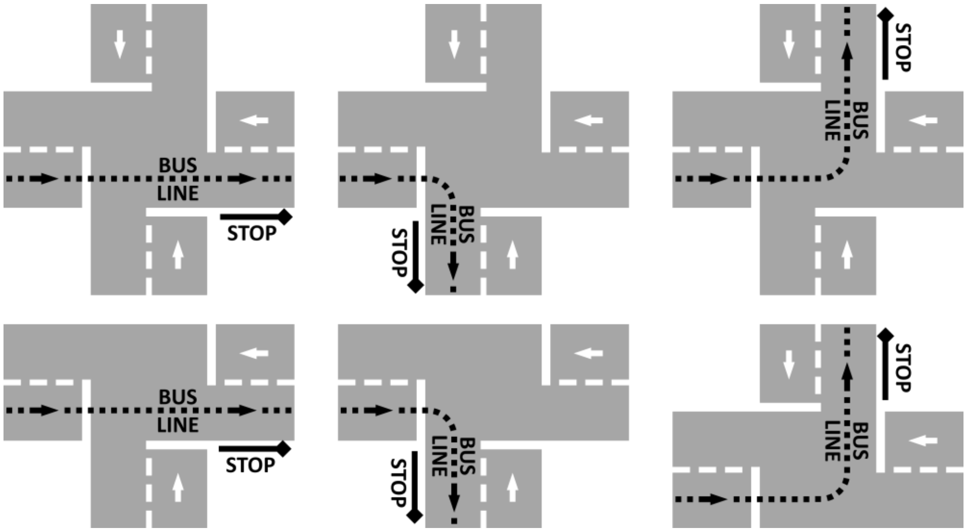

Figure 4-3 — Farside (^{L}FAR) bus-stop placements relative intersections.

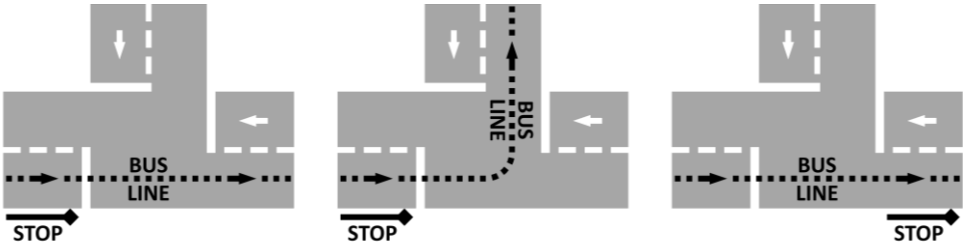

Figure 4-4 — Opposite (^{L}OPP) bus-stop placements relative to intersections.

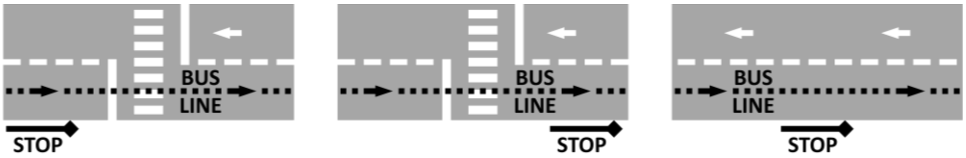

Figure 4-5 — At (${}^{L}AT$) bus-stop placements relative to intersections. ${}^{L}AT$ also includes other configurations that are not nearside (${}^{L}NEAR$), farside (${}^{L}FAR$), or opposite (${}^{L}OPP$) as shown in Figure 4-2 – Figure 4-4.

In addition, some stops may correspond to multiple values as they are found between two intersections. As an example, Figure 4-6 shows two stops that would be assigned two values: Stop (A) be both an opposite (${}^{L}OPP$) and nearside (${}^{L}NEAR$) stop and Stop (B) would be both a farside (${}^{L}FAR$) and ${}^{L}NEAR$ stop. Stops with multiple location types are one reason that all stop types may be simultaneously included in regression models. ${}^{L}NEAR_i + {}^{L}FAR_i + {}^{L}OPP_i + {}^{L}AT_i \geq 1$ for all stops $i \hookleftarrow \dot{\alpha}$.

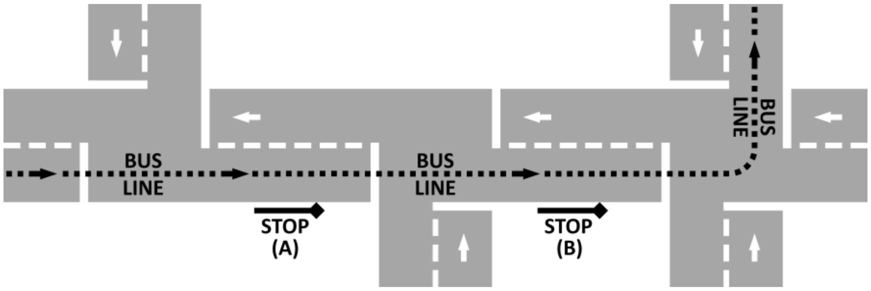

Figure 4-6 — Bus stop placements that correspond to multiple location variables.

Intersections are further complicated by traffic control features, such as signals (${}^{L}SIG$). Intersections that are signalized have a different impact on transit operations than those that are unsignalized.

Definition 4-7 — ${}^{L}SIG$ [$\mathbb{B}$] is a binary indicator for signalized intersections. ${}^{L}SIG$ is equal to 1 if signalized, 0 otherwise.

The combination of a stop location variables (${}^{L}VAR$) variables with ${}^{L}SIG$ is used to differentiate signalized locations (${}^{Ls}VAR$) and unsignalized locations (${}^{Lu}VAR$). Eight new binary variables (i.e. ${}^{Ls}NEAR$, ${}^{Ls}FAR$, ${}^{Ls}OPP$, ${}^{Ls}AT$, ${}^{Lu}NEAR$, ${}^{Lu}FAR$, ${}^{Lu}OPP$, and ${}^{Lu}AT$) allow for the effects of traffic signals to be quantified for each location type. Additionally, effects are more significant after data aggregation.

Definition 4-8 — ${}^{Ls}VAR$ [$\mathbb{B}$] is a *location-type* variable with binary units used for signalized intersections.

(4.3.1) $$\forall {}^{Ls}VAR_i \in \{{}^{Ls}VAR_i\}_{i \in J}, ({}^{Ls}VAR_i = {}^{L}VAR_i \times {}^{L}SIG_i)$$

Definition 4-9 — ${}^{Lu}VAR$ [$\mathbb{B}$] is a *location-type* variable with binary units used for unsignalized intersections.

(4.3.2) $$\forall {}^{Lu}VAR_i \in \{{}^{Lu}VAR_i\}_{i \in J}, \left({}^{Lu}VAR_i = {}^{L}VAR_i \times (1 - {}^{L}SIG_i)\right)$$

Table 4-2 gives the mean and variance for ${}^{T}\widehat{D}WL_i$ dependent on signalized and unsignalized locations, as defined by TriMet. From this table, Signalized locations (Ls) have longer ${}^{T}\widehat{D}WL_i$ than unsignalized (Lu) and ${}^{L}AT$ locations have much longer door open durations than the other stop placements. This is likely due to the fact that transit centers are generally classified as ${}^{L}AT$; however, most ${}^{L}AT$ are not transit centers. Therefore, additional variables will be useful to separate out stop locations that have distinct behaviors not defined by their relationship to intersections.

Table 4-2 — Statistics for $\left\{{}^{T}\widehat{D}WL_i : \Phi_1(\cdots) \wedge \Phi_2(\cdots) \wedge (i \in J)\right\}$, where $\Phi_1(\cdots)$ indicates intersection type and $\Phi_2(\cdots)$ indicates traffic signals.

	$\Phi_2(\cdots) \coloneqq$					
	True (Any ${}^{L}SIG_i$)		${}^{L}SIG_i = 1$		${}^{L}SIG_i = 0$	
$\Phi_1(\cdots) \coloneqq$	**Mean**	**Var**	**Mean**	**Var**	**Mean**	**Var**
True (Any Type)	14.6	455.6	16.0	450.0	13.0	460.0
${}^{L}NEAR_i = 1$	13.8	302.1	15.7	385.9	11.1	175.0
${}^{L}FAR_i = 1$	13.9	338.4	15.3	421.4	11.8	205.8
${}^{L}OPP_i = 1$	11.9	193.9	14.1	209.7	11.0	184.8
${}^{L}AT_i = 1$	25.8	2093.9	30.1	1940.8	24.5	2135

Transit Centers and Park-and-Rides

Portland has 16 transit centers $({}^{L}TC)$, that serve as hubs between multiple transit routes (Figure 4-7) (TriMet, 2020). Portland also has 61 park-and-ride $({}^{L}P\&R)$ locations where passengers may park personal vehicles and walk to nearby bus stops.

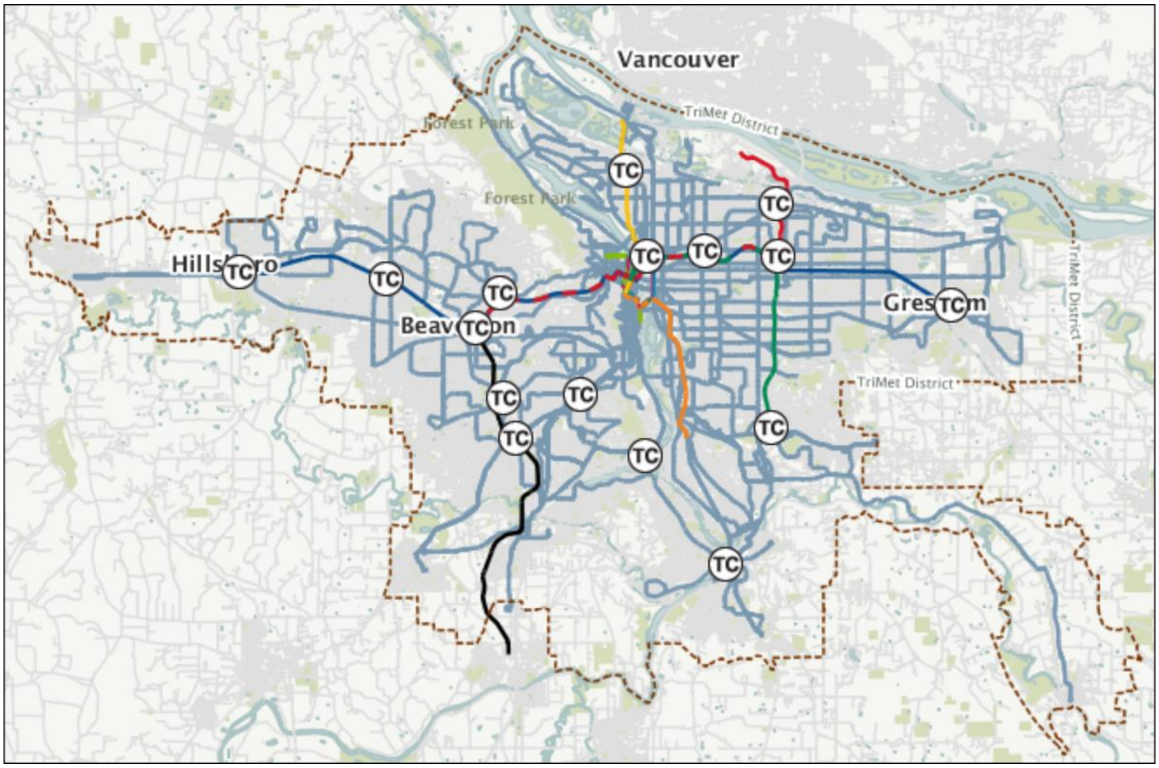

Figure 4-7 — Map of transit center located on the TriMet transit system.

Definition 4-10 — ${}^{L}TC$ [$\mathbb{B}$] is a *Transit Center* variable with binary units used to indicate if a stop is part of a transit center.

Definition 4-11 — ${}^{L}P\&R$ [$\mathbb{B}$] is a *Park & Ride* variable with binary units to indicate bus stops located within a quarter mile of a designated park-and-ride facility.

Figure 4-8 shows the histograms for service times for stops located at transit centers and stops located within a quarter mile of a park-and-ride. Both door open duration and bus-bay duration types have higher means and variances than the network as a whole (Figure 4-1). For transit centers, the tails of the histograms are longer, which indicates that that longer durations are more common.

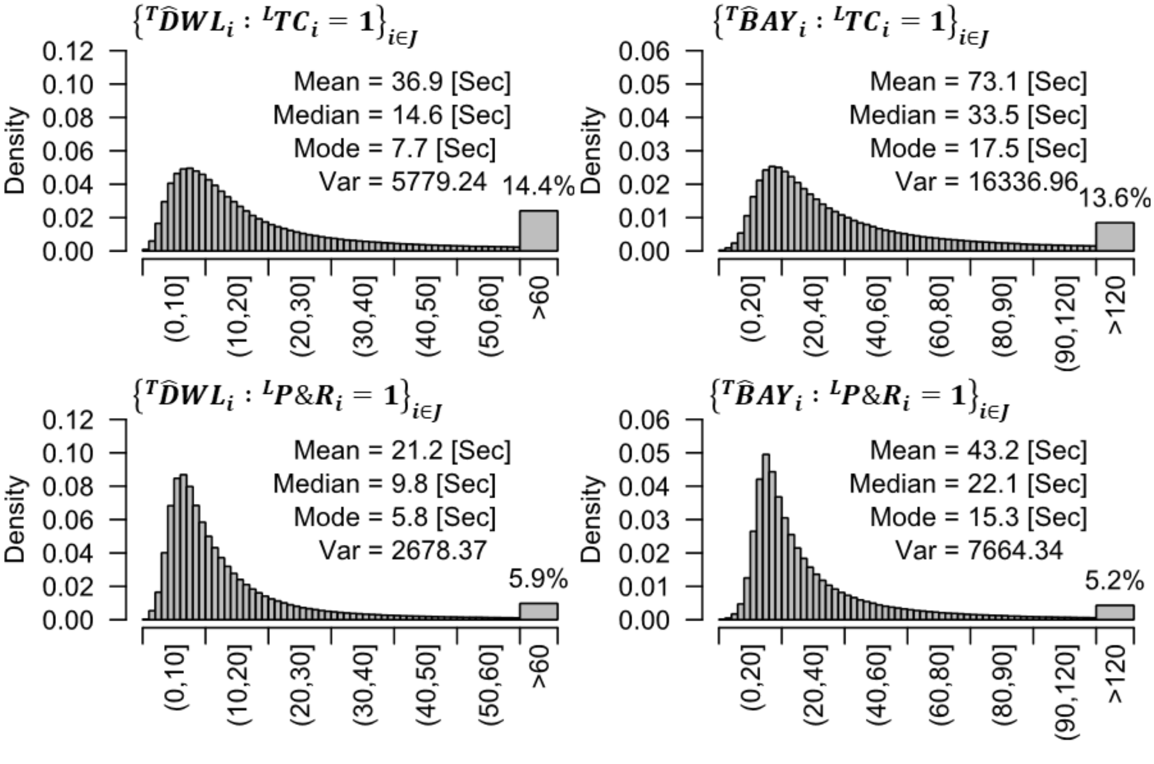

Figure 4-8 — Histograms of $\{{}^{T}\widehat{D}WL_i : i \in J\}$ (Left) and $\{{}^{T}\widehat{B}AY_i : i \in J\}$ (Right) for transit centers (${}^{L}TC \geq 1$) (Top) and park-and-rides (${}^{L}P\&R \geq 1$) (Bottom).

Downtown Transit Mall

Within Portland, bus stops on the downtown transit mall $({}^{L}MALL)$ are also known to behave differently from other locations. This is primarily due to the requirement to stop at all bus-stop location regardless of passenger activity. For transit centers and park-and-rides, the histograms for door open duration and bus-bay stop duration have similar shapes. This is not true on the downtown transit mall (Figure 4-9) (TriMet, 2020).

Definition 4-12 — ${}^{L}MALL$ [$\mathbb{B}$] is a *Transit Mall* variable with binary units used to indicate if a stop is located on 5th or 6th Avenue in downtown core of Portland.

Figure 4-10 shows that the histogram of $\left\{{}^{T}\widehat{D}WL_i : {}^{L}MALL_i = 1\right\}$ has a similar distribution to $\left\{{}^{T}\widehat{D}WL_i : {}^{L}TC_i = 1\right\}$ from Figure 4-8, but $\left\{{}^{T}\widehat{B}AY_i : {}^{L}MALL_i = 1\right\}$ is distinctly bi-modal where $\left\{{}^{T}\widehat{B}AY_i : {}^{L}TC_i = 1\right\}$ was not. The same trends are true for the data after taking the natural logarithm. The causes for this irregular distribution of stop durations may be attributed to several factors: vehicles are required to stop regardless of activity, vehicles often wait while other vehicles pass, and the downtown transit mall has a high density of signalized intersections. The influence of each of these factors is discussed as part of the regressions at the stop-event and aggregated levels.

The distinct differences in durations and distributions at both transit centers and for the transit mall provide reasons to separate these stops from the other location variables. ${}^{L}NEAR_i$, ${}^{L}FAR_i$, ${}^{L}OPP_i$, and ${}^{L}AT_i$, along with their corresponding ${}^{Ls}VAR_i$ and ${}^{Lu}VAR_i$, are assigned values of zero when ${}^{L}TC_i > 0$ or ${}^{L}MALL_i$>0. Like other location variables, ${}^{L}TC_i$ and ${}^{L}MALL_i$ can be categorized into their signalized (i.e. ${}^{Ls}TC_i$ and ${}^{Ls}MALL_i$) and unsignalized (i.e. ${}^{Lu}TC_i$ and ${}^{Lu}MALL_i$) partitions.

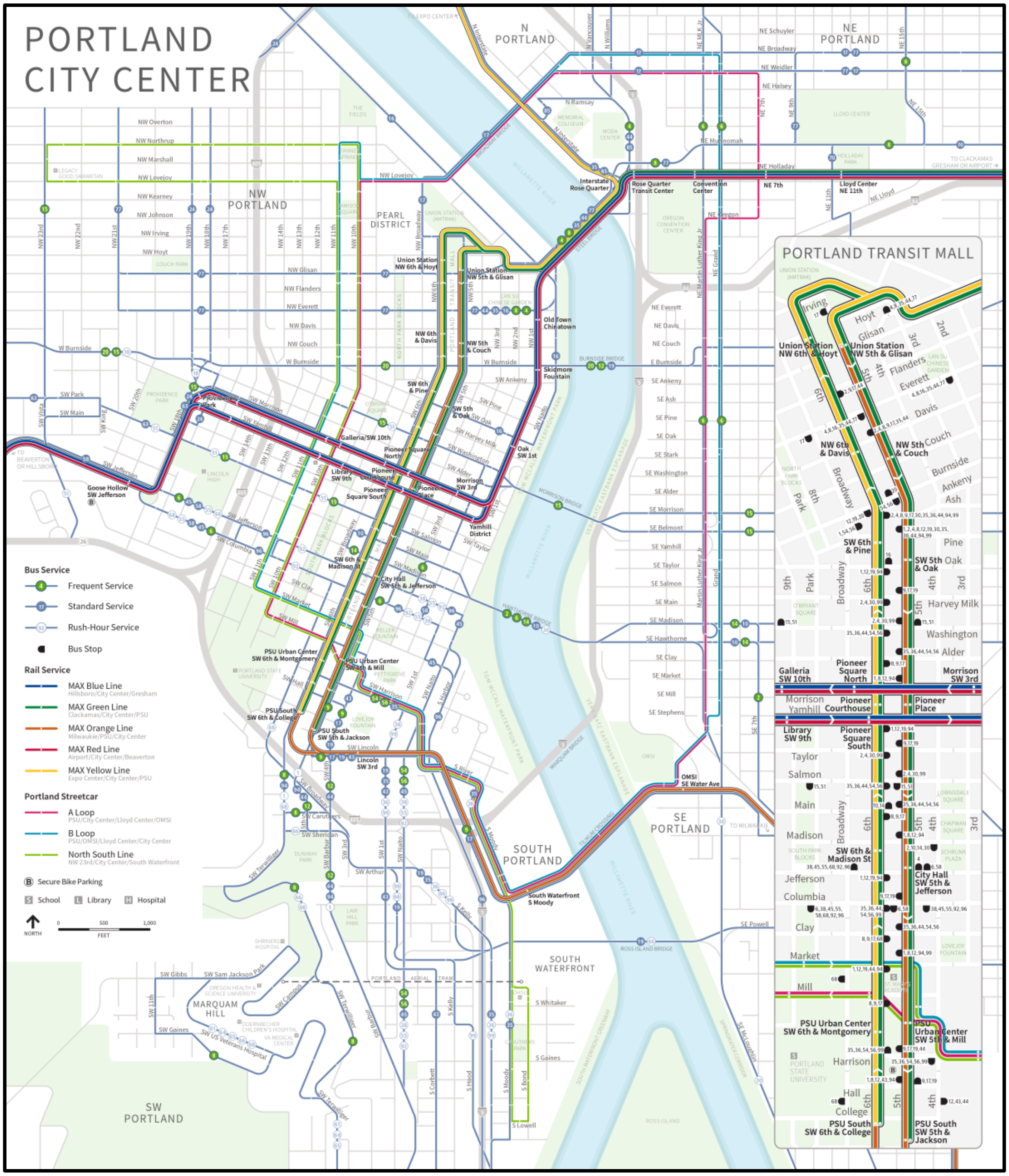

Figure 4-9 — TriMet stylized map of the downtown transit mall. Visit https://trimet.org/maps/img/citycenter.png for a full-size image.

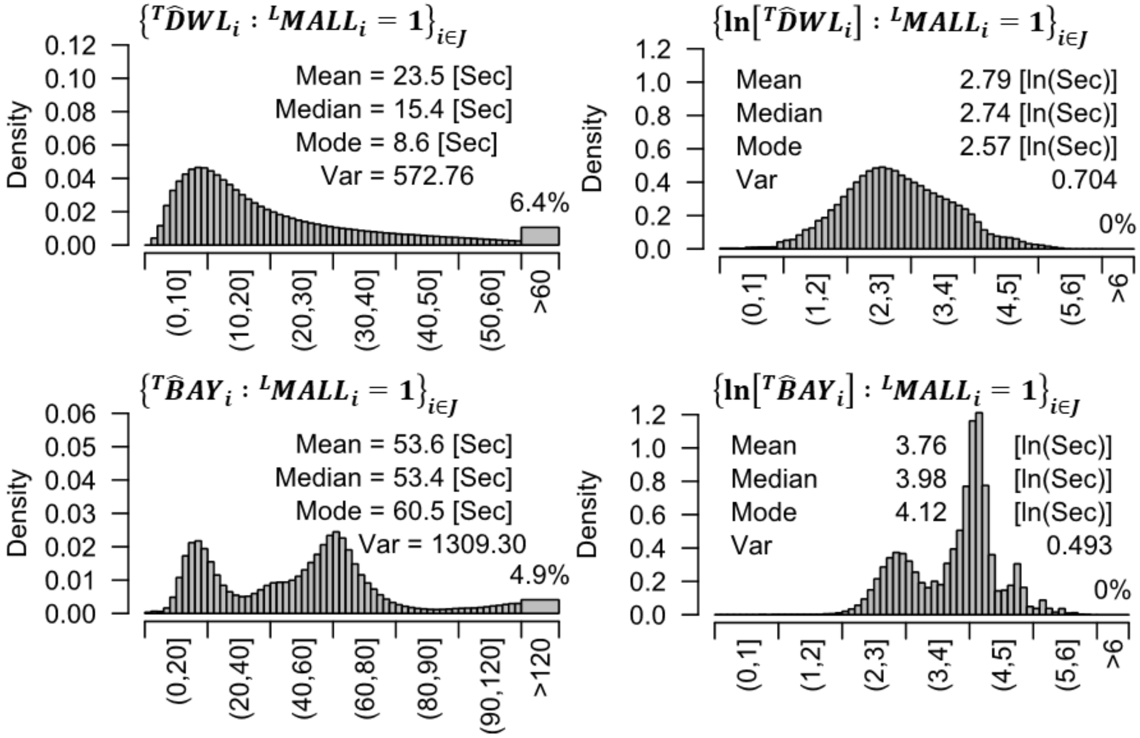

Figure 4-10 — Histograms on the transit mall for $\{{}^{T}\widehat{D}WL_i : \phi\}$ (Top-Left), $\{\ln[{}^{T}\widehat{D}WL_i] : \phi\}$ (Top-Right), $\{{}^{T}\widehat{B}AY_i : \phi\}$ (Bottom-Left), and $\{\ln[{}^{T}\widehat{B}AY_i] : \phi\}$ (Bottom-Right), *where* $\phi \coloneqq ({}^{L}MALL_i = 1) \wedge (i \in J)$.

4.3.2. Passenger Movements

Passenger movements are primary contributors to service durations within bus-bays when buses stop to serve passengers. Previous research has shown that increasing values of $\widehat{O}NS_i$ and $\widehat{O}FFS_i$ also increases ${}^{T}\widehat{D}WL_i$, but that the increase is non-linear. Each additional movements adds less time than the previous movement. With models for ${}^{T}\widehat{D}WL_i$, these economies of scale have previously been captured by including the square terms of $\widehat{O}NS_i$ and $\widehat{O}FFS_i$.

Definition 4-13 — $\widehat{O}NS_i^2$ and $\widehat{O}FFS_i^2$ [pax^2] are the square of the corresponding $\widehat{O}NS_i$ and $\widehat{O}FFS_i$, respectively. Both values are calculated from the cleaned dataset.

For all service events $({}^{E}SVC)$ in the system, the means and variance for the number of boarding and alighting passengers is approximately the same. Table 4-3 show the statistics for boardings, alightings, and their sum when ${}^{T}\widehat{D}WL_i > 0$. Approximately 40% of all stops do not have passengers entering a vehicle and 90% of stops have three or fewer. The same relationship is true when examining passengers exiting vehicles. Examined together, the average stop has between two and three passenger movements and less than 8% of stops, where the door opens, will have zero passenger movements.

Table 4-3 — Statistics for $\{\hat{O}NS_i : \phi\}$, $\{\hat{O}FFS_i : \phi\}$, and $\{\hat{O}NS_i + \hat{O}FFS_i : \phi\}$,
$where\ \phi \coloneqq \left({}^{T}\widehat{D}WL_i > 0\right) \land \left(i \in (I_0 \cup J)\right)$.

	$\{\hat{O}NS_i : \phi\}$		$\{\hat{O}NS_i : \phi\}$		$\{\hat{O}NS_i + \hat{O}FFS_i : \phi\}$	
Mean	1.235		1.236		2.471	
Variance	3.476		3.366		6.586	
Number	*Percent*	*Cumulative*	*Percent*	*Cumulative*	*Percent*	*Cumulative*
0	42.1%	42.1%	42.3%	42.3%	7.4%	7.4%
1	30.7%	72.9%	29.7%	72.0%	38.9%	46.3%
2	13.0%	85.9%	13.6%	85.6%	21.3%	67.6%
3	6.0%	91.9%	6.3%	92.0%	12.0%	79.6%
4	3.1%	95.0%	3.2%	95.1%	7.0%	86.6%
5	1.7%	96.7%	1.7%	96.9%	4.3%	90.8%
6	1.0%	97.8%	1.0%	97.9%	2.7%	93.5%
7	0.7%	98.5%	0.6%	98.5%	1.8%	95.3%
8	0.4%	98.9%	0.4%	99.0%	1.2%	96.6%
9	0.3%	99.2%	0.3%	99.2%	0.9%	97.4%
10	0.2%	99.4%	0.2%	99.4%	0.6%	98.1%
>10	0.6%	100.0%	0.6%	100.0%	1.9%	100.0%

The time-of-day and day of the week have a large influence on average passenger activity. Typically, schedules are different for weekdays and weekends as they have different number of total passengers and different usage curves. Figure 4-11 shows the total hourly statistics for each day of the week. While there are differences between commuter patterns on each weekday, the ranges of possible values overlap. This not the case for weekends. The average weekend day does not lie within the confidence interval for Saturday or Sunday. Some agencies, including TriMet, will report values for Saturday and Sunday separately or include weekend total value. For this research, weekdays and weekends will be examined separately. In addition to different demand between weekdays and weekends, the times of peak travel are different and bimodal for weekdays.

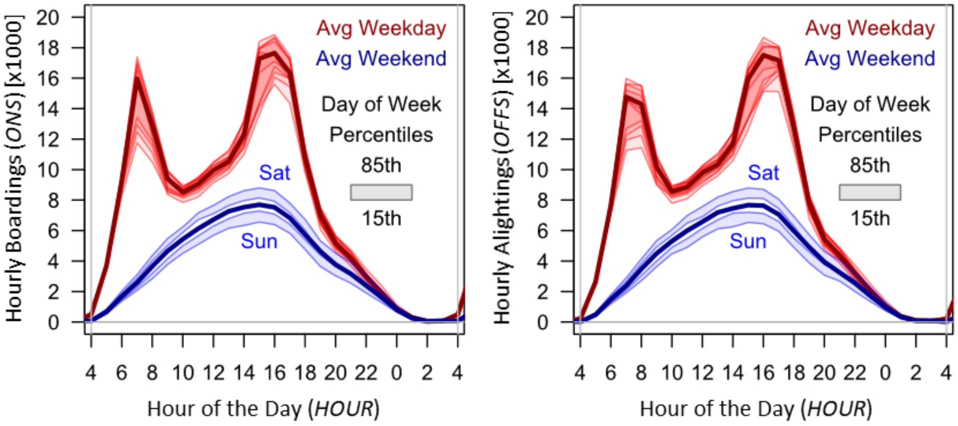

Figure 4-11 — Average hourly boardings and alightings for weekdays and weekend days with percentiles for each day of the week.

For transit centers $({}^{L}TC)$, park-and-rides $({}^{L}P\&R)$, timepoints $({}^{L}TP)$, and transit mall $({}^{L}MALL)$ stops, some of their effect on ${}^{T}\widehat{D}WL$ may be attributed to the differences in passenger movements. Table 4-4 gives the average boardings and alightings select hours of the day and Figure 4-12 plots averages verses time of day for each location type. During all times of day, average $\left\{\widehat{O}NS_i : ({}^{L}TP_i = 1) \wedge \left({}^{T}\widehat{D}WL_i > 0\right) \wedge \left(i \in (I_0 \cap J)\right)\right\}$ has more than double the average $\left\{\widehat{O}NS_i : \left({}^{T}\widehat{D}WL_i > 0\right) \wedge \left(i \in (I_0 \cap J)\right)\right\}$. The other locations also increase passenger movements, but not to the same degree.

Table 4-4 — Average boardings and alightings, dependent on time-of-day and location type (i.e. transit centers (^{L}TC), park-and-rides ($^{L}P\&R$), timepoints (^{L}TP), and the downtown transit mall ($^{L}MALL$)).

	Boardings (ONS)						**Alightings ($OFFS$)**					
Hour	**All**	^{L}TC	$^{L}P\&R$	^{L}TP	$^{L}MALL$	**Rest**	**All**	^{L}TC	$^{L}P\&R$	^{L}TP	$^{L}MALL$	**Rest**
[02:00-06:00)	1.14	2.09	1.41	1.58	1.07	*1.02*	0.79	1.92	1.11	1.41	1.86	*0.57*
[06:00-10:00)	1.21	2.62	1.58	1.82	1.22	*1.07*	1.20	2.60	1.48	2.06	3.00	*0.90*
[10:00-14:00)	1.22	2.73	1.48	2.05	1.68	*1.01*	1.21	2.68	1.63	2.02	1.98	*0.97*
[14:00-18:00)	1.39	3.44	1.60	2.58	3.15	*1.07*	1.37	3.18	1.88	2.22	1.64	*1.16*
[18:00-22:00)	1.10	2.86	1.40	2.03	2.27	*0.83*	1.21	2.28	1.48	1.71	1.16	*1.10*
[22:00-02:00)	0.93	2.33	1.27	1.61	1.73	*0.70*	1.03	1.65	1.22	1.33	0.83	*0.96*
All Hours	**1.24**	**2.88**	**1.51**	**2.11**	**2.07**	***1.00***	**1.24**	**2.64**	**1.61**	**1.98**	**1.93**	***1.02***

Examined visually, transit centers show high activity throughout the day, but the highest activity is on briefly on the downtown transit mall for one or two hours. The transit mall shows a commuter pattern, as a work destination, with high average boardings in the evening and high average alightings in the morning.

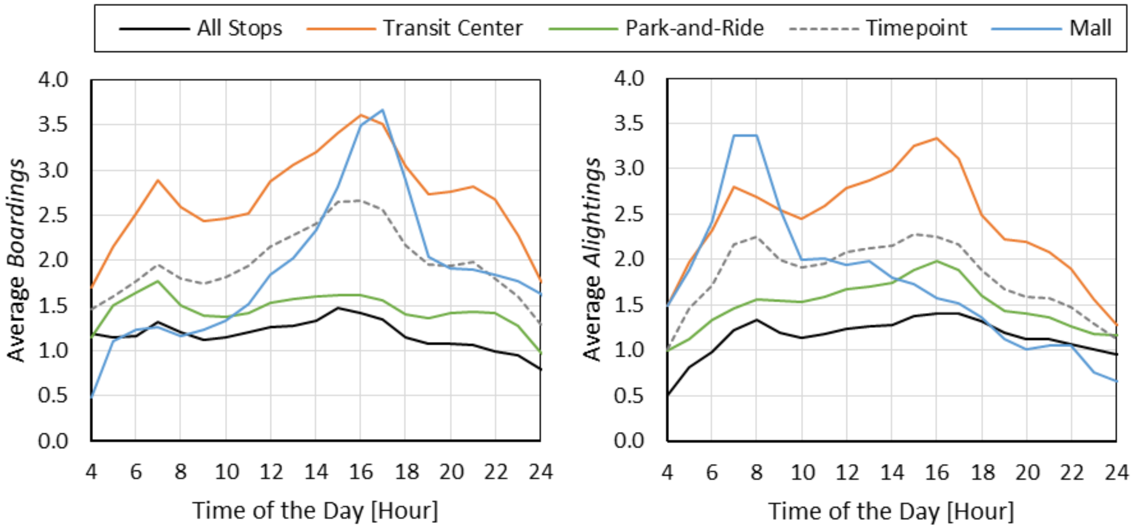

Figure 4-12 — Average boardings and alightings verses time-of-day and for transit centers (^{L}TC), park-and-rides ($^{L}P\&R$), timepoints (^{L}TP), and the downtown transit mall ($^{L}MALL$).

Wheelchair Ramps

Another type of passenger movement is the use of a wheelchair ramp. Ramps take about 30 seconds to deploy and therefore add to service times. These events are somewhat rare, but also show hourly and location-based variation. Table 4-5 shows the mean number of $\widehat{LIFT}$ per 1,000 stops in 4-hour intervals.

Definition 4-14 — $\widehat{LIFT}$ [$\mathbb{B}$] is binary variable indicating *Wheelchair Ramp Deployment.* $\widehat{LIFT}_i$ was also cleaned as the original variable was not consistently binary.

Wheelchair ramp deployment does not follow the AM/PM commuter pattern of passenger boardings and alightings; all days of the week follow a curve similar to weekend travel. Weekends have higher average activity when compared to weekdays; the average Saturday, Sunday, and weekdays record about one wheelchair lift event per 220, 235, and 260 stops, respectively.

Table 4-5 — Wheelchair ramp deployment ($\widehat{LIFT}$) per 1,000 service events (^{E}SVC) for transit centers (^{L}TC), park-and-rides ($^{L}P\&R$), timepoints (^{L}TP), and the downtown transit mall ($^{L}MALL$).

	$\widehat{LIFT}$ per 1,000 ^{E}SVC					
Hour	**All**	**^{L}TC**	**^{L}TP**	**$^{L}MALL$**	**$^{L}P\&R$**	***Rest***
[02:00-06:00)	5.10	11.64	7.79	14.35	6.66	*4.07*
[06:00-10:00)	8.86	17.79	14.62	12.93	10.6	*7.15*
[10:00-14:00)	22.73	38.60	34.77	29.99	26.11	*19.32*
[14:00-18:00)	18.15	37.63	29.60	23.16	22.54	*15.18*
[18:00-22:00)	13.48	24.30	19.69	15.08	16.35	*11.69*
[22:00-02:00)	9.23	15.83	12.44	11.36	10.96	*8.18*
All Hours	**15.43**	***28.85***	***23.95***	***19.64***	***18.56***	***13.02***

Transit centers, timepoints, the transit mall, and park-and-rides all have elevated $\widehat{LIFT}$ activity, as compared to the system average. During peak lift activity, between 10:00 and 14:00, ^{L}TC, ^{L}TP, $^{L}MALL$, and $^{L}P\&R$ stops experience an average of about one $\widehat{LIFT}$

event per 105, 115, 135, and 155 service stops, respectively. All other locations experience less than one $\hat{L}IFT$ per 200 stops during the same period.

4.3.3. Bus Interactions

The amount of time a bus spends a bus stop is affected by its physical proximity to other buses at that stop and the order of each vehicle's arrival and departure times. Previous research has provided a means to categorize these interactions into four main categories (Glick & Figliozzi, 2019). Figure 4-13 shows space-time diagrams for interaction scenarios: (1) non-interacting vehicles, (2) and (3) the four cases defined by previous research. Interaction (I) variables have values greater than 0 for vehicles and stops where ${}^{T}\hat{B}AY_i > 0$; as such, *Bus A* and *Bus B*, from interaction scenario (4), do not have a defined interaction type.

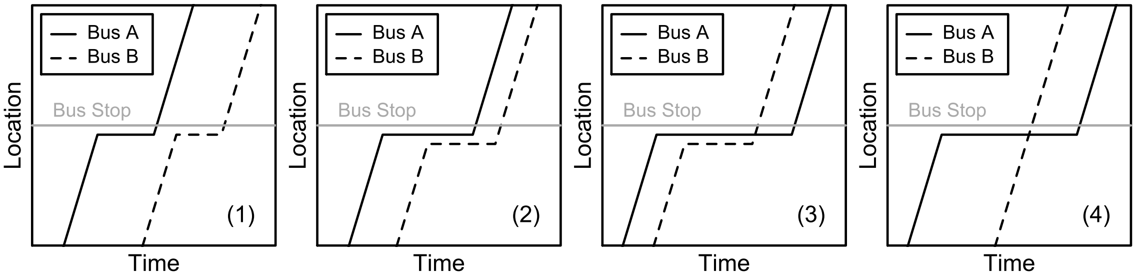

Figure 4-13 — Space-time diagrams for bus interactions scenarios. See Table 4-6 for order of events and variable classification.

Each bus-stop, for one day, will be indexed by $\mathscr{b}$, which is defined as a unique intersection of ℓ, and $\mathscr{d}$. To index ordered events in $\mathscr{b}$, the index b will be used, such that $b \in \mathscr{b} = \{b_1, b_2, \dots, b_n\}$ and $(i \mapsfrom b)$ is a function that maps index i from index b. The family of all $\mathscr{b}$ is contained in $\mathbb{b}$.

(4.3.3) $$\forall \mathscr{b} \in \mathbb{b}, (\mathscr{b} = \bigcap \mathscr{b}' : \mathscr{b}' \in \mathbb{b}'),$$

Where: $\mathbb{b}' = \mathbb{l} \times \mathbb{d} = \{(\ell, d) : ((\ell \in \mathbb{l}) \wedge (d \in \mathbb{d}))\}$.

The order of vehicle arrival times $({}^{t}ARR_{b})$ and departure times $({}^{t}DEP_{b})$ is given in Table 4-6 and the primary interaction types are defined below. The equations are applicable to vehicle, $(i \mapsfrom b)$, based on its interaction with the next vehicle's arrival and departure times (i.e. ${}^{t}ARR_{b+1}$ and ${}^{t}DEP_{b+1}$) or the previous vehicle's arrival and departure times (i.e. ${}^{t}ARR_{b-1}$ and ${}^{t}DEP_{b-1}$) at a given bus-bay.

Table 4-6 — Order of events and independent variable names for bus interaction (I) scenarios (1) – (3) in Figure 4-13.

					Bus Interaction Variable	
Scenario	Order of Events at Bus Stop				Bus A	Bus B
(1)	Bus A Arrives	Bus A Departs	Bus B Arrives	Bus B Departs	NA	NA
(2)	Bus A Arrives	Bus B Arrives	Bus A Departs	Bus B Departs	${}^{I}LEAD$	${}^{I}TAIL$
(3)	Bus A Arrives	Bus B Arrives	Bus B Departs	Bus A Departs	${}^{I}WAIT$	${}^{I}JUMP$

More than two vehicles may interact at a given stop. For all stopping vehicles, interactions are classified for each pair of interacting vehicles as binary variables. These binary variables are summed to give the total number of interactions of each type for each vehicle. The resulting variables have positive integer.

Definition 4-15 — ${}^{I}LEAD_{i \mapsfrom b}$ $[\mathbb{N}_0]$ is a *Leading Interaction* for $VEH_{i \mapsfrom b}$ and gives the number of *Bus A* interactions from Scenario (2).

Definition 4-16 — ${}^{I}TAIL_{i \mapsfrom b}$ $[\mathbb{N}_0]$ is a *Tailing Interaction* for $VEH_{i \mapsfrom b}$ and gives the number of *Bus B* interactions from Scenario (2).

Definition 4-17 — ${}^{I}WAIT_{i \mapsfrom b}$ $[\mathbb{N}_0]$ is a *Waiting Interaction* for $VEH_{i \mapsfrom b}$ and gives the number of *Bus A* interactions from Scenario (3).

Definition 4-18 — ${}^{I}JUMP_{i \mapsfrom b}$ $[\mathbb{N}_0]$ is a *Jumping Interaction* for $VEH_{i \mapsfrom b}$ and gives the number of *Bus B* interactions from Scenario (3).

$$\text{(4.3.4)} \qquad \begin{Bmatrix} \forall {}^{I}LEAD_{i \mapsfrom b} \in \{{}^{I}LEAD_b\}_{b \in \mathbb{b}}, \left({}^{I}LEAD_b = \sum_{\ddot{\boldsymbol{b}}=(b+1)}^{b+5} [\mathbf{1}_{\{\phi_{LEAD}\}}]\right) \\ \forall {}^{I}TAIL_{i \mapsfrom b} \in \{{}^{I}TAIL_b\}_{b \in \mathbb{b}}, \left({}^{I}TAIL_b = \sum_{\ddot{\boldsymbol{b}}=(b-5)}^{b-1} [\mathbf{1}_{\{\phi_{TAIL}\}}]\right) \\ \forall {}^{I}WAIT_{i \mapsfrom b} \in \{{}^{I}WAIT_b\}_{b \in \mathbb{b}}, \left({}^{I}WAIT_b = \sum_{\ddot{\boldsymbol{b}}=(b+1)}^{b+5} [\mathbf{1}_{\{\phi_{WAIT}\}}]\right) \\ \forall {}^{I}JUMP_{i \mapsfrom b} \in \{{}^{I}JUMP_b\}_{b \in \mathbb{b}}, \left({}^{I}JUMP_b = \sum_{\ddot{\boldsymbol{b}}=(b-5)}^{b-1} [\mathbf{1}_{\{\phi_{JUMP}\}}]\right) \end{Bmatrix},$$

Given that: $\ddot{\boldsymbol{b}}$ is a placeholder index; *and*, the conditionals, ϕ, from the indicator functions ($\mathbf{1}_{\{\phi\}}$) are defined as:

$$\begin{Bmatrix} \phi_{LEAD} \coloneqq \phi_{\{b,\ddot{\boldsymbol{b}}\}} \wedge ({}^{t}ARR_b < {}^{t}ARR_{\ddot{\boldsymbol{b}}} < {}^{t}DEP_b < {}^{t}DEP_{\ddot{\boldsymbol{b}}}) \\ \phi_{TAIL} \coloneqq \phi_{\{b,\ddot{\boldsymbol{b}}\}} \wedge ({}^{t}ARR_{\ddot{\boldsymbol{b}}} < {}^{t}ARR_b < {}^{t}DEP_{\ddot{\boldsymbol{b}}} < {}^{t}DEP_b) \\ \phi_{WAIT} \coloneqq \phi_{\{b,\ddot{\boldsymbol{b}}\}} \wedge ({}^{t}ARR_b < {}^{t}ARR_{\ddot{\boldsymbol{b}}} < {}^{t}DEP_{\ddot{\boldsymbol{b}}} < {}^{t}DEP_b) \\ \phi_{JUMP} \coloneqq \phi_{\{b,\ddot{\boldsymbol{b}}\}} \wedge ({}^{t}ARR_{\ddot{\boldsymbol{b}}} < {}^{t}ARR_b < {}^{t}DEP_b < {}^{t}DEP_{\ddot{\boldsymbol{b}}}) \\ \phi_{\{b,\ddot{\boldsymbol{b}}\}} \coloneqq \left({}^{T}\widehat{B}AY_b > 0\right) \wedge \left({}^{T}\widehat{B}AY_{\ddot{\boldsymbol{b}}} > 0\right) \end{Bmatrix}.$$

For each interaction categorized for vehicle b, there is only one interaction type per vehicle overlap. ${}^{I}LEAD_{i \mapsfrom b}$ and ${}^{I}WAIT_{i \mapsfrom b}$ can only apply if vehicle b is the first of the pair to arrive while ${}^{I}TAIL_{i \mapsfrom b}$and ${}^{I}JUMP_{i \mapsfrom b}$ can only apply if vehicle b is the second of the pair to arrive. In all cases, if there is no overlap between stop times or if either vehicle does not stop, then none of the interaction variables apply.

There are other interactions that need to be considered that cannot be classified into the categories above due to data limitations. Sometimes, buses record the same arrival and/or departure times at bus stops. Table 4-7 shows the order of events at a given bus stop and the interaction variable assigned when there is an exact overlap in the data. These variables, in Table 4-7, are not necessarily unique like those in Table 4-6. Cases (4) – (6), in Table 4-7 may fall into one of two possibilities.

Table 4-7 — Order of events and independent variable names for bus interactions with identical arrival and/or departure times.

		Bus Interaction Variable	
Scenario	Order of Events at Bus Stop	Bus A	Bus B
(4)	Buses A and B Arrive Bus A Departs Bus B Departs	${}^{I}SAME.LEAD$	${}^{I}SAME.TAIL$
(5)	Bus A Arrives Bus B Arrives Buses A and B Depart	${}^{I}LEAD.SAME$	${}^{I}TAIL.SAME$
(6)	Buses A and B Arrive Buses A and B Depart	${}^{I}SAME.SAME$	

For example, vehicle A, assigned ${}^{I}SAME.LEAD = 1$, had the same recorded arrival time, but earlier departure than vehicle B; given the overlap, it is unknown if vehicle A was actually a leading vehicle or jumping vehicle. None of the variables, defined in Table 4-7, are used directly. Instead, they are combined according to equation (4.3.5) into an "other" variable (i.e. ${}^{I}SAME_{i \mapsfrom b}$), which is only used to define the total count of all interactions for vehicle $i \mapsfrom b$.

Definition 4-19 — ${}^{I}INT_{i \mapsfrom b}$ [$\mathbb{N}_0$] is the sum of all *Interactions* for $VEH_{i \mapsfrom b}$ for one bus-bay stopping event.

(4.3.5)

$$\forall {}^{I}INT_{i \mapsfrom b} \in \{{}^{I}INT_b\}_{b \in \mathbb{b}} \,, \quad \left({}^{I}INT_b = ({}^{I}LEAD_b + {}^{I}TAIL_b + {}^{I}WAIT_b + {}^{I}JUMP_b + {}^{I}SAME_b)\right),$$

Where: ${}^{I}SAME_b = \sum_{\ddot{\boldsymbol{b}}=(b-5)}^{b-1}[\mathbf{1}_{\{\phi_{SAME}\}}] + \sum_{\ddot{\boldsymbol{b}}=(b+1)}^{b+5}[\mathbf{1}_{\{\phi_{SAME}\}}]$; *and, given that*:

$$\left\{\begin{matrix} \phi_{SAME} \coloneqq \phi_{\{b,\ddot{\boldsymbol{b}}\}} \land \left(({}^{t}ARR_b = {}^{t}ARR_{\ddot{\boldsymbol{b}}}) \lor ({}^{t}DEP_b = {}^{t}DEP_{\ddot{\boldsymbol{b}}})\right) \\ \phi_{\{b,\ddot{\boldsymbol{b}}\}} \coloneqq \left({}^{T}\widehat{B}AY_b > 0\right) \land \left({}^{T}\widehat{B}AY_{\ddot{\boldsymbol{b}}} > 0\right) \end{matrix}\right\}.$$

Additionally, buses may come from the same routes or from different routes. Each interaction variable is divided between two additional variables that apply when buses are from the same route (Is) or from two different routes (Id). The superscript, such that ${}^{I}LEAD_i = {}^{Is}LEAD_i + {}^{Id}LEAD_i$. If interactions are not broken down by type, the variables, ${}^{I}INT$, ${}^{Is}INT$, and ${}^{Id}INT$ are used to represent interactions for vehicles of all routes, same routes, and different routes, respectively.

Definition 4-20 — ${}^{Is}INT_{i\hookleftarrow b}$ $[\mathbb{N}_0]$ is the sum of all *Same-Route Interactions* for $VEH_{i\hookleftarrow b}$ for one bus-bay stopping event.

Definition 4-21 — ${}^{Id}INT_{i\hookleftarrow b}$ $[\mathbb{N}_0]$ is the sum of all *Different-Route Interactions* for $VEH_{i\hookleftarrow b}$ for one bus-bay stopping event.

4.3.4. Headways

Where $b \in \mathscr{b} = \{b_1, b_2, \dots, b_n\}$ is the set of ordered events for one stop on one day, $\dot{b} \in \dot{\mathscr{b}} = \{\dot{b}_1, \dot{b}_2, \dots, \dot{b}_n\}$ is the set of ordered events for one stop, one day, and for one route direction. $\dot{\mathscr{b}}$, is defined as a unique intersection of ℓ, $\mathscr{r}_d$, and $\mathscr{d}$ and the family of all $\dot{\mathscr{b}}$ is contained in $\dot{\mathbb{b}}$.

(4.3.6) $$\forall \dot{\mathscr{b}} \in \dot{\mathbb{b}}, \left(\dot{\mathscr{b}} = \bigcap \dot{\mathscr{b}}' : \dot{\mathscr{b}}' \in \dot{\mathbb{b}}'\right),$$

Where: $\dot{\mathbb{b}}' = \mathbb{l} \times \mathbb{r}_{\mathrm{d}} \times \mathbb{d} = \left\{(\ell, \mathscr{r}_{\mathrm{d}}, \mathscr{d}) : \left((\ell \in \mathbb{l}) \wedge (\mathscr{r}_{\mathrm{d}} \in \mathbb{r}_{\mathrm{d}}) \wedge (\mathscr{d} \in \mathbb{d})\right)\right\}$.

The gap between scheduled times $({}^{t}SKD)$ or observed service times $({}^{t}ARR$ or ${}^{t}DEP)$ for two consecutive vehicles from the same route, is a headway (HW). Headways calculated from ${}^{t}SKD$ will be denoted with the superscript (S); headways calculated from

${}^{t}ARR$ or ${}^{t}DEP$ will be denoted with the superscripts (A) or (D), respectively. For this research, headways are defined at specific locations.

Definition 4-22 — ${}^{A}H_{i\mapsfrom\dot{b}}$, ${}^{D}H_{i\mapsfrom\dot{b}}$, and ${}^{S}H_{i\mapsfrom\dot{b}}$ [sec] are the time differences for *Arrivals*, *Departures*, and *Scheduled Service* between two consecutive vehicles of the same route servicing a given stop.

$$(4.3.7) \qquad \left\{\begin{array}{l} \forall {}^{A}H_{i\mapsfrom\dot{b}} \in \{{}^{A}H_i\}_{i\in\bigcup\dot{\mathbb{b}}}\,, \left({}^{A}H_{\dot{b}} = {}^{t}ARR_{\dot{b}} - {}^{t}ARR_{\dot{b}-1} \,:\, \dot{b} \neq \dot{b}_1\right) \\ \forall {}^{D}H_{i\mapsfrom\dot{b}} \in \{{}^{D}H_i\}_{i\in\bigcup\dot{\mathbb{b}}}\,, \left({}^{D}H_{\dot{b}} = {}^{t}DEP_{\dot{b}} - {}^{t}DEP_{\dot{b}-1} \,:\, \dot{b} \neq \dot{b}_1\right) \\ \forall {}^{S}H_{i\mapsfrom\dot{b}} \in \{{}^{S}H_i\}_{i\in\bigcup\dot{\mathbb{b}}}\,, \left({}^{S}H_{\dot{b}} = {}^{t}SKD_{\dot{b}} - {}^{t}SKD_{\dot{b}-1} \,:\, \dot{b} \neq \dot{b}_1\right) \end{array}\right\}$$

By the formulation in equations (4.3.7), the first vehicle to reach a stop on a given day, for one route, will not have an associated headway. While there is technically a measurable gap between the first vehicle on a given day and the last vehicle on the previous day, it is not considered a headway for this study.

On average, ${}^{D}H_{i\mapsfrom\dot{b}}$ is about eight seconds longer than ${}^{A}H_{i\mapsfrom\dot{b}}$ at locations with a service event $({}^{E}SVC)$. However, at timepoints, where vehicles arrive early, ${}^{D}H_i$ is an average of 20 seconds longer, but four seconds shorter when vehicles are running late. The differences between the average arrival and departure headways of early versus late vehicles is weak evidence of schedule maintenance ${}^{L}TP$ stops.

4.3.5. Congestion

A method for measuring congestion using stop event data was outlined in a publication by Furth and Halawani (2018). That research described five components that may be used quantify the financial impact of congestion for transit agencies and transit users. The general methodology from that publication will be adapted for this research. The variable names and notation from that publication will be updated to match the system used in this research.

The methodology from Furth and Halawani requires a baseline time-period that serves as the comparison for other time periods. As such trips will be sorted based on the hour in which they began; however, trips beginning before 06:00 PLT or after 20:00 PLT will be grouped together and will serve as the baseline time-period. The index $\mathscr{r}_w$ is defined as a subset of I that includes all observations, VAR_i, that occurred within for the same route-direction on either weekdays or weekends. The set of all $\mathscr{r}_w$ are contained in $\mathbb{r}_w$.

(4.3.8) $$\forall \mathscr{r}_w \in \mathbb{r}_w, \left(\mathscr{r}_w = \bigcap \mathscr{r}'_w : \mathscr{r}'_w \in \mathbb{r}'_w\right),$$

Where: $\mathbb{r}'_w = \mathbb{r}_d \times \mathbb{w} = \left\{(\mathscr{r}_d, \mathscr{w}) : \left((\mathscr{r}_d \in \mathbb{r}_d) \wedge (\mathscr{w} \in \mathbb{w})\right)\right\}$.

For each $\mathscr{r}_w$, we need a set of unique time-periods. $\dot{\mathscr{h}}$ is a modified index for hours defined by $\mathscr{h} \in \mathbb{h}$ corresponding to the start of each transit trip, $\mathscr{a}$. $\dot{\mathscr{h}}$ is further modified such that off-peak time periods are combined into a single index.

(4.3.9) $$\dot{\mathscr{h}} \in \dot{\mathbb{h}} = \left\{\begin{matrix} \dot{\mathscr{h}}_0 = \bigcup\{\mathscr{h}_a \in (\{0, 1, \dots, 5, 20, \dots, 23\} \subset \mathbb{h}_a)\} \\ \dot{\mathscr{h}}_6 = (\mathscr{h}_a = 6) \\ \vdots \\ \dot{\mathscr{h}}_{19} = (\mathscr{h}_a = 19) \end{matrix}\right\},$$

Where: $\forall \mathscr{h}_a \in \mathbb{h}_a = \{0, 1, 2, \dots, 23\}, \left(\mathscr{h}_a \coloneqq (\mathscr{h} \ni \{\min[{}^tDEP_i] : i \in \mathscr{a}\})\right)$.

Unique intersections of $\mathscr{r}_w$ and $\dot{\mathscr{h}}$ are used to calculate congestion. And for each unique $\mathscr{r}_w$, there will be a time-period, $\dot{\mathscr{h}}_0$, that will serve as a baseline for all other $\left(\dot{\mathscr{h}} \neq \dot{\mathscr{h}}_0\right)$. An additional index $\mathscr{p}$ is defined by the intersection of $\mathscr{r}_w$ and $\dot{\mathscr{h}}$, such that the complete set of $\mathscr{p}$ is contained in $\mathbb{p}$.

(4.3.10) $$\forall \mathcal{p} \in \mathbb{p}, (\mathcal{p} = \cap \mathcal{p}' : \mathcal{p}' \in \mathbb{p}')\ ,$$

Where: $\mathbb{p}' = \mathbb{r}_w \times \dot{\mathbb{h}} = \left\{(\mathcal{r}_w, \dot{\mathcal{h}}) : \left((\mathcal{r}_w \in \mathbb{r}_w) \wedge (\dot{\mathcal{h}} \in \dot{\mathbb{h}})\right)\right\}$.

Congestion variables (i.e. ${}^{C}VAR_{\mathcal{p}}$) denote an increase in period $\mathcal{p}$ over the baseline period ($\mathcal{p} = \mathcal{p}_0$). For each period, ${}^{CT}VAR_{\mathcal{p}}$ will denote an increase in elapsed time and ${}^{C\$}VAR_{\mathcal{p}}$ will denotes an increase in costs.

Agencies

For agencies, there is an increased cost due to running times (i.e. ${}^{C\$}RUN$) and a secondary increase due to the recovery time between trips (i.e. ${}^{C\$}RCV$). The simple regression model given in Table 4-8 is not a highly effective model, but provides $\gamma = 21.415$, $\alpha = 2.534$, and $\beta = 4.069$, which represent average time per stop event, average time per alighting passenger, and average time be boarding passenger, respectively.

Table 4-8 — Simple linear regression model, using passenger movements only, for non-zero bus-bay stop durations from all service stops at all times of day. $\forall {}^{T}\widehat{B}AY_i \in \left\{{}^{T}\widehat{B}AY_i : (i \in J) \wedge \left({}^{T}\widehat{B}AY_i < 180\right)\right\}$.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	$\gamma = Intercept$	21.415	0.00455		
Passenger Movements	$\beta = \hat{O}NS$	4.069	0.00179	9.596%	73.14%
	$\alpha = \hat{O}FFS$	2.534	0.00180	3.524%	26.86%
	$n = 45{,}490{,}831$		Adjusted R-Squared = 13.12%		
	p-value ≪ 0.001 for all variables				

For this research, ${}^{CT}\tilde{R}UN_{\mathcal{p}}$ is defined as the estimated running time not related to passenger movements. Accounting for the passenger movements requires coefficients of a simple linear regression model predicting bus-bay stop duration.

Definition 4-23 — ${}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}}$ [sec] is the average time of congestion per trip in period $\mathscr{p}$ resulting from an increase in running time. ${}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}}$ is a intermediary step in calculating time increases.

(4.3.11)

$$\forall\, {}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}} \in \left\{ {}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}} \right\}_{\mathscr{p}\in\mathbb{p}}, \left({}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}} = {}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}} - {}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}_0} : \mathscr{p} \mapsfrom \left(\mathscr{r}_w \cap \dot{\mathscr{h}} \right) \right),$$

Such that: $\mathscr{p}$ and $\mathscr{p}_0$ always refer to the same $\mathscr{r}_w$; *where:*

$$\forall\, {}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}} \in \left\{ {}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}} \right\}_{\mathscr{p}\in\mathbb{p}},$$
$$\left({}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}} = {}_{avg}^{T}\tilde{T}RVL_{\mathscr{p}} - (\gamma)\, {}_{avg}^{E}SVC_{\mathscr{p}} - (\beta)\, {}_{avg}\hat{O}NS_{\mathscr{p}} - (\alpha)\, {}_{avg}\hat{O}FFS_{\mathscr{p}} \right).$$

And where: ${}_{avg}^{T}\tilde{T}RVL_{\mathscr{p}}$, ${}_{avg}^{E}SVC_{\mathscr{p}}$, ${}_{avg}\hat{O}NS_{\mathscr{p}}$, and ${}_{avg}\hat{O}FFS_{\mathscr{p}}$ are the average of trip totals (i.e. summations for each $\mathscr{a} \in \mathbb{a}$ during a given $\mathscr{p}$) for travel time, number of stops, number of boardings, and number of alightings.

By the formulation in equation (4.3.11), there are never increases for off-peak periods. Unfortunately, the formulation does allow for problematic estimates for low usage routes. Specifically, the coefficients of the regression model overestimate the impact of stops and passenger movements, such that run times estimates sometimes come out negative. Luckily this issue can be easily corrected using this research's data set. Equation (4.3.12) is a modification to (4.3.11) that directly calculates the travel time that is not caused by passenger movements.

(4.3.12)

$$\forall\, {}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}} \in \left\{ {}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}} \right\}_{\mathscr{p}\in\mathbb{p}}, \left({}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}} = {}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}} - {}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}_0} : \mathscr{p} \mapsfrom \left(\mathscr{r}_w \cap \dot{\mathscr{h}} \right) \right),$$

Such that: $\mathscr{p}$ and $\mathscr{p}_0$ always refer to the same $\mathscr{r}_w$, *where:*

$$\forall\, {}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}} \in \left\{ {}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}} \right\}_{\mathscr{p}\in\mathbb{p}}, \left({}_{avg}^{C'T}\tilde{R}UN_{\mathscr{p}} = {}_{avg}\left({}^{T}\widetilde{M}OVE + {}^{T}\tilde{D}STB \right)_{\mathscr{p}} \right),$$

Given that: ${}_{avg}\left({}^{T}\widetilde{M}OVE + {}^{T}\tilde{D}STB \right)_{\mathscr{p}}$ is the average of the sum for all trips.

Multiplying ${}^{CT}_{avg}\tilde{R}UN_{\wp}$ by the agency operation costs per unit time, results in the average monetary impact of running time from congestion per trip, assuming matching units after appropriate conversions. Within TriMet for 2019, which was a typical year, the cost to operate a bus per hour is \$139.20 per hour (TriMet, 2019).

$$\forall\, {}^{C\$}_{avg}RUN_{\wp} \in \left\{ {}^{C\$}_{avg}RUN_{\wp} \right\}_{\wp \in \mathbb{p}}, \left(k \cdot \$139.20 \left[\frac{\text{USD}}{\text{hour}}\right] \cdot {}^{CT}_{avg}\tilde{R}UN_{\wp}[\text{sec}]\right), \tag{4.3.13}$$

Where: k is a constant used to convert units.

The secondary consideration for agencies is recovery time. The methodology proposed is based on the requirements of the Massachusetts transit authority that the trips scheduled run time and recovery time is based on the 90th percentile of run time. That methodology was tested, but unfortunately does not directly apply to TriMet operations. TriMet recovery time is based on the service time. TriMet worker regulations require five-minutes of recovery for each one-hour of travel time. As such a conversion factor of $5/60 = 1/12$ may be used and equations (4.3.14) and (4.3.15) define the recovery time and costs as they apply to $\wp$ in TriMet's network.

$$\forall\, {}^{CT}_{avg}\tilde{R}CV_{\wp} \in \left\{ {}^{CT}_{avg}\tilde{R}CV_{\wp} \right\}_{\wp \in \mathbb{p}}, \left({}^{CT}_{avg}\tilde{R}CV_{\wp} = \left(\frac{1}{12}\right) {}^{CT}_{avg}\tilde{R}UN_{\wp} \right) \tag{4.3.14}$$

$$\forall\, {}^{C\$}_{avg}RCV_{\wp} \in \left\{ {}^{C\$}_{avg}RCV_{\wp} \right\}_{\wp \in \mathbb{p}}, \left({}^{C\$}_{avg}RCV_{\wp} = \left(\frac{1}{12}\right) {}^{C\$}_{avg}RUN_{\wp} \right) \tag{4.3.15}$$

Given that the costs are the same and both paid by the agencies, the run time and recovery times may be simplified as a single average "Running & Recovery" variable, (i.e. ${}^{CT}_{avg}\tilde{R}\&R_{\wp}$), and its associated cost (i.e. ${}^{C\$}_{avg}R\&R_{\wp}$).

(4.3.16) $$\forall\, {}_{avg}^{CT}\tilde{R}\&R_{\mathscr{p}} \in \left\{ {}_{avg}^{CT}\tilde{R}\&R_{\mathscr{p}} \right\}_{\mathscr{p}\in\mathbb{p}}, \left({}_{avg}^{CT}\tilde{R}\&R_{\mathscr{p}} = \left(\frac{13}{12}\right) {}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}} \right)$$

(4.3.17) $$\forall\, {}_{avg}^{C\$}R\&R_{\mathscr{p}} \in \left\{ {}_{avg}^{C\$}R\&R_{\mathscr{p}} \right\}_{\mathscr{p}\in\mathbb{p}}, \left({}_{avg}^{C\$}R\&R_{\mathscr{p}} = \left(\frac{13}{12}\right) {}_{avg}^{C\$}RUN_{\mathscr{p}} \right)$$

Estimating the totals for one $\mathscr{p}$ for a day, week, month, or year requires knowing or estimating the number of trips for each $\mathscr{p}$ over that time-frame. These totals are multiplied by their associated averages then summed over the target time-frame.

Passengers

For passengers, there are three values to consider, the ride time within the vehicle, the waiting time at bus stops, and a buffer time impact, which the authors assumed to be based on the 95th percentile of travel time. Riding time was assumed to be 40% of ${}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}}$ based on a previous publication (Furth, 2005) and the authors assumed that buffer time was 75% of the recovery time due to the tendency for passenger to alight towards the end of a trip. Given that recovery time and ride time are proportional to each other within TriMet's network, an average "Ride & Buffer" time (i.e. ${}_{avg}^{CT}\tilde{R}\&B_{\mathscr{p}}$) and associated average cost per passenger (i.e. ${}^{C\$}R\&B_{\mathscr{p}}$) may be defined together.

(4.3.18) $$\forall\, {}_{avg}^{CT}\tilde{R}\&B_{\mathscr{p}} \in \left\{ {}_{avg}^{CT}\tilde{R}\&B_{\mathscr{p}} \right\}_{\mathscr{p}\in\mathbb{p}}, \left({}_{avg}^{CT}\tilde{R}\&B_{\mathscr{p}} = \left(\frac{2}{5} + \frac{1}{12}\cdot\frac{3}{4}\right) {}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}} = \left(\frac{37}{80}\right) {}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}} \right)$$

(4.3.19) $$\forall\, {}^{C\$}R\&B_{\mathscr{p}} \in \left\{ {}^{C\$}R\&R_{\mathscr{p}} \right\}_{\mathscr{p}\in\mathbb{p}}, \left({}^{C\$}R\&R_{\mathscr{p}} = k \cdot \$12 \left[\frac{\text{USD}}{\text{hour}}\right] \cdot \left(\frac{2}{5} + \frac{1}{12}\cdot\frac{3}{4}\cdot\frac{3}{4}\right) \cdot {}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}}[\text{sec}] = k \cdot \$12 \left[\frac{\text{USD}}{\text{hour}}\right] \cdot \left(\frac{143}{320}\right) \cdot {}_{avg}^{CT}\tilde{R}UN_{\mathscr{p}}[\text{sec}] \right),$$

Where: k is a constant used to convert units.

Lastly, the excess waiting time may be calculated considering both long and short transit headways (i.e. frequent and non-frequent service). The excess waiting (i.e. ${}_{avg}^{CT}\tilde{E}XW_{\wp}$) is defined in equation (4.3.20) as the difference between actual expected wait (i.e. ${}_{actual}^{C'T}\tilde{E}XW_{\wp}$) and the ideal expected wait (i.e. ${}_{ideal}^{C'T}\tilde{E}XW_{\wp}$) for period $\wp$. Attention should be given to the use of subscripts $\wp$ and $\wp_0$, as they are both used.

$$(4.3.20)\quad \forall\, {}_{avg}^{CT}\tilde{E}XW_{\wp} \in \left\{{}_{avg}^{CT}\tilde{E}XW_{\wp}\right\}_{\wp\in\mathbb{p}}, \left({}_{avg}^{CT}\tilde{E}XW_{\wp} = {}_{actual}^{C'T}\tilde{E}XW_{\wp} - {}_{ideal}^{C'T}\tilde{E}XW_{\wp}\right),$$

Where:

$${}_{actual}^{C'T}\tilde{E}XW_{\wp} = \left\{\begin{matrix} \dfrac{{}_{var}^{D}H_{\wp}}{(2)\,{}_{avg}^{D}H_{\wp}}\,, & \text{if } {}_{avg}^{S}H_{\wp} < 15\ [\text{min}] \\ mean\left(\{{}_{dev}^{D}H_i\}_{i\in\wp}\right) - P_{0.02}\left(\{{}_{dev}^{D}H_i\}_{i\in\wp}\right), & \text{if } {}_{avg}^{S}H_{\wp} \geq 15\ [\text{min}] \end{matrix}\right\},$$

Given that: ${}_{avg}^{D}H_{\wp}$ and ${}_{var}^{D}H_{\wp}$ are the mean and variances of $\{{}^{D}H_i : i \in \wp\}$, the departure headway; *and*, ${}_{dev}^{D}H_i = |{}^{D}H_i - {}^{S}H_i|$ is the departure deviation; *and where:*

$${}_{ideal}^{C'T}\tilde{E}XW_{\wp} = \left\{\begin{matrix} \dfrac{\begin{pmatrix} 1 + 2(0.25\gamma)^2\left({}_{var}^{E}SVC_{\wp} - {}_{var}^{E}SVC_{\wp_0}\right) \\ + 2(0.5\beta + 0.15\alpha)^2\left({}_{avg}\hat{O}NS_{\wp} - {}_{avg}\hat{O}NS_{\wp_0}\right) \end{pmatrix}}{(2)\,{}_{avg}^{D}H_{\wp}}\,, & \text{if } {}_{avg}^{S}H_{\wp} < 15\ [\text{min}] \\ {}_{actual}^{C'T}\tilde{E}XW_{\wp_0}\,, & \text{if } {}_{avg}^{S}H_{\wp} \geq 15\ [\text{min}] \end{matrix}\right\},$$

Given that: ${}_{var}^{E}SVC_{\wp}$ is the variance of the number of service stops per trip $\mathcal{a}$, for all $\mathcal{a} \cap \wp$; *and*, ${}_{avg}\hat{O}NS_{\wp}$ is the average number of boardings per trip $\mathcal{a}$, for all $\mathcal{a} \cap \wp$.

The value of passenger time was set at $18 per hour while waiting (Furth & Muller, 2006), 1.5 times the in-vehicle riding time of $12. The average cost per passenger, resulting from waiting time in $\wp$ is defined by equation (4.3.21).

(4.3.21)

$$\forall\, {}_{avg}^{C\$}EXW_{\wp} \in \left\{{}_{avg}^{C\$}EXW_{\wp}\right\}_{\wp\in\mathbb{p}}, \left({}_{avg}^{C\$}EXW_{\wp} = k \cdot \$18\left[\frac{\text{USD}}{\text{hour}}\right] \cdot {}_{avg}^{CT}\tilde{E}XW_{\wp}[\text{sec}]\right),$$

Where: k is, again, a constant used to convert units.

The total passenger costs, over a desired time-frame, is calculated by multiplying the number of boarding passengers in each $\mathscr{p}$ by relevant cost averages then summing over the target time-frame. The formulation outlined is not original research. It is an application of a modified version of the methodology from Furth and Halawani. The method will be used with event level data set to provide a means to compare the estimates of cost increases due to congestion when using the aggregated data. While there is some ability to compare different time periods for the ELD, this is primarily a route level methodology.

4.4. Aggregated Variables Definitions

The aggregated variables represent timepoint segments (TPS) with one timepoint location (${}^{L}TP$) and one route-direction, span one hour, and are created using the variables from the Event Level Data (ELD). The index $\mathscr{t}$ is defined as a subset of I that contains all i that were recorded within one unique timepoint segment. The family of all $\mathscr{t}$ is containded in $\mathbb{t}$. The division of timepoint segments is partially formulaic. The separations between timepoint-segments are defined as the middle stop between timepoints. If there are two middle stops, the lowest usage stop is selected as a boundary. The index $\dot{\mathscr{t}}$ is defined as a subset of $\mathscr{t}$ that includes all i recorded at each unique stop. For each unique stop in a timepoint segment, $\dot{t} \in \dot{\mathscr{t}} = \{\dot{t}_1, \dots, \dot{t}_n\}$ is the set of ordered events. While most data within a timepoint segment is aggregated for the entire segment, headways calculations focus the first and last stops only (i.e. $\dot{t}_1$ and $\dot{t}_n$).

$$\dot{t} \in \dot{\mathscr{t}} = \{\dot{t}_1, \dots, \dot{t}_n\} \in \{\dot{\mathscr{t}}_1, \dots, \dot{\mathscr{t}}_n\} : (\forall \dot{\mathscr{t}}, \dot{\mathscr{t}} \subset \mathscr{t}) \tag{4.4.1}$$

4.4.1. Summations

A left-superscript (Σ) is added to variables to indicate that ${}^{\Sigma}VAR$ is part of the aggregated data set. It does not directly indicate a sum; rather, Σ is used to help differentiate variables. Summations will continue to use the notation introduced in Appendix Section A.3. For example, ${}^{\Sigma}\hat{O}NS_{t}$ and ${}^{\Sigma T}\hat{D}WL_{t}$ are the sum of all boardings and the sum of door open duration within one timepoint-segment, t, respectively. The distribution of $\{{}^{\Sigma T}\hat{D}WL_{t} : t \in \mathbb{t}\}$ and $\{{}^{\Sigma T}\hat{B}AY_{t} : t \in \mathbb{t}\}$ are shown in Figure 4-14. Both distributions are heavily skewed towards shorter durations, peak at values under one-minute, then decay with long tails.

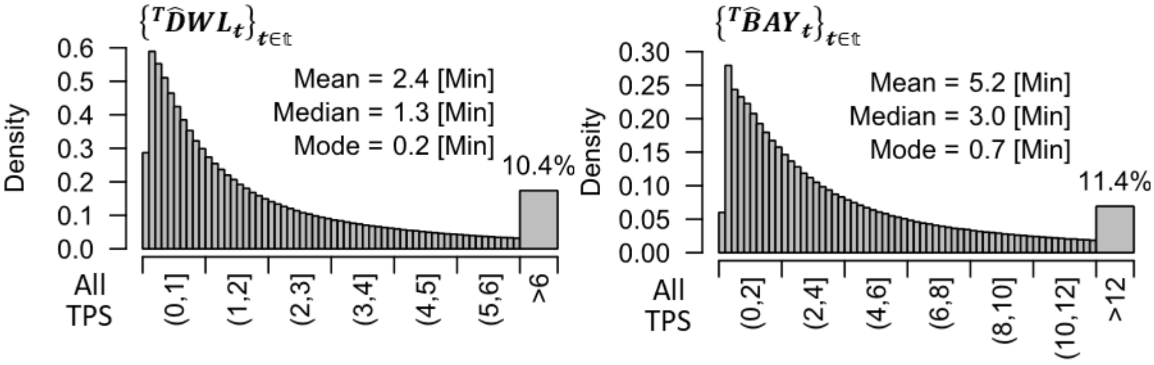

Figure 4-14 — Histogram for all timepoint-segments of $\{{}^{T}\hat{D}WL_{t} : t \in \mathbb{t}\}$ (Left) and $\{{}^{T}\hat{B}AY_{t} : t \in \mathbb{t}\}$ (Right).

Number of Vehicles

The aggregated variables have different distributions depending on the number vehicles per TPS (${}^{\Sigma}VEH_{t}$).

Definition 4-24 — ${}^{\Sigma}VEH_{t}$ is the *Number of Unique Vehicles* in one timepoint segment.

Table 4-9 shows the number, percent, and cumulative percent of observations for values of ${}^{\Sigma}VEH_t$. About a quarter of TPS have just one vehicle and about one-third have two. TPS with one or two vehicles represent about half of the aggregated data points, but just one-third of vehicles, one-quarter of service stops (${}^{E}SVC$), and one-fifth of passenger movements.

Table 4-9 — Number of timepoint segments given ${}^{\Sigma}VEH_t = \{1,2,\dots,12,>12\}$.

${}^{\Sigma}VEH_t$	Observations				${}^{\Sigma}VEH_t$	Observations		
	Number	*Percent*	*Cum. %*			*Number*	*Percent*	*Cum. %*
1	1,037,796	22.9%	22.9%		**8**	26,259	0.6%	99.1%
2	1,400,160	30.9%	53.9%		**9**	15,947	0.4%	99.4%
3	696,440	15.4%	69.3%		**10**	11,491	0.3%	99.7%
4	682,503	15.1%	84.3%		**11**	8,658	0.2%	99.9%
5	460,039	10.2%	94.5%		**12**	4,096	0.1%	100.0%
6	124,158	2.7%	97.2%		**>12**	2,135	0.0%	100.0%
7	56,119	1.2%	98.5%		**Total**	**4,525,801**	**100%**	**100%**

Figure 4-15 shows the average hourly number of TPS by number of vehicles per segment. The times-of-the day, where ${}^{\Sigma}VEH_t \geq 6$, are concentrated during the AM and PM peak periods corresponding to the high demand.

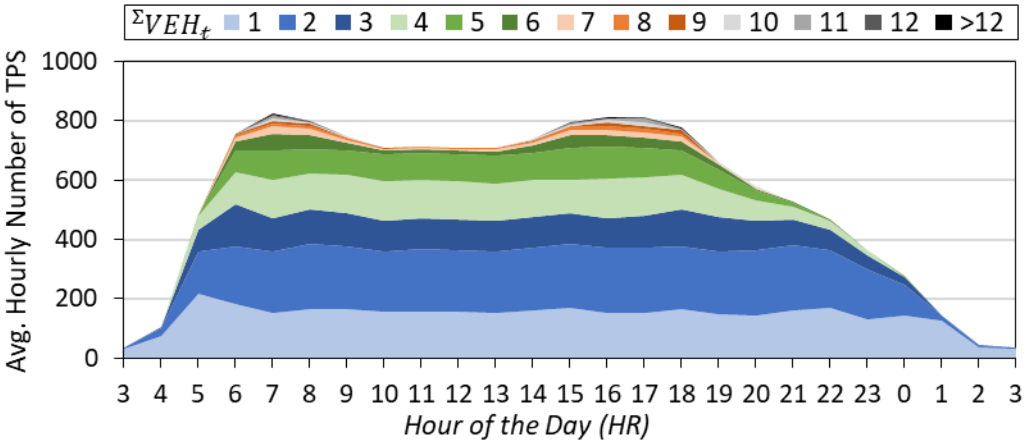

Figure 4-15 — Average hourly number of timepoint segments (TPS) by number of vehicles (${}^{\Sigma}VEH_t$) within segment.

Not all RTEs operate during all HRs, which changes the total number of segments throughout the day. More important to the number of segments is the activity within segments. Figure 4-16 shows the average hourly ${}^{\Sigma}\hat{O}NS_t + {}^{\Sigma}\hat{O}FFS_t$ for different number of ${}^{\Sigma}VEH_t$ per TPS.

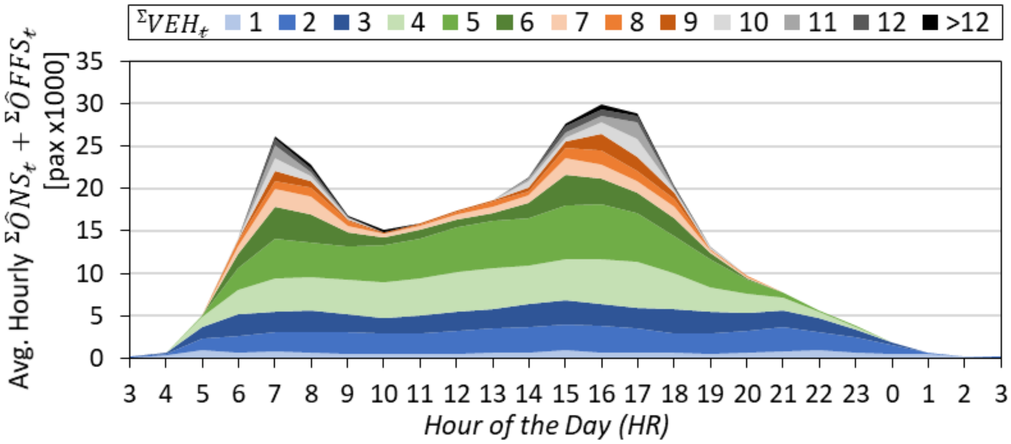

Figure 4-16 — Average hourly ${}^{\Sigma}\hat{O}NS_t + {}^{\Sigma}\hat{O}FFS_t$ for timepoint segments (TPS) by number of vehicles $({}^{\Sigma}VEH_t)$ within TPS.

Like is shown in Figure 4-11, passenger movements are heavily concentrated during those peak hours. During the AM and PM peak periods, TPS with ${}^{\Sigma}VEH_t \geq 6$ per TPS account for 35% of passenger movements while representing just 10% of TPS during those HRs. In contrast, TPS with one vehicle operate throughout the day, account for 23% of segments, but represent less than 5% of all passenger movements and less than 3% during the AM and PM peaks.

Normalized Summations

In addition to the summations, the sum of a variable, $({}^{\Sigma}VAR_t)$, divided by: one, the number of scheduled stops (i.e. ${}^{\Sigma E}SKD_t$); two, the number serviced stops (i.e. ${}^{\Sigma E}SVC_t$); or three, the number of vehicles (i.e. ${}^{\Sigma}VEH_t$). These averages will be denoted by ${}_{\mu(skd)}^{\Sigma}VAR_t$ ${}_{\mu(svc)}^{\Sigma}VAR_t$, and ${}_{\mu(veh)}^{\Sigma}VAR_t$, respectively, where μ is considered an average function. Table 4-10 shows means and variances for TPS service durations (i.e. ${}_{\mu(svc)}^{\Sigma T}\widehat{D}WL_t$ and ${}_{\mu(svc)}^{\Sigma T}\widehat{B}AY_t$) and passenger movements (i.e. ${}_{\mu(svc)}^{\Sigma}\hat{O}NS_t$ and ${}_{\mu(svc)}^{\Sigma}\hat{O}FFS_t$) by the number of vehicles per TPS (i.e. ${}^{\Sigma}VEH_t$).

Table 4-10 — Means and variances for service average ($\mu(svc)$) of stop durations $\left\{{}_{\mu(svc)}^{\Sigma T}\widehat{D}WL_t : \phi\right\}$ and $\left\{{}_{\mu(svc)}^{\Sigma T}\widehat{B}AY_t : \phi\right\}$, and passenger movements $\left\{{}_{\mu(svc)}^{\Sigma}\hat{O}NS_t : \phi\right\}$ and $\left\{{}_{\mu(svc)}^{\Sigma}\hat{O}FFS_t : \phi\right\}$, where ϕ is dependent on ${}^{\Sigma}VEH_t$.

$\phi := (t \in \mathbb{t}) \wedge$	$\left\{{}_{\mu(svc)}^{\Sigma T}\widehat{D}WL_t : \phi\right\}$		$\left\{{}_{\mu(svc)}^{\Sigma T}\widehat{B}AY_t : \phi\right\}$		$\left\{{}_{\mu(svc)}^{\Sigma}\hat{O}NS_t : \phi\right\}$		$\left\{{}_{\mu(svc)}^{\Sigma}\hat{O}FFS_t : \phi\right\}$	
$({}^{\Sigma}VEH_t < \cdots)$	***Mean***	***Var***	***Mean***	***Var***	***Mean***	***Var***	***Mean***	***Var***
1	15.0	822.3	31.5	1780.8	1.026	1.858	0.973	1.605
2	14.7	530.8	31.8	1426.7	1.080	1.327	1.066	1.243
3	14.7	403.4	32.4	1178.7	1.122	1.226	1.140	1.097
4	15.2	242.0	33.2	831.8	1.183	0.901	1.155	0.868
5	16.2	214.5	34.8	848.0	1.246	0.721	1.264	0.781
6	15.3	106.9	33.8	639.1	1.289	0.794	1.387	0.997
7	15.1	77.7	32.9	335.4	1.281	0.809	1.445	1.380
8	14.8	43.0	32.7	201.8	1.329	1.297	1.565	1.557
9	15.0	46.1	33.3	197.8	1.521	1.947	1.636	2.397
10	15.8	48.4	34.7	184.3	1.622	2.026	1.730	3.025
11	14.9	35.6	33.9	161.4	1.474	1.435	1.657	1.756
12	15.6	29.7	35.0	149.7	1.472	0.712	1.701	1.812
>12	15.2	25.4	34.1	129.3	1.386	0.606	1.586	1.189

As averages per service event (i.e. $\mu(svc)$), the means of $\left\{{}_{\mu(svc)}^{\Sigma T}\widehat{D}WL_t : \phi\right\}$ and $\left\{{}_{\mu(svc)}^{\Sigma T}\widehat{B}AY_t : \phi\right\}$ do not show large fluctuations for different ${}^{\Sigma}VEH_t$. However, their variances tend to decrease as ${}^{\Sigma}VEH_t$ increases. This trend is likely the result of taking averages over an increasing number of vehicles. The same trend is not observed for passenger movements. Variances decrease initially, then increase; the maximum variances are observed when ${}^{\Sigma}VEH_t$ is equal to ten.

Service Statistics

The distribution of services is dependent on the number of the vehicles in each timepoint-segment, the hour-of-the-day, or the specific location. Figure 4-17 is the first of a series of plots that uses both violin and box-and-whisker plots. The violin portion shows the density function of 99% of data points. A box-plot is plotted on top of the violin, giving the interquartile range (IQR) (i.e. the 25th to 75th percentile) the median, and whiskers showing either: 1.5 times the IQR or the maximum/minimum value. Due to the smoothing of the density function, violin plots can sometimes extend beyond the maximum or minimum value.

Figure 4-17 and Figure 4-18 shows the average scheduled stops per bus, dependent on the number of the vehicles (${}^{\Sigma}VEH_t$) in each TPS and the time-of-day, respectively. In both figures, an important takeaway is that the distribution of average scheduled stops per vehicle is reasonably consistent. The ranges tend to decrease as the number of vehicles increases and the distribution tends to be wider during the early morning and late evening.

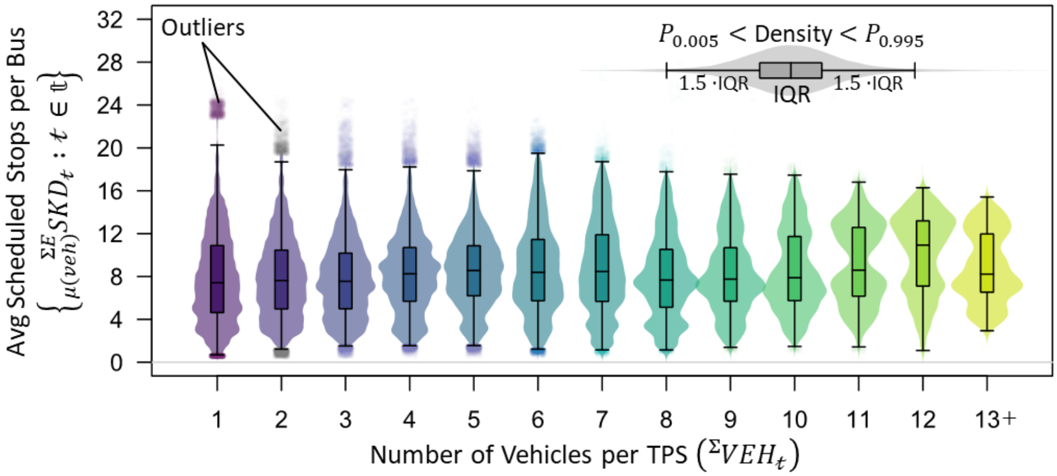

Figure 4-17 — Violin and box-plots for all TPS. Average scheduled stops per vehicle, $\left\{ {}^{\Sigma E}_{\mu(veh)}SKD_t : t \in \mathbb{t} \right\}$, given number of vehicles (${}^{\Sigma}VEH_t$).

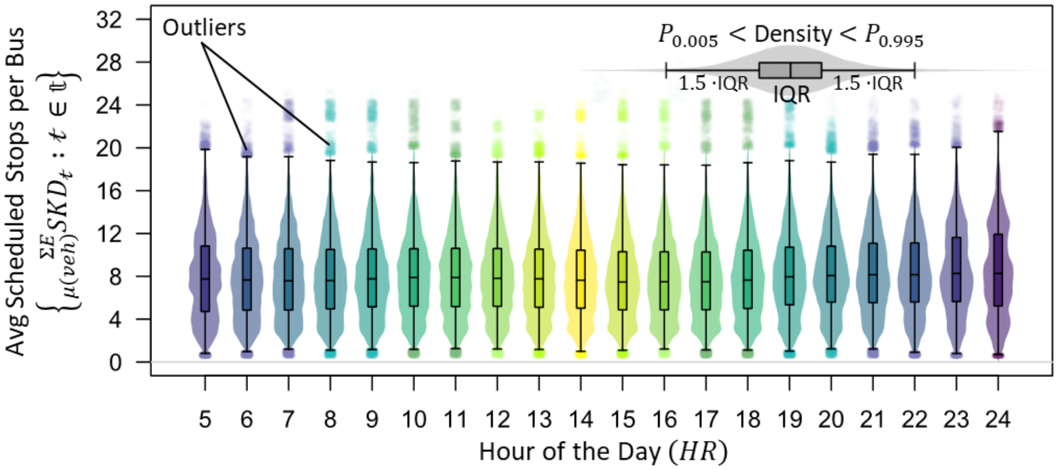

Figure 4-18 — Violin and box-plots for all timepoint segments. Average scheduled stops per vehicle, $\left\{ {}_{\mu(veh)}^{\Sigma E}SKD_t : t \in \mathfrak{t} \right\}$, given hour-of-the-day (HR_t).

The tends observed in these graphics are potentially a reflection of effective planning and consistent schedules. In Figure 4-19 and Figure 4-20, which show the percent of stops serviced, consistency is not the main trend. By the number of vehicles, a higher proportion of scheduled stops tend to be serviced as the number of vehicles increases.

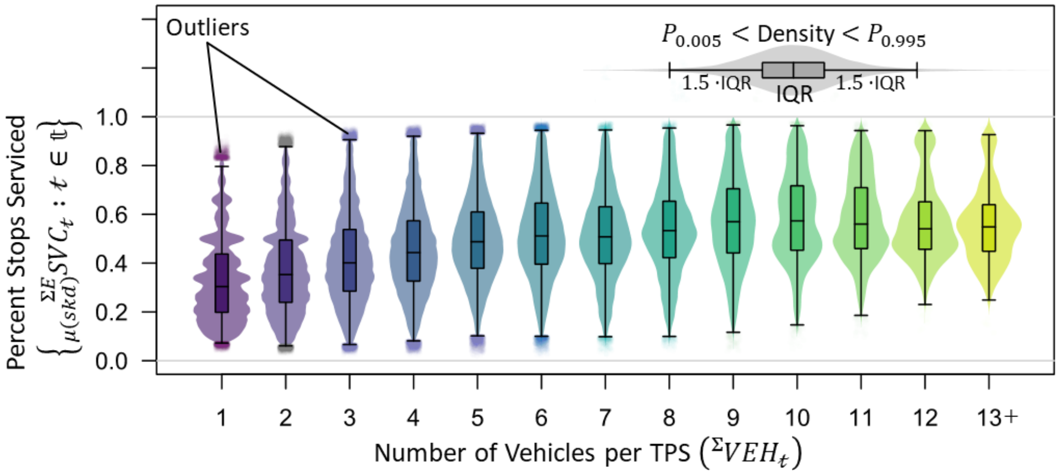

Figure 4-19 — Violin and box-plots for all timepoint segments. Percent of stops serviced by TPS, $\left\{ {}_{\mu(skd)}^{\Sigma E}SVC_t : t \in \mathfrak{t} \right\}$, given number of vehicles $\left({}^{\Sigma}VEH_t \right)$.

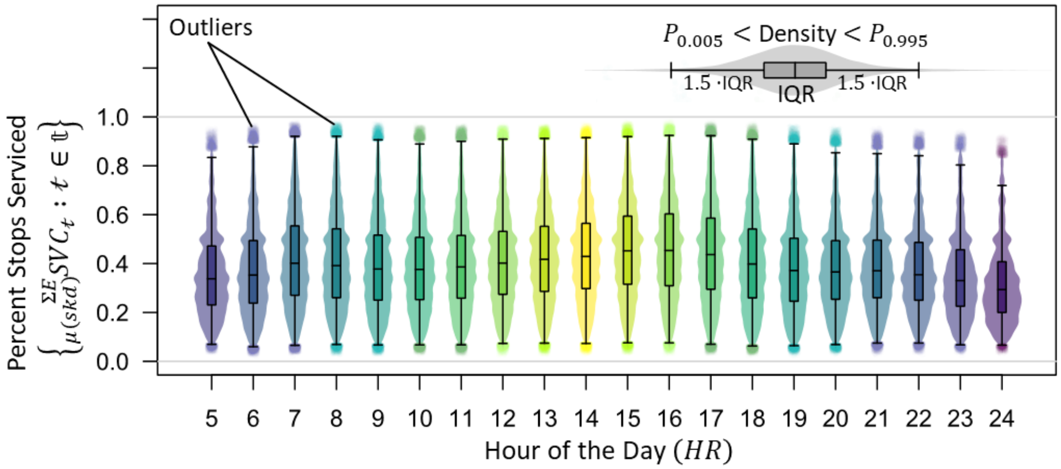

Figure 4-20 — Violin and box-plots for all timepoint segments. Percent of stops serviced by TPS, $\left\{{}^{\Sigma E}_{\mu(skd)}SVC_t : t \in \mathbb{t}\right\}$, given hour-of-the-day (HR_t).

For TPS with six or more vehicles, 55% of segments have service percentages above 50%. Less than 5% of these TPS serve less than 20% of stops on their schedules. For TPS with less than six vehicles, just 30% of segments have service percentages above 50%. By time-of-day (Figure 4-20), trends towards higher service percentages seem to follow a typical commuter pattern. The highest IQRs are observed during the AM and PM peaks.

Location Aggregation

Figure 4-14, which shows distribution of service durations, considers all TPSs; however, the distributions of service times are location dependent at the stop level, which also carries forward to the aggregated level. Location variables are aggregated in three parts for each type: first, the number of locations on the service schedule (i.e. ${}^{\Sigma L}_{skd}VAR_t$); second, the number of service events (i.e. ${}^{\Sigma L}VAR_t$); and third, the number of ${}^{E}THRU$ $\left({}^{\Sigma L}_{thru}VAR_t = {}^{\Sigma L}_{skd}VAR_t - {}^{\Sigma L}VAR_t\right)$. These categories will be shown to have different overall effects on variable coefficients.

At the aggregated level, transit centers and the transit mall also include other locations types within the TPS. Figure 4-21 and Figure 4-22 show the distributions of ${}^{T}\widehat{D}WL_{t}$ and ${}^{T}\widehat{B}AY_{t}$ for TPS where ${}^{L}TC_{t} \geq 1$ and ${}^{L}MALL_{t} \geq 1$, respectively.

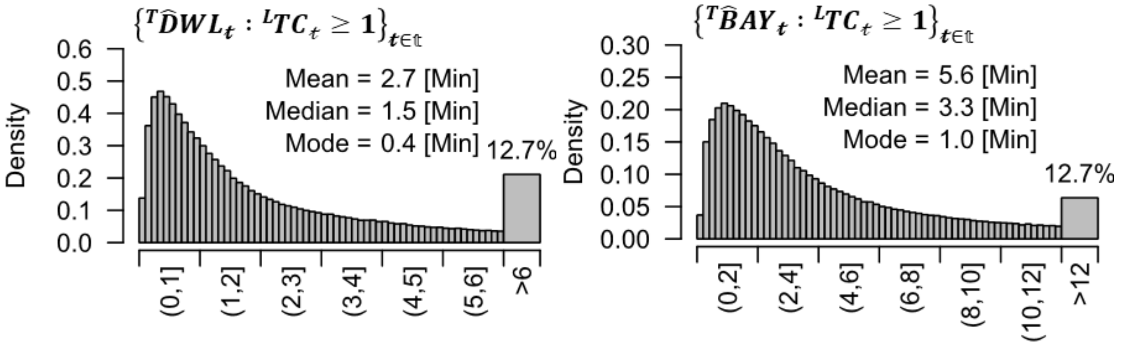

Figure 4-21 — Histogram of $\{{}^{T}\widehat{D}WL_{t} : ({}^{L}TC_{t} \geq 1) \wedge (t \in \mathbb{t})\}$ (Left) and $\{{}^{T}\widehat{B}AY_{t} : ({}^{L}TC_{t} \geq 1) \wedge (t \in \mathbb{t})\}$ (Right).

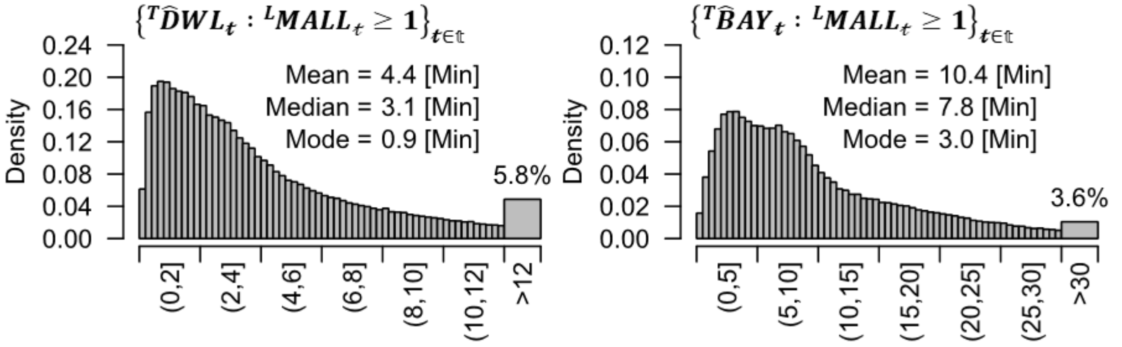

Figure 4-22 — Histogram of $\{{}^{T}\widehat{D}WL_{t} : ({}^{L}MALL_{t} \geq 1) \wedge (t \in \mathbb{t})\}$ (Left) and $\{{}^{T}\widehat{B}AY_{t} : ({}^{L}MALL_{t} \geq 1) \wedge (t \in \mathbb{t})\}$ (Right).

TPSs with a transit center are more similar to the system as a whole than TPSs with stops on the transit mall, but both experience longer aggregated service durations. For bus-bay stop durations on the mall, the mean, median, and mode are double the metrics on the rest of the network. Given the unique features and service requirements of vehicles (e.g. the requirement to stop at all stops) on the downtown transit mall, such differences are not unexpected. The aggregated distribution for the mall also reduces the bimodal distribution seen in Figure 4-10.

4.4.2. Headways

While headways may be calculated any consecutive vehicles from the same route, for the data aggregation, only vehicles within one TPS are included. As such, two vehicles are needed for one headway and three are needed for any additional statistics. For this aggregation, ${}^{S}H_i$, ${}^{A}H_i$, and ${}^{D}H_i$ are used using the previously defined formulation. The first vehicle in each $ŧ$ will technically have a headway, but it corresponds to a vehicle outside the segment and is therefore not part of calculations.

Headway performance metrics are calculated at the first and last stop of each segment. The first stop will use the arrival headways (i.e. ${}^{A}H_i$) and the last stop will use the departure headways (i.e. ${}^{D}H_i$) to differentiate calculations and variable names. Functionally, there are limited differences between the two headways.

Definition 4-25 — ${}_{avg}^{A}H_{ŧ}$ and ${}_{avg}^{D}H_{ŧ}$ are the *Mean* (avg) *Headway* between vehicles *arriving* at the first stop or *departing* the last stop of a timepoint-segment.

(4.4.2)

$$\left\{\begin{matrix} \forall\, {}_{avg}^{A}H_{ŧ} \in \left\{ {}_{avg}^{A}H_{ŧ} \right\}_{ŧ \in \mathbb{t}}, \Big({}_{avg}^{A}H_{ŧ} = {}_{avg}^{A}H_{\dot{ŧ}_1 \in ŧ} \\ \qquad = \left\{\begin{matrix} \frac{1}{\|\dot{ŧ}_1\|-1} \sum_{(\forall \dot{t} \neq \dot{t}_1) \in \dot{ŧ}_1} \left[{}^{A}H_{i \leftarrow \dot{t}} \right], & \text{if } \|\dot{ŧ}_1\| \geq 3 \\ \emptyset, & \text{otherwise} \end{matrix}\right\} \Big) \\ \forall\, {}_{avg}^{D}H_{ŧ} \in \left\{ {}_{avg}^{D}H_{ŧ} \right\}_{ŧ \in \mathbb{t}}, \Big({}_{avg}^{D}H_{ŧ} = {}_{avg}^{D}H_{\dot{ŧ}_n \in ŧ} \\ \qquad = \left\{\begin{matrix} \frac{1}{\|\dot{ŧ}_n\|-1} \sum_{(\forall \dot{t} \neq \dot{t}_n) \in \dot{ŧ}_n} \left[{}^{A}H_{i \leftarrow \dot{t}} \right], & \text{if } \|\dot{ŧ}_n\| \geq 3 \\ \emptyset, & \text{otherwise} \end{matrix}\right\} \Big) \end{matrix}\right\},$$

Where $\|\dot{ŧ}_1\|$ and $\|\dot{ŧ}_n\|$ are the number of elements in $\dot{ŧ}_1$ and $\dot{ŧ}_n$, respectivley; *and*, $i \leftarrow \dot{t}$ is a function mapping i from the index $\dot{t}$.

For the average headways, the formulation of the model requires at least three vehicles for an average headway. For all subsequent calculations, it can be assumed that non-null values were calculated from three or more vehicles. Average headways provide useful information, but are not useful to compare performance across segments with different scheduled headways. Useful metrics for such comparisons are the mean absolute deviation (defined in Definition 4-26), which helps to quantify consistency; and, a headway deviation index (defined in Definition 4-27), which is a unitless ratio created by normalizing the mean absolute deviation by the mean headway.

Definition 4-26 — ${}_{mad}^{A}H_{t}$ and ${}_{mad}^{D}H_{t}$ [sec] are the *Mean Absolute Deviation* (mad) for arrivals at the first stop and departures at the last stop of a TPS. ${}_{mad}^{\{A,D\}}H_{t}$ are defined as the absolute difference between headways and mean headway.

$$(4.4.3)\quad \left\{\begin{aligned} &\forall\, {}_{mad}^{A}H_{t} \in \{{}_{mad}^{A}H_{t}\}_{t\in\mathbb{t}}, \Big({}_{mad}^{A}H_{t} = {}_{mad}^{A}H_{\dot{t}_1\in t} \\ &\qquad = \frac{1}{\|\dot{t}_1\|-1}\sum\nolimits_{(\forall \dot{t}\neq \dot{t}_1)\in \dot{t}_1}\Big[\big|{}^{A}H_{i\leftarrow \dot{t}} - {}_{avg}^{A}H_{t}\big|\Big]\Big) \\ &\forall\, {}_{mad}^{D}H_{t} \in \{{}_{mad}^{D}H_{t}\}_{t\in\mathbb{t}}, \Big({}_{mad}^{D}H_{t} = {}_{mad}^{D}H_{\dot{t}_n\in t} \\ &\qquad = \frac{1}{\|\dot{t}_n\|-1}\sum\nolimits_{(\forall \dot{t}\neq \dot{t}_n)\in \dot{t}_n}\Big[\big|{}^{A}H_{i\leftarrow \dot{t}} - {}_{avg}^{D}H_{t}\big|\Big]\Big) \end{aligned}\right\}$$

Definition 4-27 — ${}_{idx}^{A}H_{t}$ and ${}_{idx}^{D}H_{t}$ [unitless ratio] are the *Headway Deviation Indexes* (idx) for arrivals at the first stop and departures at the last stop of a TPS.

$$(4.4.4)\quad \left\{\begin{aligned} &\forall\, {}_{idx}^{A}H_{t} \in \{{}_{idx}^{A}H_{t}\}_{t\in\mathbb{t}}, \Big({}_{idx}^{A}H_{t} = \frac{{}_{mad}^{A}H_{t}}{{}_{avg}^{A}H_{t}} = \frac{{}_{mad}^{A}H_{\dot{t}_1\in t}}{{}_{avg}^{A}H_{\dot{t}_1\in t}}\Big) \\ &\forall\, {}_{idx}^{D}H_{t} \in \{{}_{idx}^{D}H_{t}\}_{t\in\mathbb{t}}, \Big({}_{idx}^{D}H_{t} = \frac{{}_{mad}^{D}H_{t}}{{}_{avg}^{D}H_{t}} = \frac{{}_{mad}^{D}H_{\dot{t}_n\in t}}{{}_{avg}^{D}H_{\dot{t}_n\in t}}\Big) \end{aligned}\right\}$$

In addition, it is useful to understand how these deviations and deviation indexes relate to the scheduled headways. For these comparisons, the mean, mean absolute deviation, and the headway deviation index need to be calculated for scheduled headways of vehicles at the first (SA) and last (SD) stops of each TPS.

Definition 4-28 — ${}_{avg}^{SA}H_{t}$ and ${}_{avg}^{SD}H_{t}$ [sec] are the *Mean Headway* for *scheduled* arrivals at the first stop and *scheduled* departures at the last stop of a TPS.

(4.4.5) $$\left\{\begin{array}{l} \forall\, {}_{avg}^{SA}H_{t} \in \left\{{}_{avg}^{SA}H_{t}\right\}_{t\in\mathbb{t}}, \left({}_{avg}^{SA}H_{t} = {}_{avg}^{SA}H_{\dot{t}_1\in t} = \frac{1}{\|\dot{t}_1\|-1}\sum_{(\forall\dot{t}\neq\dot{t}_1)\in\dot{t}_1}\left[{}^{S}H_{i\mapsfrom\dot{t}}\right]\right) \\ \forall\, {}_{avg}^{SD}H_{t} \in \left\{{}_{avg}^{SD}H_{t}\right\}_{t\in\mathbb{t}}, \left({}_{avg}^{SD}H_{t} = {}_{avg}^{SD}H_{\dot{t}_n\in t} = \frac{1}{\|\dot{t}_n\|-1}\sum_{(\forall\dot{t}\neq\dot{t}_n)\in\dot{t}_n}\left[{}^{S}H_{i\mapsfrom\dot{t}}\right]\right) \end{array}\right\}$$

Definition 4-29 — ${}_{mad}^{SA}H_{t}$ and ${}_{mad}^{SD}H_{t}$ [sec] are the *Mean Absolute Deviation* (mad) for *scheduled* arrivals at the first stop and *scheduled* departures at the last stop.

(4.4.6)

$$\left\{\begin{array}{l} \forall\, {}_{mad}^{SA}H_{t} \in \{{}_{mad}^{SA}H_{t}\}_{t\in\mathbb{t}}, \left({}_{mad}^{SA}H_{t} = {}_{mad}^{SA}H_{\dot{t}_1\in t}\right. \\ \qquad\qquad\qquad\qquad\qquad \left. = \frac{1}{\|\dot{t}_1\|-1}\sum_{(\forall\dot{t}\neq\dot{t}_1)\in\dot{t}_1}\left[\left|{}^{S}H_{i\mapsfrom\dot{t}} - {}_{avg}^{S}H_{t}\right|\right]\right) \\ \forall\, {}_{mad}^{SD}H_{t} \in \{{}_{mad}^{SD}H_{t}\}_{t\in\mathbb{t}}, \left({}_{mad}^{SD}H_{t} = {}_{mad}^{SD}H_{\dot{t}_n\in t}\right. \\ \qquad\qquad\qquad\qquad\qquad \left. = \frac{1}{\|\dot{t}_n\|-1}\sum_{(\forall\dot{t}\neq\dot{t}_n)\in\dot{t}_n}\left[\left|{}^{S}H_{i\mapsfrom\dot{t}} - {}_{avg}^{S}H_{t}\right|\right]\right) \end{array}\right\}$$

Definition 4-30 — ${}_{idx}^{SA}H_{t}$ and ${}_{idx}^{SD}H_{t}$ [unitless ratio] *Headway Deviation Indexes* (idx) for *scheduled* arrivals at the first stop and *scheduled* departures at the last stop of a TPS.

(4.4.7) $$\left\{\begin{array}{l} \forall\, {}_{idx}^{SA}H_{t} \in \left\{{}_{idx}^{SA}H_{t}\right\}_{t\in\mathbb{t}}, \left({}_{idx}^{SA}H_{t} = \frac{{}_{mad}^{SA}H_{t}}{{}_{avg}^{SA}H_{t}} = \frac{{}_{mad}^{SA}H_{\dot{t}_1\in t}}{{}_{avg}^{SA}H_{\dot{t}_1\in t}}\right) \\ \forall\, {}_{idx}^{SD}H_{t} \in \left\{{}_{idx}^{SD}H_{t}\right\}_{t\in\mathbb{t}}, \left({}_{idx}^{SD}H_{t} = \frac{{}_{mad}^{SD}H_{t}}{{}_{avg}^{D}H_{t}} = \frac{{}_{mad}^{SD}H_{\dot{t}_n\in t}}{{}_{avg}^{SD}H_{\dot{t}_n\in t}}\right) \end{array}\right\}$$

Finally, the indexes for observed and scheduled headways may be combined to produce adjusted deviation indexes (*adj*). These indexes provide the means to compare overlapping segments in terms of their own scheduled headways.

Definition 4-31 — ${}_{adj}^{A}H_{\mathcal{t}}$ and ${}_{adj}^{D}H_{\mathcal{t}}$ [unitless ratio] are the *Adjusted Deviation Indexes* for Arrivals at the first stop and Departures at the last stop, respectively.

(4.4.8)
$$\begin{Bmatrix} \forall\, {}_{adj}^{A}H_{\mathcal{t}} \in \left\{{}_{adj}^{A}H_{\mathcal{t}}\right\}_{\mathcal{t}\in\mathbb{t}}, \left({}_{adj}^{A}H_{\mathcal{t}} = \frac{{}_{idx}^{A}H_{\mathcal{t}} - {}_{idx}^{SA}H_{\mathcal{t}}}{1-{}_{idx}^{SA}H_{\mathcal{t}}} = \frac{{}_{idx}^{A}H_{\dot{\mathcal{t}}_1\in\mathcal{t}} - {}_{idx}^{SA}H_{\dot{\mathcal{t}}_1\in\mathcal{t}}}{1-{}_{idx}^{SA}H_{\dot{\mathcal{t}}_1\in\mathcal{t}}}\right) \\ \forall\, {}_{adj}^{D}H_{\mathcal{t}} \in \left\{{}_{adj}^{D}H_{\mathcal{t}}\right\}_{\mathcal{t}\in\mathbb{t}}, \left({}_{adj}^{D}H_{\mathcal{t}} = \frac{{}_{idx}^{D}H_{\mathcal{t}} - {}_{idx}^{SD}H_{\mathcal{t}}}{1-{}_{idx}^{SD}H_{\mathcal{t}}} = \frac{{}_{idx}^{D}H_{\dot{\mathcal{t}}_1\in\mathcal{t}} - {}_{idx}^{SD}H_{\dot{\mathcal{t}}_1\in\mathcal{t}}}{1-{}_{idx}^{SD}H_{\dot{\mathcal{t}}_1\in\mathcal{t}}}\right) \end{Bmatrix}$$

4.4.3. Congestion

The aggregated data set provides a means to calculate the effects of congestion. The values calculated the route level will serve as a baseline for comparisons. The main differences in the calculations are: first, timepoint-segments are considered; second, the run-time estimate, from equation (4.3.11), are now calculated directly from the average moving time (i.e.) and average disturbance time (i.e.) in different time-periods; and third, excess wait time is based on the average between the mean absolute deviation for arrival and departures at the first and last stop of each timepoint-segment. The conversions factors, from Section 4.4.3, that relate run-time to buffer time and passenger ride-and-recovery time will still be used; as will the assumptions of costs.

A key difference is that the periods, $\mathcal{p}$, which previously represented $\mathcal{r}_w \cap \dot{\mathcal{h}}$, now need to represent $\mathcal{t} \cap \dot{\mathcal{h}}$. As such, the index $\mathcal{t}_p$ is defined by the intersection of $\mathcal{t}$ and $\dot{\mathcal{h}}$, such that the complete set of $\mathcal{t}_p$ is contained in $\mathbb{t}_p$. The key different between $\mathcal{t}_p$ and $\mathcal{t}$ is that off-peak hours have been grouped together.

(4.4.9)
$$\forall \mathcal{t}_p \in \mathbb{p}, \left(\mathcal{t}_p = \bigcap \mathcal{t}_p' : \mathcal{t}_p' \in \mathbb{t}_p'\right),$$

Where: $\mathbb{t}_p' = \mathbb{t} \times \dot{\mathbb{h}} = \left\{(\mathcal{t}, \dot{\mathcal{h}}) : \left((\mathcal{t} \in \mathbb{t}) \wedge (\dot{\mathcal{h}} \in \dot{\mathbb{h}})\right)\right\}$.

Agencies

For agencies, both the running time and recovery time are considered as one value, due to TriMet's requirement that recovery time must be five-minutes per one-hour of service time.

Definition 4-32 — ${}^{\Sigma CT}_{avg}\tilde{R}\&R_{t_p}$ [sec] is the average time of congestion per trip in period $\mathscr{p}$ resulting from an increase in running time.

(4.4.10)

$$\forall {}^{\Sigma CT}_{avg}\tilde{R}\&R_{t_p} \in \left\{{}^{\Sigma CT}_{avg}\tilde{R}\&R_{t_p}\right\}_{t_p \in \mathbb{t}_p},$$

$$\left({}^{\Sigma CT}_{avg}\tilde{R}\&R_{t_p} = \left(\frac{13}{12}\right) \cdot \left({}^{\Sigma C'T}_{avg}\tilde{R}UN_{t_p} - {}^{\Sigma C'T}_{avg}\tilde{R}UN_{(t_p = t_{p,0})}\right) : \mathscr{p} \mapsfrom \left(t \cap \dot{\hbar}\right)\right),$$

Such that: t_p and $t_{p,0}$ always refer to the same t; *where:*

$$\forall {}^{\Sigma C'T}_{avg}\tilde{R}UN_{t_p} \in \left\{{}^{\Sigma C'T}_{avg}\tilde{R}UN_{t_p}\right\}_{t_p \in \mathbb{t}_p}, \left({}^{\Sigma C'T}_{avg}\tilde{R}UN_{t_p} = {}_{\mu(veh)}\left({}^{\Sigma T}\widetilde{M}OVE + {}^{\Sigma T}\widetilde{D}STB\right)_{t_p}\right),$$

And, given that: ${}_{\mu(veh)}\left({}^{\Sigma T}\widetilde{M}OVE + {}^{\Sigma T}\widetilde{D}STB\right)_{t_p}$ is the average of the sum for all trips.

By the formulation in equation (4.4.9), there are never increases for off-peak periods, the same as for event level data. Multiplying ${}^{\Sigma CT}_{avg}\tilde{R}\&R_{t_p}$ by the agency operation costs per unit time, results in the average monetary impact of running time from congestion per trip, assuming matching units after appropriate conversions. The same value of \$139.20 per hour will be used. (TriMet, 2019).

(4.4.11) $\forall {}^{\Sigma C\$}_{avg}R\&R_{t_p} \in \left\{{}^{\Sigma C\$}_{avg}\tilde{R}\&R_{t_p}\right\}_{t_p \in \mathbb{t}_p}, \left(k \cdot \$139.20 \left[\frac{\text{USD}}{\text{hour}}\right] \cdot {}^{\Sigma CT}_{avg}\tilde{R}\&R_{t_p}[\text{sec}]\right),$

Where: k is a constant used to convert units.

Passengers

For passengers, the in-vehicle time be based ono the increase in running and recovery time, the main difference is that estimated passenger load will be used directly to calculate costs instead of the 40% of boardings estimate. The 75% estimate for recovery time and value per passenger will still be used.

$$\text{(4.4.12)} \quad \forall {}^{\Sigma C\$}_{avg}R\&B_{t_p} \in \left\{ {}^{\Sigma C\$}_{avg}\tilde{R}\&B_{t_p} \right\}_{t_p \in \mathbb{t}_p}, \left({}^{\Sigma C\$}_{avg}\tilde{R}\&B_{t_p} = \left(\frac{12}{13} + \frac{1}{13} \cdot \frac{3}{4}\right) \cdot {}^{\Sigma CT}_{avg}\tilde{R}\&R_{t_p} \right.$$
$$\left. = \left(\frac{51}{52}\right) \cdot {}^{\Sigma CT}_{avg}\tilde{R}\&R_{t_p} \right)$$

$$\text{(4.4.13)} \quad \forall {}^{\Sigma C\$}_{avg}R\&B_{t_p} \in \left\{ {}^{\Sigma C\$}_{avg}R\&B_{t_p} \right\}_{t_p \in \mathbb{t}_p},$$
$$\left({}^{\Sigma C\$}_{avg}R\&B_{t_p} = k \cdot \$12 \left[\frac{\text{USD}}{\text{hour}}\right] \cdot \left(\frac{12}{13} + \frac{1}{13} \cdot \frac{3}{4} \cdot \frac{3}{4}\right) \cdot {}^{\Sigma CT}_{avg}\tilde{R}\&R_{t_p}[\text{sec}] \right.$$
$$\left. = k \cdot \$12 \left[\frac{\text{USD}}{\text{hour}}\right] \cdot \left(\frac{201}{208}\right) \cdot {}^{\Sigma CT}_{avg}\tilde{R}\&R_{t_p}[\text{sec}] \right),$$

Where: k is a constant used to convert units.

The excess waiting time is calculated considering the average of the mean absolute deviation from the first and last stop of each timepoint segment. This formulation in equation (4.4.14) is intended to capture increased headway variability between each period and the off-peak time-period. The value of passenger time remains set at $18 per hour while wait. The average cost per passenger, resulting from waiting time in t_p is defined by equation (4.4.15).

(4.4.14) $\forall {}^{\Sigma CT}_{avg}\widetilde{E}XW_{t_p} \in \left\{ {}^{\Sigma CT}_{avg}\widetilde{E}XW_{t_p} \right\}_{t_p \in \mathbb{t}_p},$

$$\left({}^{\Sigma CT}_{avg}\widetilde{E}XW_{t_p} = \max \left[\begin{matrix} 0, \\ \left({}^{\Sigma C'T}_{avg}\widetilde{E}XW_{t_p} - {}^{\Sigma C'T}_{avg}\widetilde{E}XW_{t_{p,0}} \right) \end{matrix} \right] \right),$$

Where: ${}^{\Sigma C'T}_{avg}\widetilde{E}XW_{t_p} = \frac{1}{2}\left({}^{A}_{mad}H_{t_p} + {}^{D}_{mad}H_{t_p} \right).$

(4.4.15)

$$\forall {}^{\Sigma CT}_{avg}EXW_{t_p} \in \left\{ {}^{\Sigma CT}_{avg}EXW_{t_p} \right\}_{p \in \mathbb{p}}, \left({}^{\Sigma CT}_{avg}EXW_{t_p} = k \cdot \$18 \left[\frac{\text{USD}}{\text{hour}} \right] \cdot {}^{\Sigma CT}_{avg}\widetilde{E}XW_{t_p} [\text{sec}] \right),$$

Where: k is again a constant used to convert units.

The total passenger costs, over a desired time-frame, is calculated by multiplying the number of boarding passengers in each t_p by relevant cost averages then summing over the target time-frame.

Travel Speeds

Average transit moving speeds within a timepoint segment generally decrease during peak periods and during midday, as compared to off-peak times. The highest speeds are seen in the early hours of the morning. Across all TPS for the network, the variance and spread of TPS speeds remains fairly consistent throughout a day. The maximum variance is 34 between 1:00 and 2:00 AM and the minimum is 28 between 6:00 and 7:00 AM. Confidence intervals for moving speeds within TPS throughout a day are shown in Figure 4-23.

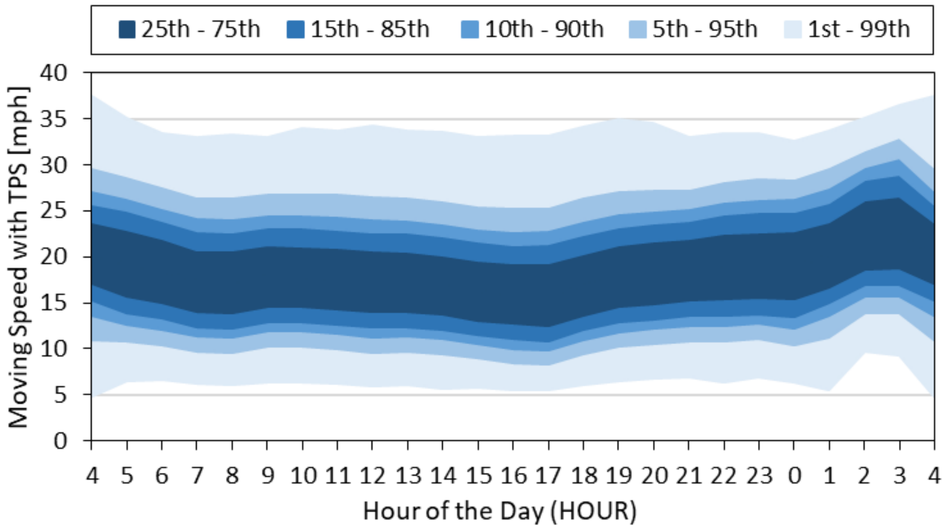

Figure 4-23 — Percentile windows for moving speed (mph) for all vehicles within timepoint segments (TPS).

4.5. Data Verification

A final part of a full data description is the verification that the cleaning process results in variables that represent the actual network. One useful tool is to estimate performance metrics from the data that are reported by TriMet for individual years (TriMet, 2019). The data set used for this analysis spans multiple years, as such interpolated values were estimated based on the proportion of the data set in each year (Table 4-11).

Table 4-11 — TriMet reported system performance metrics.

TriMet Ridership Report (Bus Only)	***2017***	***2018***	***Weighted***
Total Yearly Boarding Rides	57,820,520	56,737,466	56,971,478
Average Weekday Boarding Rides	186,800	183,800	184,449
Revenue Hours	1,529,532	1,552,044	1,547,180
Revenue Miles	20,923,103	21,354,739	21,261,477
Passenger Miles	214,823,255	203,687,503	206,093,566

The estimates created from the data are based on a total of 361 days. The values are therefore scaled by a factor of $\frac{365}{361}$ to get estimates. Table 4-12 includes estimates using cleaned (corrected) values and the original (uncorrected) values for the same set of ELD.

Table 4-12 — Performance metrics estimates using corrected data and original (i.e. uncorrected) data.

	Original		**Corrected**	
Performance Metric	*Estimate*	*% Error*	*Estimate*	*% Error*
Total Yearly Boarding Rides	56,118,684	-1.50%	57,465,222	0.87%
Average Weekday Boarding Rides	179,361	-2.76%	1,552,648	-0.29%
Revenue Hours	1,552,648	0.35%	*-NA-*	
Revenue Miles	21,160,004	-0.48%	*-NA-*	
Passenger Miles	179,761,978	-12.78%	197,179,979	-4.33%

For each of these metrics, which are those that can be calculated without additional sources, the cleaned data produces values that approximate the official reports. The exception is passenger miles, which is low, but still improved from the original, uncorrected data.

Another point of verification is stop or location specific. Outputs were created to verify performance at individual bus stops and groups of stops (e.g. entire routes and transit centers. Table 4-13 is an example of one of these outputs that is specific to transit centers. The “real” values were taken from TriMet reports for total usage including the MAX, Portland Streetcar, and other transit agencies. The data set used in this research is specific to buses and therefore other modes needed to be removed. For most transit centers, estimates for non-bus stop usage were available and could be removed, but not for all.

Table 4-13 — Percent error for transit center passenger usage estimates.

	Average Passengers ($\hat{O}NS + \hat{O}FFS$) Percent Error (95% Confidence Interval*)*	
Transit Center	***Weekdays***	***Weekends***
Barbur Blvd T	(-8.7% , 1.9%)	(-8.3% , 8.8%)
Beaverton TC	(-2.6% , 1.8%)	(-6.9% , -3.1%)
Clackamas Town Center TC	(-5.0% , 0.4%)	(-3.2% , 1.9%)
Gateway / NE 99th Ave TC	(-11.6% , -3.4%)	(-18.9% , -10.6%)
Gresham Central TC	(-9.3% , -2.8%)	(-6.9% , -0.6%)
Hillsboro Central/SE 3rd Ave TC	(-0.8% , 8.7%)	(-6.2% , 2.6%)
Hollywood / NE 42nd Ave TC	(-4.9% , 4.0%)	(-1.6% , 6.9%)
Lake Oswego TC	(-7.3% , 7.0%)	(-8.1% , 9.4%)
N Lombard TC	(-6.0% , 1.7%)	(-2.3% , 4.6%)
Oregon City TC	(-10.7% , -2.8%)	(-7.2% , 2.6%)
Parkrose / Sumner TC	(-1.7% , 6.6%)	(-3.8% , 3.5%)
Rose Quarter TC	(-7.1% , -0.9%)	(-5.4% , 1.4%)
Sunset TC	(-15.1% , -7.4%)	(-13.3% , -4.2%)
Tigard TC	(-10.0% , -2.9%)	(0.4% , 7.1%)
Washington Square TC	(-6.5% , 1.8%)	(-3.6% , 3.6%)
Willow Creek / SW 185th Ave TC	(-9.2% , -0.4%)	(-7.9% , 2.0%)

The data set used in this research is specific to buses and therefore other modes needed to be removed. For most transit centers, estimates for non-bus stop usage were available and could be removed, but not for all. At the locations where the 95% confidence of the percent error does not contain a zero, other usage often remained. For example, The Gateway Transit Center estimates of passenger movements include the Columbia Area transit and the Colombia George Express. The Sunset Transit Center includes The Point, The Wave, Forrest Heights Shuttle, and the PCC Shuttle. In these cases, usage was verified by examining each bus stop location individually.

4.6. Conclusion

Chapter 4 builds on the definitions and datasets introduces in Chapter 3. Where the previous chapter establishes a broadly applicable approach to data cleaning and may be used independently, Chapter 4 relies on the stochastic cleaning methodology from Chapter 3. More traditional methods of data cleaning, such as excluding outliers, would results in unacceptable underestimations for the timepoint-segment aggregation. Yet, with the data cleaning methodology from the previous chapter, the data aggregation becomes a potentially powerful tool for examining transit.

Chapter 4 does not model performance, but shows how even simple histograms of transit operations, at the TPS level, have the potential to show operational trends and smooth the high variability seen in event-level data. For example, the bi-modal distribution of stop times on the downtown transit mall, which is not present at the TPS level. While the source of the bi-modal distribution may be useful information to understand specific stops, the TPS aggregation helps show that overall performance is more normally distributed and difference between nearby locations may not be having notable negative effects on overall operations. Overall, Chapter 4 establishes the variables needed for more detailed analysis and provides the ability to examine transit performance at a mesoscopic and microscopic level to compare the results and evaluate the tradeoffs between analysis levels.

CHAPTER 5 — RESULTS: SERVICE DURATION MODELING

5.1. Introduction

Chapter 5 is the first of the two “Results” chapters in this dissertation and will focus on service duration modeling. The chapter first introduces the process used to create the final regression models in Section 5.2, then evaluates service durations with bus bays using the event level data and aggregated data in Sections 5.3 and 5.4, respectively. The aggregated data is further utilized to evaluate moving time and stopped time between bus stops in Section 5.5; then, total travel time in 5.6. That section is also used to compare the effectiveness of the aggregated models by comparing results back to the regression results using the event level data.

An additional focus of Chapter 5 is given to evaluating the tradeoffs between sample size and usable results. As the number of available data sets and the sizes of those datasets increase, it useful to evaluate the tradeoffs between quantity of the inputs and quality of the results. While including more data in an analysis is likely to reduce the variance of results, it also increases the computational burden. As such, it is important to define how much data is needed for consistent results and how that quantity may change depending on the type of data.

5.2. Service Duration Modeling

A critical step in assessing the performance of the aggregated prediction models is to provide a basis for comparison. Using previous publications to guide variable selection, linear and log-linear models were tested for all stops and on important subsets of those stops. The ELD contains more than 45.7 million entries for stop services. For each run of each model, all applicable data points had an equal chance of being included. Models with less than 4.5 million applicable points included all values; models with more than 4.5 million applicable points included a random subset of 4.5 to 4.8 million values. The exact number of included values was also random.

Definition 5-1 — $\Psi_{m<\|\mathscr{s}\|}(\mathscr{s} \subseteq I)$ is a function to define a random sample of size m, taken without replacement, from a non-strict subset of the complete index set (i.e. $\mathscr{s} \subseteq I$). For all $i \in I$, there is an equal probability of each index, i, being included in the sample. The size of the sample is strictly less than the number of elements in $\mathscr{s}$ (i.e. $m < \|\mathscr{s}\|$); therefore, $\Psi_m(\mathscr{s})$ is defined as a strict subset of $\mathscr{s}$.

In Definition 5-1, the sample size, m, is a user defined number. For this research, the actual number of data points included was sometimes randomly defined. When m is defined randomly or according to a function, $\psi(m)$ will be included in noation.

Definition 5-2 — $\psi(m)$ is a function to generate a sample size.

For each run of each model, all applicable data points had an equal chance of being included. Models with less than 4.5 million applicable points included all values; models with more than 4.5 million applicable points included a random subset of 4.5 to 4.8 million values according to equation (5.2.1).

$$\psi_1(m) = \left\{ \begin{array}{cc} \|\mathscr{s}\|, & \text{if } \|\mathscr{s}\| \leq A \\ U_{(A, \min[B, \|\mathscr{s}\|])}, & \text{otherwise} \end{array} \right\}, \tag{5.2.1}$$

Given that: $U_{(a,b)} \sim Uniform(a,b), A = 4.5(10)^6, B = 4.8(10)^6,$ *and* $\|\mathscr{s}\|$ is the number of elements in $\mathscr{s}$, a subset of the complete index set I.

5.2.1. Variable Selection

The independent variables, included in the first run of each model, were selected manually and were based on previous research and preliminary tests. Variables included potentially relevant variables related to passengers, vehicles, locations, times, and other indicators. To create each of the final models, the insignificant variables were removed step-wise, such that only the most insignificant variable was removed on a given run. Each run of the stepwise function included a different random subset of applicable values. The remaining variables were tested for their contribution to the model explanatory power as the first-variable and the last-variable. Given the number of data points, many variables are significant, but do not provide practical usefulness. Variables were additionally removed if they contributed less than 0.01% or 0.0001% as the first or last variable, respectively.

Relative Contributions

Finally, the contribution of each variable to the R-squared and its relative contribution was calculated for the resulting models. Unlike first-variable and last-variable, this estimate considers the correlations between variables. However, the computational complexity of the relative contribution functions required bootstrapping to provide useful information in a timely way. While optimized for efficient calculations, the *R* package “relaimpo” has limitations caused by the underlying formulas, which multiplicatively scale with each added independent variable (Grömping, 2006). Additionally, the computation time increases linearly with the number of data points up to about 140,000 data points. As such, a randomly selected 120 thousand values were included in each run such that each data point had an equal probability of being selected one time for up to fifty runs. The reported contributions are the average of those runs.

However, for models with more than about 15 independent variables, computation times increased beyond practicality, even with limited sample sizes. In these cases, and within the stepwise loops, a piecewise approach was used to estimate the contribution. The “relaimpo” package allows for a subset of independent variables to considered as single group, thus lowering the effective number of variables. The set of independent variables were partitioned into two groups of similar variables and two calculations sets were performed using one partition as a group. The partition average from the first/second set was proportionally divided between the relative contribution of the second/first set. These results are not the same as a complete relative contribution calculations, but provided a means to compare.

For the contributions reported in tables throughout this dissertation, a different approach was used that again is an estimate of the relative contribution. Table 5-1 shows an example using three contribution estimates. Reported contributions from each set are the average of five runs of about 500,000 data points. Running the three sets five times each takes less than half the time of running the complete model five times. Italicized numbers are calculated manually after the runs. An independent variable representing the sum of a specific variable type is substituted for its constituent parts. The first run includes all substituted variables and each subsequent run breaks apart one of those variables. In the example, there are two substituted variables. Final contribution begins as a mean of the runs, but also divides the substituted variables proportionally. Second, the scaled model adjusts the values based on the adjusted R-squared of the model using all datapoints.

Table 5-1 — Contribution and relative contribution calculation example for simplified $\left\{{}^{\Sigma T}\widehat{D}WL_t : t \in \mathbb{t}\right\}$ aggregated linear regression model. Adjusted R-squared of model using all datapoints is 0.7327.

	Contribution Partitions			**Final Contributions**			**Relative Contrib.**	
Variable	**Set 1**	**Set 2**	**Set 3**	***Mean***	**Scaled**	**Actual**	**Scaled**	**Actual**
${}^{\Sigma}VEH$	7.6%	6.9%	7.3%	*7.3%*	7.33%	6.59%	10.00%	9.00%
${}^{\Sigma}\hat{O}NS$	14.1%	13.5%	13.9%	*13.8%*	13.90%	13.24%	18.97%	18.06%
${}^{\Sigma}\hat{O}FFS$	8.4%	7.7%	8.1%	*8.1%*	8.12%	7.45%	11.08%	10.17%
$({}^{\Sigma}\hat{O}NS)^2$	5.9%	5.5%	5.7%	*5.7%*	5.75%	5.46%	7.85%	7.45%
$({}^{\Sigma}\hat{O}FFS)^2$	3.6%	3.2%	3.4%	*3.4%*	3.43%	3.15%	4.69%	4.30%
${}^{\Sigma}\hat{L}IFT$	3.7%	3.5%	3.6%	*3.6%*	3.60%	3.38%	4.92%	4.62%
$\sum[{}^{\Sigma L}VAR]$	14.3%	*18.8%*	14.3%	*15.8%*				
${}^{\Sigma L}TC$		1.1% *(6.0%)*		*(1.0%)*	0.96%	1.07%	1.31%	1.46%
${}^{\Sigma L}MALL$		1.4% *(7.4%)*		*(1.2%)*	1.18%	1.90%	1.62%	2.59%
${}^{\Sigma L}NEAR$		7.7% *(40.8%)*		*(6.5%)*	6.49%	7.46%	8.86%	10.18%
${}^{\Sigma L}FAR$		5.3% *(28.2%)*		*(4.5%)*	4.49%	5.09%	6.13%	6.95%
${}^{\Sigma L}OPP$		2.0% *(10.7%)*		*(1.7%)*	1.71%	1.91%	2.33%	2.60%
${}^{\Sigma L}AT$		1.3% *(6.7%)*		*(1.1%)*	1.07%	1.15%	1.46%	1.58%
$\sum[{}^{\Sigma Ls}VAR]$	10.8%	10.2%	*12.3%*	*11.1%*				
${}^{\Sigma Ls}TC$			0.6% *(4.6%)*	*(0.5%)*	0.52%	0.52%	0.70%	0.71%
${}^{\Sigma Ls}NEAR$			5.8% *(47.6%)*	*(5.3%)*	5.31%	5.64%	7.24%	7.70%
${}^{\Sigma Ls}FAR$			4.3% *(34.7%)*	*(3.9%)*	3.87%	3.97%	5.28%	5.41%
${}^{\Sigma Ls}OPP$			0.7% *(5.8%)*	*(0.6%)*	0.65%	0.62%	0.88%	0.85%
${}^{\Sigma Ls}AT$			0.9% *(7.3%)*	*(0.8%)*	0.81%	0.80%	1.11%	1.10%
$FREQ$	2.1%	1.8%	1.9%	*1.9%*	1.96%	1.72%	2.67%	2.35%
W_1^{AM}	0.1%	0.1%	0.1%	*0.1%*	0.10%	0.09%	0.14%	0.13%
W_1^{PM}	0.2%	0.2%	0.2%	*0.2%*	0.22%	0.19%	0.30%	0.26%
W_0^{P}	0.0%	0.0%	0.0%	*0.0%*	0.05%	0.05%	0.06%	0.06%
${}^{\Sigma Id}INT$	1.2%	1.2%	1.4%	*1.3%*	1.31%	1.41%	1.79%	1.92%
${}^{\Sigma Is}INT$	0.5%	0.4%	0.4%	*0.4%*	0.45%	0.40%	0.61%	0.54%

Table 5-1 shows the contribution and relative contribution (based on the above process) compared to a calculated version using the average of five runs of 500,000 data points. Some of the differences may be attributed to the different samples used in each run. While the values are not exactly the same, they are close enough for its intended practical application. For the types of comparisons made through Chapter 5, such small differences will not change the conclusions.

Final Models

The step-wise process was run in loops that tested variables for many different date, time, and location specific models. These estimates are a useful metric for differentiating models representing different subsets of the transit system. Following these looped runs, additional tests for new variables or combination of variables were tested manually using a stepwise process. The final models, reported as tables in this dissertation, were created based on those results, using all applicable data points.

5.2.2. Variance Inflation Factors

Many variables that could be used in regression models are highly correlated. For the set of variables used in the models, variance inflation factors (VIF) were calculated using the library "car" (Fox & Weisberg, 2018). In the event level models, the VIF remained low (i.e. less than five) for all included variables. $\hat{O}NS_i$ had the highest VIF, at just over 3. It is most strongly correlated with its square term.

For the aggregated regression models, the VIF of ${}^{\Sigma}\hat{O}NS_t$ increased to about 12 depending on the included variables. After aggregations, ${}^{\Sigma}\hat{O}NS_t$ shows a high correlation (>0.5) to other passenger movements, their square terms, and the number of service stops.

Of the variables related to passenger boardings. ${}^{\Sigma}\hat{O}FFS_t$ is the next highest value at about seven, followed by ${}^{\Sigma L}NEAR_t$ (i.e. the number of nearside stops serviced) and $({}^{\Sigma}\hat{O}NS_t)^2$, both at six. All other variables were below five It is not a reasonable limitation to exclude passenger movements or the number of stops from the aggregated models; as such, some inflation of the variance will occur. However, passenger boardings did not have the highest variance inflation factor of the aggregated variables.

A few related variables will not be included in models due to their extremely high VIF and correlations with other variable. The signalized versus unsignalized variable pairs for the downtown transit mall (i.e. $({}^{\Sigma L}MALL_t, {}^{\Sigma Ls}MALL_t)$ and $({}^{\Sigma L}_{skd}MALL_t, {}^{\Sigma Ls}_{skd}MALL_t)$) have VIF values greater than fifty when included. The pairs are also correlated at upwards of 90%. As such, the signalized version of the variables (i.e. ${}^{\Sigma Ls}MALL_t$ and ${}^{\Sigma Ls}_{skd}MALL_t$) will not be included in any final models. Other L and Ls variable pairs did not have such correlations and were tested.

5.2.3. Sample Sizes

In an effort to define the relationship between data sizes and usefulness of the results, this research will run regressions, at varied sample sizes, for door open duration (using event level data) and total travel time (using aggregated data). The estimated coefficients, from many runs of a model, may be plotted against the input sample size, which was defined using equation (5.2.2). The following examples are based on a linear regression model for door open duration that includes some variables excluded in the final model given by Table 5-2. Results specific to that model are discussed in Section 5.3.3.

$$\psi_2(m) = \left\lfloor \exp\left[U_{(\ln[N_0],\ln[(2)N_1])}\right] \right\rfloor, \tag{5.2.2}$$

Given that: $U_{(a,b)} \sim Uniform(a,b)$ *and* $\lfloor X \rfloor$ is a floor function for X; *and where:* N_0 is the smallest allowed sample size defined as: $N_0 = 10(N_{VAR} + 1)$, N_{VAR} is the number of independent variables, *and* N_1 is the user defined plot window.

Two main axis ranges were selected, which plot samples sizes up to about 10,000 or 100,000. Moving forward, N_1, from equation (5.2.2), will be given as N_{10} or N_{100} for plot windows up to $m = 10{,}000$ or $m = 100{,}000$, respectively. The largest sample sizes are defined as twice N_1, because plot windows extend beyond the highest labeled value. Results from sample sizes less than 12,000 or 130,000 will be largely visible for N_{10} and N_{100} plots, respectively.

Figure 5-1 is the first example plot. Both the left and right plots show the same data using the same y-axis range. The range for the y-axis was formulaically defined to show approximately 95% of the data points. Two horizonal lines are included: one, a grey line indicating the zero-point of the y-axis; and two, a colored coefficient line (color based on p-value) indicating the value of the coefficient from the complete regression model (using all available points). The minimum sample size is denoted by a vertical dashed line. For a model with 19 independent variables, the minimum sample size, N_0, is defined as 200. The color of each point is defined by the p-value of the coefficient. Insignificant coefficients (i.e. p-value ≥ 0.05) are gray, while green, blue, and purple points indicate increasing significance (i.e. decreasing p-values). Missing and NA coefficients are plotted on the coefficient as a red "×". Finally, the percent of non-zero values for the given independent variable and the 95% confidence interval (CI) for number of non-zero values for a N_1 sample size is given.

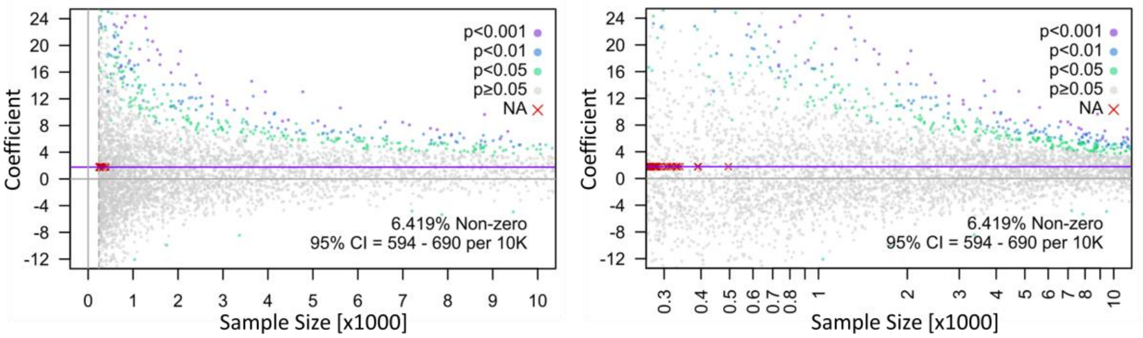

Figure 5-1 — Coefficients for ${}^{L}AT$ versus sample size (N_{10}) with a linear x-axis (left) and logarithmic x-axis (right). $\{{}^{T}\widehat{D}WL_i : i \in \Psi_{\psi_2(m)}(I)\}$ linear regression model using independent variable inputs shown in Table 5-2.

Figure 5-2 is the same data as provided in Figure 5-1 with an overlay of three equal width boxes. The label in each box is the approximate number of data point within the range of $\psi_2(m)$ values. With a linear x-axis, the number of data points within each equal-width section decreases as the sample size increases. With a logarithmic x-axis, the number of data point is approximately the same for any two equal width sections. The actual number varies because the sample size was selected randomly.

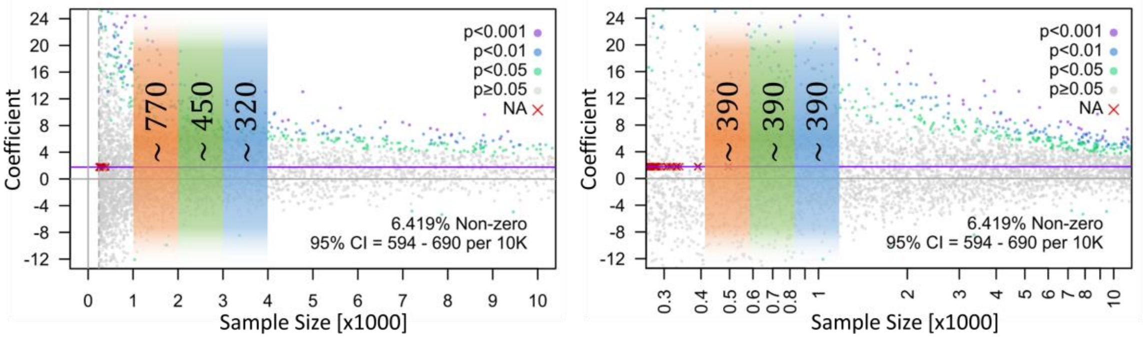

Figure 5-2 — Approximate number of data points for equal width plot regions for linear x-axis (left) and logarithmic x-axis (right). Same data as Figure 5-1.

Both the linear plots with linear axis and logarithmic axis display approximately 4,000 of the 5,000 regression runs. The same 5,000 runs are used for plots of different coefficients; as such, the specific data points outside the plot window changes depending on the given variable. Figure 5-3 is the same coefficient as the previous two figures, but uses N_{100}. A different 5,000 runs were performed for the N_{10} and N_{100} plots.

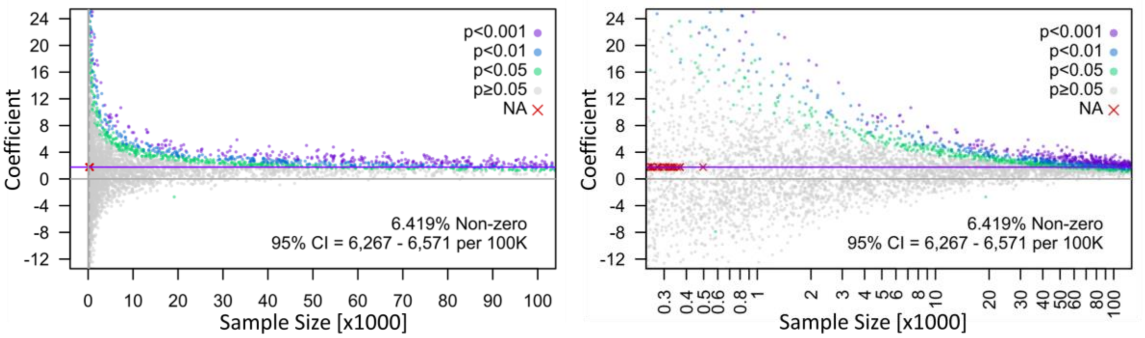

Figure 5-3 — Coefficients for ${}^{L}AT$ versus sample size (N_{100}) with a linear x-axis (left) and logarithmic x-axis (right). $\{{}^{T}\widehat{D}WL_i : i \in \Psi_{\psi_2(m)}(I)\}$ linear model with inputs from Table 5-2.

Final a third set of 5,000 results were calculated. Each result is the average of 10 independent runs with the same sample size. The 5,000 results are based on 50,000 runs total. The averages are reported as missing/NA if any one of the ten are missing/NA. The p-value is also the average of the ten runs. In Figure 5-4, the left plot shows the results from one run and the right shows average results of ten runs. The y-axis is the same for both plots. This example shows that while the range of possible values of the coefficients is reduced, nearly all of the results are insignificant. Furthermore, the number of missing data points is larger.

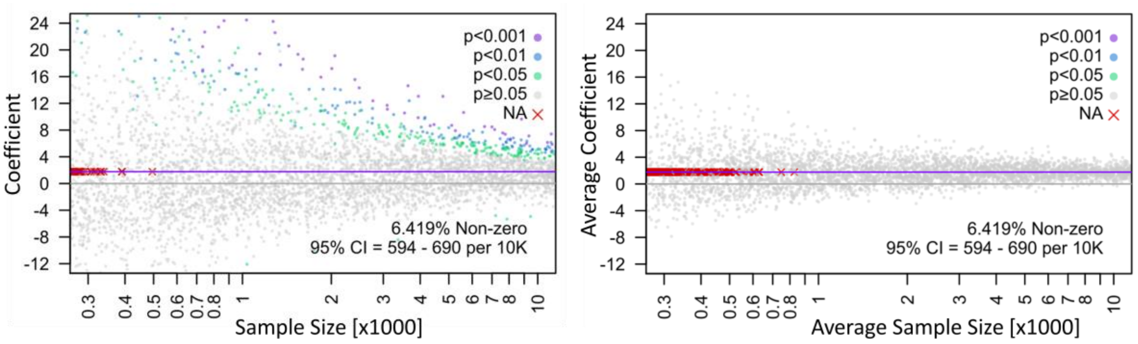

Figure 5-4 — Coefficients for one run (left) and average of ten runs (right) for ^{L}AT versus sample size (N_{10}). $\left\{{}^{T}\widehat{D}WL_i : i \in \Psi_{\psi_2(m)}(I)\right\}$ linear model with inputs from Table 5-2.

As a final example, Figure 5-5 gives the coefficient of the intercept and passenger boardings to highlight how a different number of data points are needed to produce consistent and significant results. While $\hat{O}NS$ is likely to be significant with a sample of just 300, the intercept is not consistently significant without a sample size of 30,000 (i.e. 100× larger). Additionally, the zero-line does not appear for $\hat{O}NS$, indicating a consistent sign, while negative results are possible (and sometimes significant) for the intercept. The plots (Figure 5-1 - Figure 5-5) introduced in this section are intended to explain how to read similar plots introduced later in this chapter and potential interpretations.

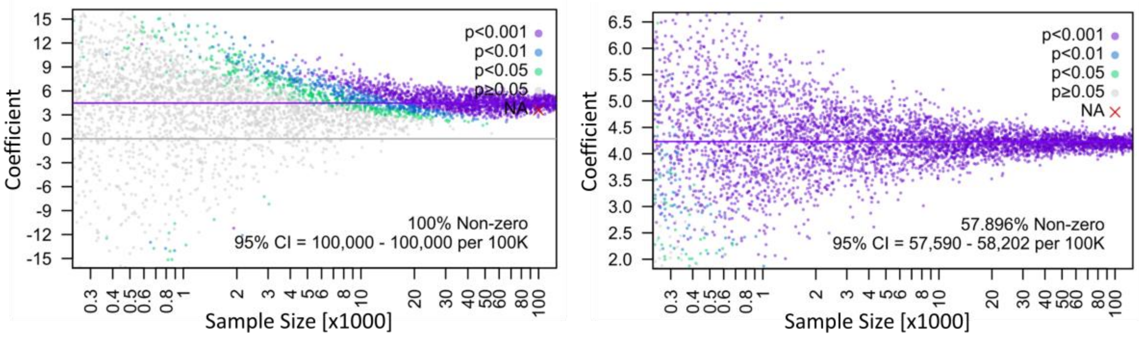

Figure 5-5 — Coefficients of intercept (left) and passenger boardings (i.e. $\hat{O}NS$) (right) versus sample size (N_{100}). $\left\{{}^{T}\widehat{D}WL_i : i \in \Psi_{\psi_2(m)}(I)\right\}$ linear model with inputs from Table 5-2.

Computation Times

Smaller samples decrease computation times. For R-studio and the libraries used in this research, samples and regressions of data-sets under 1,000,000 rows could be performed almost instantly. A regression of 1,000,000 data points is complete in about 1 second. However, a regression of 10,000,000 data points takes much longer than 10 seconds and the same is true for taking samples. A complete regression of 45,000,000 data points takes 5 to 10 minutes depending on the number of variables.

Model Explanatory Power

The adjusted R-squared and residual standard error are also dependent on the sample size of a model. Smaller sample sizes may potentially imply better performance than a model may actually provide. In the following example, sample sizes below 10,000 give results that generally over-state the adjusted R-squared and under-state the residual standard error. By $m = 100{,}000$, the model estimates are more evenly distributed around the full model values and the range values is much narrower.

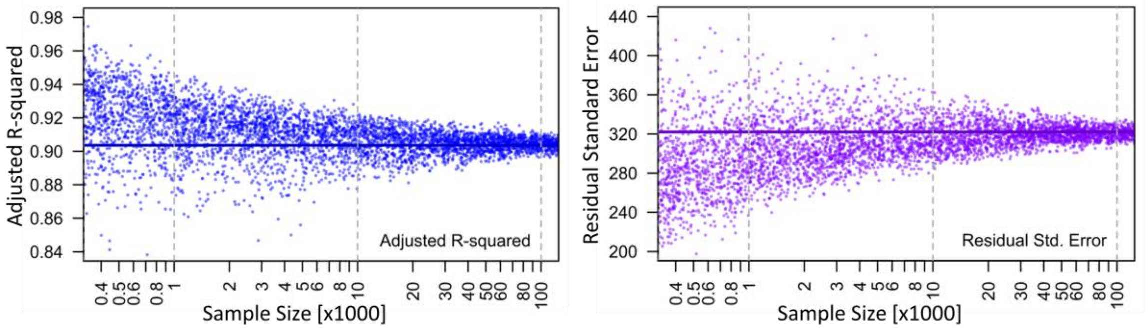

Figure 5-6 — Adjusted R-squared (left) and residual standard error (right) versus sample size (N_{100}). $\left\{ {}^{\Sigma T}\widetilde{T}RVL_{t} : t \in \Psi_{\psi_2(m)}(\mathbb{t}) \right\}$ aggregated linear regression using independent variable inputs shown in Table C-8.

5.3. Event-Level Bus Bay Service

Within a bus-bay, door open duration (i.e. ${}^{T}\widehat{D}WL$) and bus-bay stop duration (i.e. ${}^{T}\widehat{B}AY$) are modeled using linear and log-linear regressions.

5.3.1. Door Open Duration

Door open durations, ${}^{T}\widehat{D}WL$, has been well studied in the past. Using the ELD from this research, the variable coefficients and approximate contributions of resulting models are not dissimilar from previous publications. Table 5-2 shows the resulting linear model for all stops on weekdays and weekends. As in previous publications, passenger movements account for more than 70% of variable contributions to the Adjusted R-Squared. Passenger boardings $\left(\hat{O}NS\right)$ increase ${}^{T}\widehat{D}WL$ more than alightings $\left(\hat{O}FFS\right)$, and both see beneficial economies of scale, which are indicated by the negative coefficients for $\hat{O}NS^2$ and $\hat{O}FFS^2$. Of the stop locations, ${}^{L}TC$, ${}^{L}MALL$, ${}^{L}TP$, ${}^{L}P\&R$, and ${}^{L}AT$ remained in the model. The other stop location variables were either insignificant or non-practical. For vehicle interactions, leading $({}^{I}LEAD)$ and tailing $({}^{I}TAIL)$ vehicles each add about four seconds; waiting $({}^{I}WAIT)$ vehicles add about 15 seconds. Vehicles interacting from the same routes $({}^{I_S}INT)$ decreased these interaction times by about one second for each time one occurred.

Table 5-2 — Door open duration linear regression model for all service stops at all times of day. $\forall {}^{T}\widehat{D}WL_i \in \left\{{}^{T}\widehat{D}WL_i : i \in J\right\}$.

Variable Type	**Variable**	**Coefficient**	**Std Error**	**Contrib.**	**Rel-Imp**
Calculated Intercept	Intercept	4.91	0.0059		
Passenger Movements	$\hat{O}NS$	4.18	0.0024	7.31%	29.81%
	$\hat{O}FFS$	1.37	0.0019	1.69%	6.87%
	$\hat{O}NS^2$	-0.160	0.0002	2.25%	9.18%
	$\hat{O}FFS^2$	-0.003	0.0002	0.67%	2.72%
	$\hat{L}IFT$	34.69	0.0203	5.74%	23.39%
Bus Stop Locations (L)	${}^{L}TP$	7.54	0.0072	3.95%	16.10%
	${}^{L}TC$	5.45	0.0138	1.45%	5.91%
	${}^{L}MALL$	4.00	0.0132	0.45%	1.83%
	${}^{L}P\&R$	1.32	0.0104	0.04%	0.17%
	${}^{L}AT$	-0.22	0.0093	0.10%	0.42%
Traffic Signal	${}^{L}SIG$	1.24	0.0052	0.27%	1.10%
High-Frequency RTE	$FREQ$	0.79	0.0052	0.08%	0.31%
Weekdays	W_1^{AM}	-1.66	0.0078	0.08%	0.32%
	W_1^{PM}	-0.21	0.0065	0.01%	0.04%
Weekends	W_0^{P}	0.74	0.0089	0.02%	0.10%
Vehicle Interactions at Bus Stops (I)	${}^{I}LEAD$	3.73	0.0266	0.12%	0.47%
	${}^{I}TAIL$	3.88	0.0264	0.09%	0.37%
	${}^{I}WAIT$	16.02	0.0558	0.20%	0.81%
Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (Is)	${}^{Is}INT$	-1.12	0.0254	0.01%	0.04%
	$n = 45{,}616{,}055$			Adjusted R-Squared = 24.52%	
	p-value ≪ 0.001 for all variables				

The model in Table 5-2 applies to all service events $({}^{E}SVC)$. Additional models may be tailored to weekdays or weekends, to peak (AM, PM, or both) or off-peak times, and to specific location types. The model explanatory power change minorly by times and more substantively by location. Table 5-3 model explanatory power to predict ${}^{T}\widehat{D}WL$.

Table 5-3 — Door open duration linear regression model for for temporally and location specific models $\forall {}^{T}\widehat{D}WL_i \in \left\{ {}^{T}\widehat{D}WL_i : \phi \land i \in J \right\}$.

	Hours	**All Stops**	$\boldsymbol{^{L}MALL_i = 1}$	$\boldsymbol{\phi \coloneqq}$ $\boldsymbol{^{L}TC_i = 1}$	$\boldsymbol{^{L}MALL_i +}$ $\boldsymbol{^{L}TC_i = 0}$
All Days	*All*	24.52%	30.30%	7.19%	26.03%
Weekday	*All*	25.12%	31.00%	7.18%	26.58%
	AM-Peak	21.04%	26.96%	5.74%	20.52%
	PM-Peak	28.41%	31.56%	8.95%	28.80%
	Peak (AM & PM)	26.05%	32.43%	7.63%	26.37%
	Off-Peak	24.87%	30.37%	7.34%	26.94%
Weekend	*All*	22.73%	27.19%	8.13%	24.68%
	Peak	24.59%	29.26%	9.50%	26.38%
	Off-Peak	20.24%	24.23%	6.70%	22.31%

Each cell of this table is a unique model that applies to its specific location and times. While a model for transit centers (^{L}TC) during the *PM-Peak* can only account for 8.95% of data variability, these stop account for just 0.86% of the total stops in the transit system. In total, ^{L}TC and $^{L}MALL$ stops account for just 5.12% and 4.14% of data points, respectively. With the percent of total stops considered, ^{L}TC and $^{L}MALL$ models only account for 1.6% of total variability in the data. Models for all other stops locations account for the remainder.

Tailoring models may be useful for a narrow focus; but, to examine the entire system, the additional complexity warrants consideration. For example, examining weekends and weekdays separately for all stops requires two models which represent 17.6% and 82.4% of data points, respectively. The sum of their individual model performances, scaled to the number of data points, accounts for 24.70% of data variability. Yet there are only minor differences between the included variables and coefficients of the models. A gain of 0.18% must be weighed against the added complexity.

Figure 5-7 is the first of many similar plots that shows the combined model explanatory power using time and location specific models. Each regression model is scaled to the number of bus stops serviced and recombined. Each colored stack is a different set of temporally specific models including: (1) all data points; (2), weekdays and weekends; (3) peak and off-peak for weekdays and weekends; and (4) the AM-peak, PM-peak, and off-peak for weekdays and peak and off-peak for weekends. The grey stacks include the same temporal divisions as the stack to their left, but further divides each model into three based on location: (1) transit centers; (2); the downtown transit mall; and (3); all other locations.

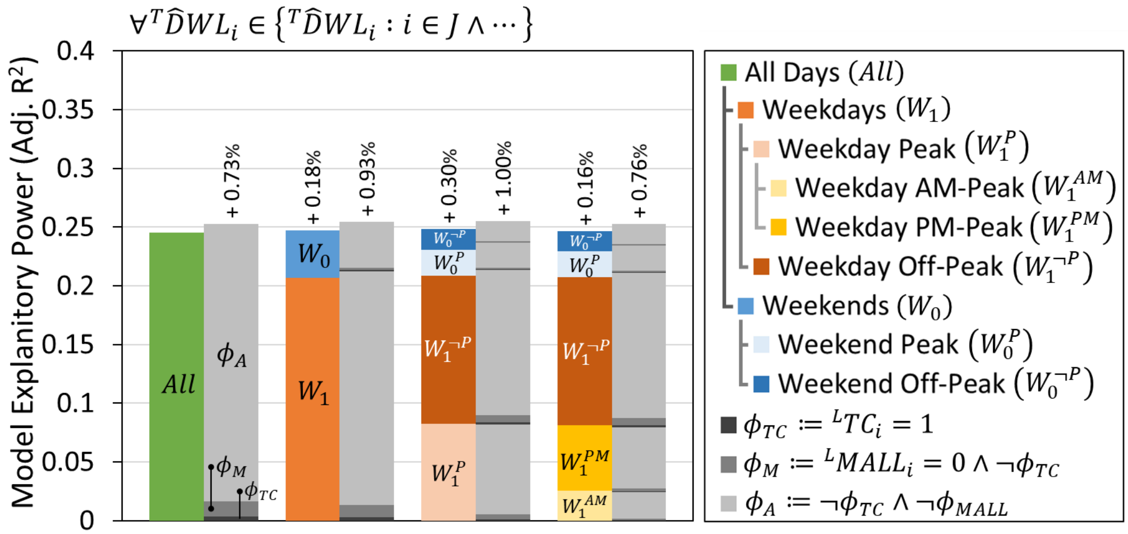

Figure 5-7 — Multiple linear regression models predicting door open duration (i.e. $\forall {}^{T}\widehat{D}WL_i \in \{ {}^{T}\widehat{D}WL_i : i \in J \}$. Location and temporally specific models are combined and scaled based on the number of data points.

In Figure 5-7, the maximum improvement of 1.00% is achieved using 12 models (i.e. peak and off-peak for weekdays and weekends broken down by location). There are other combinations, not shown in the figure that may further improve explanatory power.

Considering all models and the number of stop events they represent, a maximum of 25.57% of data variability is accounted for by using weekday peak, weekday off-peak, and weekend models for each location. Using these nine models, 1.05% of predictive power may be gained at a complexity expense of eight more models than baseline. However, the added complexity needs to be weighed against potential benefits.

When additional graphics, like Figure 5-7, are used for other independent variables. It should be assumed that other model combinations were tested. However, rather than viewing the set of stacked models as the complete set of model combinations, each graph should serve as an overview of how different model combinations compare.

Economies of Scale

An important check when including non-linear (i.e. squared) terms is to determine the limits of the estimated coefficients. In Figure 5-8, a plot of the total time given each additional passenger boarding and passenger alighting is plotted.

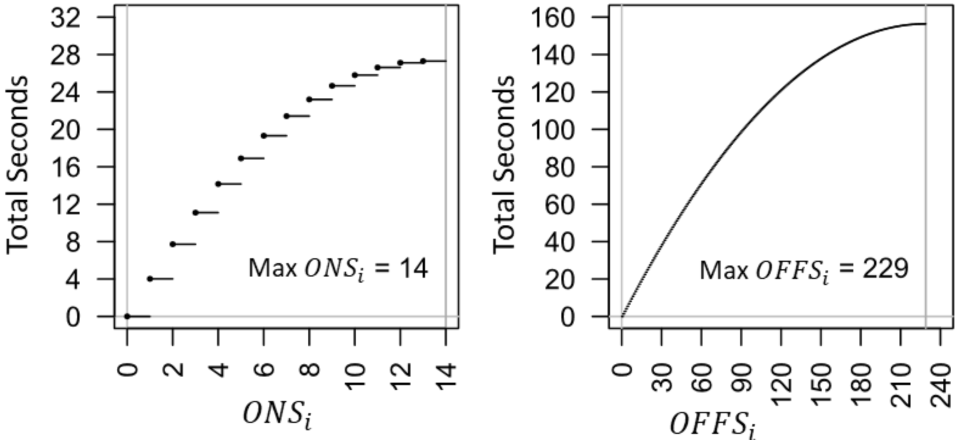

Figure 5-8 — Economies of scale for passenger movement coefficients from $\forall\, {}^{T}\widehat{D}WL_i \in \left\{{}^{T}\widehat{D}WL_i : i \in J\right\}$ linear regression model in Table 5-2

For passenger boardings, the estimated coefficients are no longer accurate given more 14 or more boardings at a single stop. In section 3.3.2, it was established that about 1/1000 stops will experience 19 or more boardings. As such, the coefficients for passenger boardings are questionable for high-usage stop events. However, other variables that identify high-usage stops, such as ${}^{L}TP$, are likely to capture some of the increases that should be attributed to more boardings.

For passenger alightings, the economies of scale are not particularly noticeable for a typical use case. For 19 alighting passengers, the savings are about 1 second. Even in rare high-usage cases where 30, 50, or 70 passengers (i.e. a completely full bus) were to alight at the same time, the savings would be about 3, 8, and 15 seconds, respectively. These absolute savings amount to 7%, 11%, and 15% of the alighting time. Given the issues with high-usage stops, previous research has shown that excluding high-usage stops improves model performance for other locations and that bus-stop specific models can be used to better estimate performance at stops with atypical usage (Glick & Figliozzi, 2017).

Log-linear Regressions

A more substantial gain may be achieved by using log-linear regression modeling for $\ln\left[{}^{T}\widehat{D}WL\right]$ (Table 5-4). Using only one model that includes all times and locations, 33.50% of the variability is captured. Using log-linear regression, the signs and relative magnitude of the independent variable coefficients are consistent with the linear models and results are similar to previous publications. While the individual variable contributions are different, passenger movements account for approximately 70% of variable contribution to the adjusted R-squared for both model types. The contribution of timepoints $({}^{L}TC)$ has increased while they have decreased for the transit mall $({}^{L}MALL)$.

Table 5-4 — Door open duration log-linear regression model for all service stops at all times of day. $\forall \ln\left[{}^{T}\widehat{D}WL_i\right] \in \left\{\ln\left[{}^{T}\widehat{D}WL_i\right] : i \in J\right\}$.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	1.8282	0.0009		
Passenger Movements	$\hat{O}NS$	0.2597	0.0003	14.08%	33.58%
	$\hat{O}FFS$	0.0950	0.0002	3.11%	10.71%
	$\hat{O}NS^2$	-0.0119	0.0000	3.69%	10.81%
	$\hat{O}FFS^2$	-0.0017	0.0000	0.97%	5.82%
	$\hat{L}IFT$	1.1406	0.0023	5.05%	11.69%
Timepoints	${}^{L}TP$	0.2367	0.0008	3.57%	13.61%
Bus Stop Locations	${}^{L}TC$	0.0534	0.0015	0.35%	4.99%
	${}^{L}MALL$	0.1837	0.0015	0.81%	3.34%
Weekday	$WDAY$	-0.0400	0.0007	0.13%	0.04%
Traffic Signal	${}^{L}SIG$	0.0870	0.0006	1.03%	2.72%
Frequent-Service $ROUTE$	$FREQ$	0.0411	0.0006	0.15%	0.47%
Vehicle Interactions at Bus Stops (I)	${}^{I}LEAD$	0.1525	0.0030	0.16%	0.90%
	${}^{I}TAIL$	0.2116	0.0030	0.26%	0.78%
	${}^{I}WAIT$	0.3419	0.0064	0.12%	0.44%
Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	${}^{Is}INT$	-0.0475	0.0029	0.01%	0.11%
	n =4,780,751 p-value $\ll$ 0.001 for all variables			Adjusted R-Squared=33.50%	

Table 5-5 shows adjusted R-squared for spatially and temporally specific log-linear models. The explanatory power for log-linear models somewhat mirrors the changes to linear models, with some exceptions. For example, the location specific models underperformed some models that didn't specify locations. In particular, models for weekdays peaks and ϕ_A lost more than 1% each.

Table 5-5 — Door open duration log-linear regression model for temporally and location specific models $\forall \ln\left[{}^{T}\widehat{D}WL_i\right] \in \left\{\ln\left[{}^{T}\widehat{D}WL_i\right] : i \in J\right\}$.

	Hours	**All Stops**	$\boldsymbol{{}^{L}MALL_i = 1}$	$\boldsymbol{\phi := {}^{L}TC_i = 1}$	$\boldsymbol{{}^{L}MALL_i + {}^{L}TC_i = 0}$
All Days	*All*	33.50%	39.95%	12.56%	33.26%
Weekday	*All*	34.39%	40.49%	12.49%	33.85%
	AM-Peak	31.92%	35.97%	10.48%	30.66%
	PM-Peak	39.44%	41.84%	15.45%	38.07%
	Peak (AM & PM)	36.92%	42.35%	13.38%	35.63%
	Off-Peak	33.05%	39.48%	12.68%	32.98%
Weekend	*All*	30.33%	37.08%	14.15%	31.27%
	Peak	32.72%	38.74%	16.36%	33.55%
	Off-Peak	27.42%	34.65%	12.03%	28.39%

Figure 5-9 show that multiple models for the network resulted in a maximum gain of just 0.25%. But, the figure also demonstrates how adding models will not always improve predictive power.

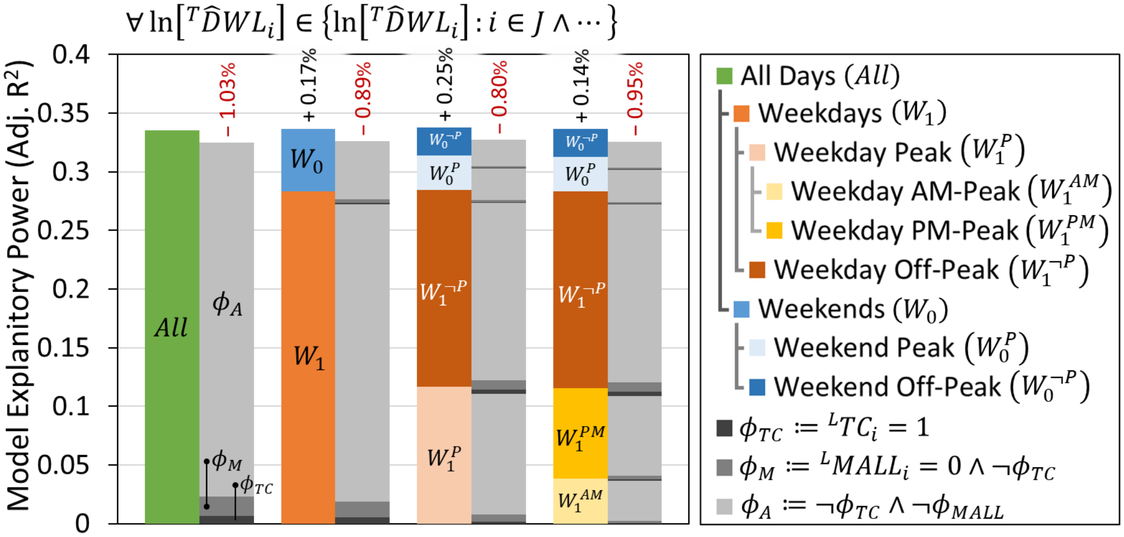

Figure 5-9 — Multiple log-linear regression models predicting door open duration (i.e. $\forall \ln\left[{}^{T}\widehat{D}WL_i\right] \in \left\{\ln\left[{}^{T}\widehat{D}WL_i\right] : i \in J\right\}$). Location and temporally specific models are combined and scaled based on the number of data points.

An additional implication of the lower adjusted R-squared is that much of the benefit of log-linear models is concentrated in the urban center of Portland and at transit centers. Bus operations, for both location types, have operational and logistical requirements that are not typical for other stops, like the requirement to stop at all mall stops, regardless of passenger activity.

5.3.2. Bus-Bay Stop Durations

Stop durations, ${}^{T}\widehat{B}AY$, may be modeled using the same variables as ${}^{T}\widehat{D}WL.$ Using linear and log-linear regression, 25.78% and 29.81% of data variability are, respectively, accounted for in model of ${}^{T}\widehat{B}AY$ for all times and locations. Table 5-6 shows the linear model predicting ${}^{T}\widehat{B}AY$ for all days and times. The primary differences between the models are intuitive. ${}^{T}\widehat{D}WL$ is part of ${}^{T}\widehat{B}AY$, but not the reverse. As such, variables that applied to the former are likely to apply to the latter.

The main changes from the ${}^{T}\widehat{D}WL$ and ${}^{T}\widehat{B}AY$ models are the: one, the inclusion of farside locations $({}^{L}FAR)$; two, the increased coefficients of the *intercept*, ${}^{L}SIG$, and interaction (I) variables; three, the exclusions of ${}^{L}AT$ and ${}^{Is}INT$; and four, changes to relative contributions of the variables. Passenger movements account for a slim majority of contribution to explanatory power when predicting ${}^{T}\widehat{B}AY$, versus 70% for ${}^{T}\widehat{D}WL.$ ${}^{L}SIG$ accounts for more than 10% of the R-Squared, up from 1%. The increase from 15 seconds to 35 seconds for waiting $({}^{I}WAIT)$ vehicles may be accounted for by common driver behaviors while waiting. Often, drivers will be prepared to move as soon as the other vehicle passes by having the doors closed and pulling slightly forward or out from the curb.

Table 5-6 — Bus-bay stop duration linear regression model for all service stops at all times of day. $\forall {}^{T}\widehat{B}AY_i \in \{{}^{T}\widehat{B}AY_i : i \in J\}$.

Variable Type	**Variable**	**Coefficient**	**Std Error**	**Contrib.**	**Rel-Imp**
Calculated Intercept	Intercept	15.70	0.0085		
Passenger Movements	$\hat{O}NS$	5.04	0.0034	5.60%	21.71%
	$\hat{O}FFS$	1.71	0.0027	1.64%	6.35%
	$\hat{O}NS^2$	-0.194	0.0003	1.74%	6.77%
	$\hat{O}FFS^2$	-0.006	0.0004	0.62%	2.39%
	$\hat{L}IFT$	36.86	0.0290	3.22%	12.48%
Timepoint	${}^{L}TP$	13.83	0.0103	5.31%	20.60%
Bus Stop Locations	${}^{L}TC$	6.85	0.0198	1.30%	5.03%
	${}^{L}MALL$	10.97	0.0192	1.51%	5.85%
	${}^{L}FAR$	-5.20	0.0080	0.78%	3.04%
	${}^{L}P\&R$	-1.33	0.0133	0.06%	0.24%
Traffic Signal	${}^{L}SIG$	7.82	0.0074	2.47%	9.57%
High-Frequency RTE	$FREQ$	0.79	0.0074	0.04%	0.14%
Weekdays	W_1^{AM}	-1.23	0.0112	0.03%	0.12%
	W_1^{PM}	1.52	0.0093	0.07%	0.26%
Weekends	W_0^{P}	1.80	0.0127	0.04%	0.15%
Vehicle Interactions at Bus Stops (I)	${}^{I}LEAD$	10.47	0.0381	0.35%	1.36%
	${}^{I}TAIL$	14.05	0.0378	0.46%	1.79%
	${}^{I}WAIT$	35.31	0.0802	0.45%	1.75%
	${}^{I}JUMP$	2.91	0.0805	0.01%	0.06%
Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (Is)	${}^{Is}INT$	-1.63	0.0367	0.04%	0.14%
	$n = 45{,}514{,}643$			Adjusted R-Squared = 25.78%	
	p-value $\ll 0.001$ for all variables				

Breaking down linear models (Table 5-7) and log-linear models (Table 5-8) into multiple locations and times, also does not result in large gains to the system as a whole. At best, an absolute increase 0.25% may be gained by using three models: weekday peak, weekday off-peak, and weekends.

Table 5-7 — Bus-bay stop duration linear regression models for temporally and location specific models $\forall {}^{T}\widehat{B}AY_i \in \left\{ {}^{T}\widehat{B}AY_i : \phi \wedge i \in J \right\}$.

				$\boldsymbol{\phi \coloneqq}$	$\boldsymbol{^{L}MALL_i +}$
	Hours	**All Stops**	$\boldsymbol{^{L}MALL_i = 1}$	$\boldsymbol{^{L}TC_i = 1}$	$\boldsymbol{^{L}TC_i = 0}$
All Days	*All*	25.78%	27.21%	8.75%	23.92%
Weekday	*All*	26.37%	27.53%	8.62%	24.09%
	AM-Peak	23.61%	22.09%	8.81%	19.36%
	PM-Peak	28.31%	27.84%	9.21%	25.02%
	Peak (AM & PM)	26.74%	27.32%	8.95%	23.25%
	Off-Peak	26.18%	27.14%	8.79%	24.72%
Weekend	*All*	24.31%	25.21%	10.59%	23.47%
	Peak	24.98%	27.19%	12.13%	24.62%
	Off-Peak	22.53%	22.30%	9.08%	21.77%

Table 5-8 — Bus-bay stop duration log-linear regression models for temporally and location specific models $\forall \ln\left[{}^{T}\widehat{B}AY_i\right] \in \left\{ \ln\left[{}^{T}\widehat{B}AY_i\right] : \phi \wedge i \in J \right\}$.

				$\boldsymbol{\phi \coloneqq}$	$\boldsymbol{^{L}MALL_i +}$
	Hours	**All Stops**	$\boldsymbol{^{L}MALL_i = 1}$	$\boldsymbol{^{L}TC_i = 1}$	$\boldsymbol{^{L}TC_i = 0}$
All Days	*All*	29.77%	26.94%	11.57%	27.49%
Weekday	*All*	30.24%	27.14%	11.43%	27.72%
	AM-Peak	28.11%	21.54%	11.42%	23.94%
	PM-Peak	33.21%	27.74%	12.60%	29.82%
	Peak (AM & PM)	31.42%	26.73%	11.96%	27.72%
	Off-Peak	29.62%	26.74%	11.66%	27.61%
Weekend	*All*	27.52%	25.55%	13.76%	26.46%
	Peak	29.20%	26.87%	15.51%	27.97%
	Off-Peak	25.65%	23.62%	12.20%	24.56%

Figure 5-10 and Figure 5-11 highlight how losses for the total explanatory power, for the location specific variables, are more pronounced for ${}^{T}\widehat{B}AY_i$ than they were for ${}^{T}\widehat{D}WL_i$. Using one weekday and one weekend model for results in a small gain for linear regressions, but almost none for the log-linear. In general, the log-linear model for ${}^{T}\widehat{D}WL_i$ provided a larger benefit than for ${}^{T}\widehat{B}AY_i$. The differences may be accounted for by the different shapes of the distributions as was shown in Figure 4-1.

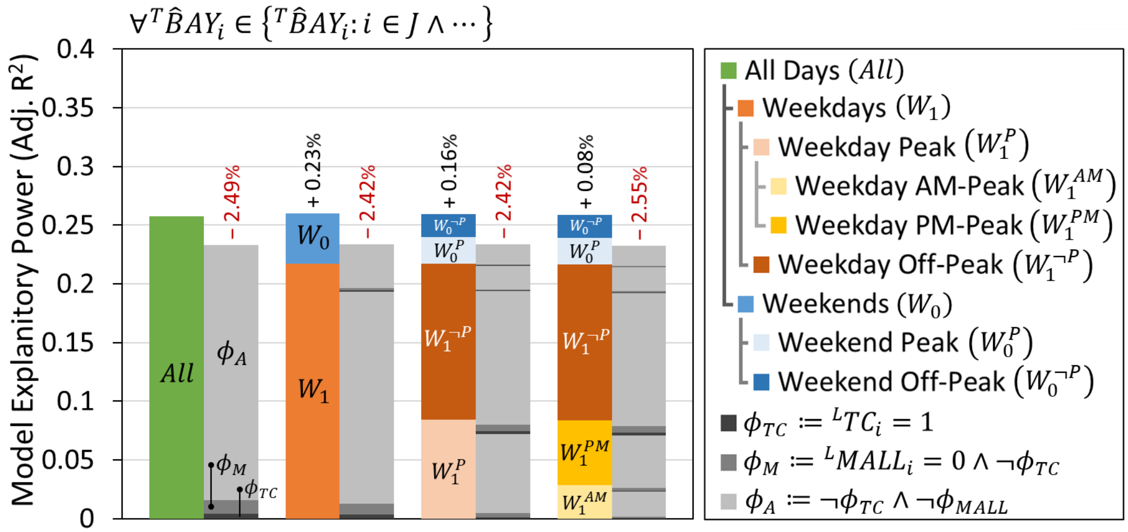

Figure 5-10 — Multiple linear regression models predicting bus-bay stop duration (i.e. $\forall {}^T\widehat{B}AY_i \in \{{}^T\widehat{B}AY_i : \phi \wedge i \in J\}$). Location and temporally specific models are combined and scaled based on the number of data points.

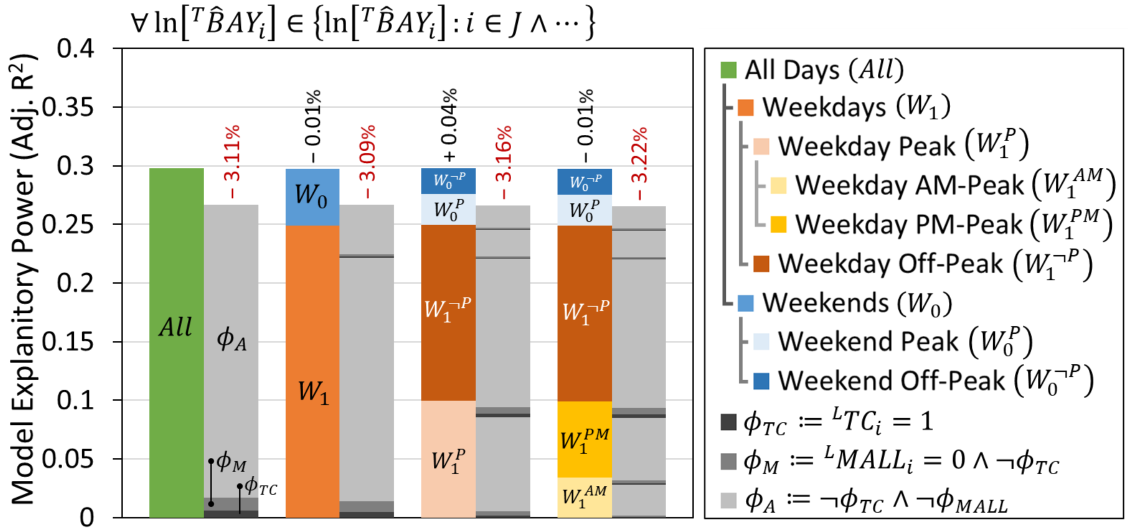

Figure 5-11 — Multiple log-linear regression models predicting bus-bay stop duration (i.e. $\forall \ln[{}^T\widehat{B}AY_i] \in \{\ln[{}^T\widehat{B}AY_i] : \phi \wedge i \in J\}$). Location and temporally specific models are combined and scaled based on the number of data points.

The passenger movement economies of scale are similar the trends observed for door open duration. Maximum boardings are the same, but the alightings savings are increased due to the coefficient difference of -0.003 for ${}^{T}\widehat{D}WL$ and -0.006 for ${}^{T}\widehat{B}AY$. For the extreme cases of 30, 50, and 70 alightings at a single stop, the time (and percentage) savings are 5 (11%), 15 (18%), and 30 (25%), respectively.

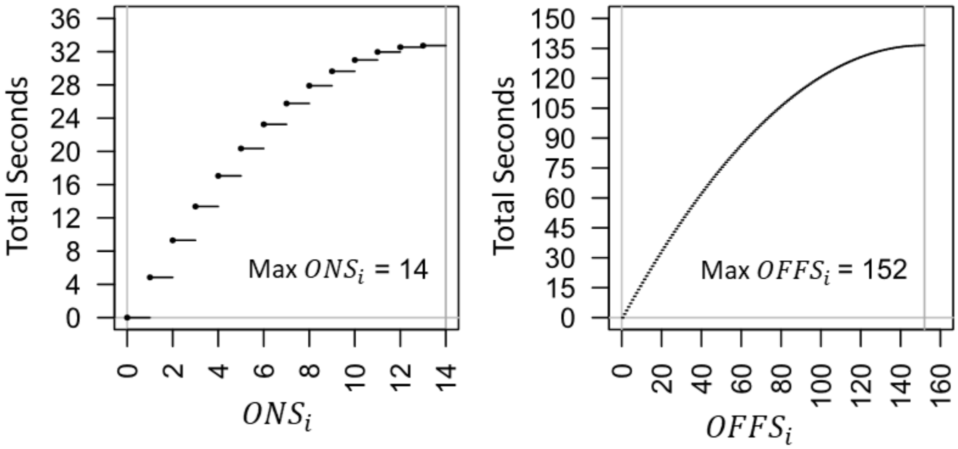

Figure 5-12 — Economies of scale for passenger movement coefficients from $\forall {}^{T}\widehat{B}AY_i \in \{{}^{T}\widehat{B}AY_i : i \in J\}$ linear regression model in Table 5-6.

5.3.3. Sample Sizes

To discuss sample sizes for the stop event data, the door open duration will be the focus. The general trends observed may be reasonable assumed to apply to other models and may be checked by rerunning the regressions with any specific model.

Passenger Movements

For the ELD, Figure 5-13 shows the coefficients for passenger movements using the independent variable inputs given by Table 5-2. Nearly all sample sizes give significant results for $\widehat{O}NS$ (left), but sample sizes above $m =$ 1,000 are needed before $\widehat{O}FFS$ (right) results are consistently significant. In both cases, the range of estimated coefficients are generally positive and narrow with increasing sample size. Above $m =$ 10,000, the estimated coefficients are generally within the range of values provided in previous literature. Also, there were no missing coefficients within 5,000 runs.

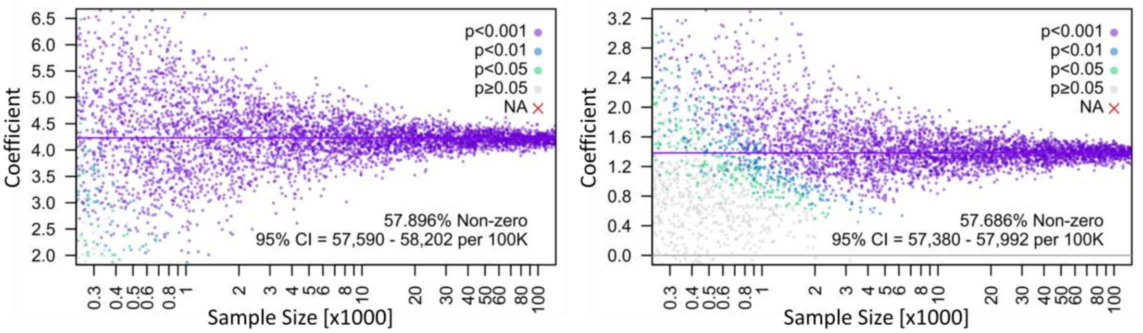

Figure 5-13 — Coefficients for $\hat{O}NS$ (left) and $\hat{O}FFS$ (right) versus sample size (N_{100}). $\left\{{}^{T}\widehat{D}WL_i : i \in \Psi_{\psi_2(m)}(I)\right\}$ linear model with inputs from Table 5-2.

Related to passenger boardings is the square terms. Figure 5-14 shows that the coefficients of for the square terms of passenger boardings and alightings do not follow the same trends.

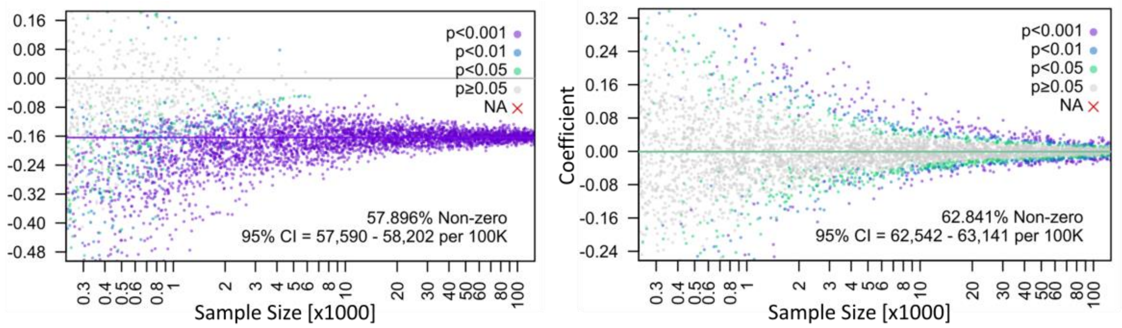

Figure 5-14 — Coefficients for $\hat{O}NS^2$ (left) and $\hat{O}FFS^2$ (right) versus sample size (N_{100}). $\left\{{}^{T}\widehat{D}WL_i : i \in \Psi_{\psi_2(m)}(I)\right\}$ linear model with inputs from Table 5-2.

Looking at $m > 10{,}000$, $\hat{O}NS^2$ (left) is consistently significant and negative; in contrast, $\hat{O}FFS^2$ (right) fluctuates around zero and gives significant negative and positive results even for $m > 100{,}000$. While previous research has shown that passenger alightings do benefit from economies of scale, large sample sizes are needed before results are consistently negative. This plot may indicate that including $\hat{O}FFS^2$ may be problematic even for large samples and results should be evaluated using metrics other than significance (e.g. contribution and relative contribution).

A final passenger movement is $\hat{L}IFT$. Figure 5-15 shows a comparison between one run (left) and the average of ten runs (right). In both cases, the coefficients are consistently significant with $m < 1{,}000$, but the range of values is much narrower using the average. However, when $m < 500$, missing results are much more common, which is an effect of the low percentage of non-zero observations.

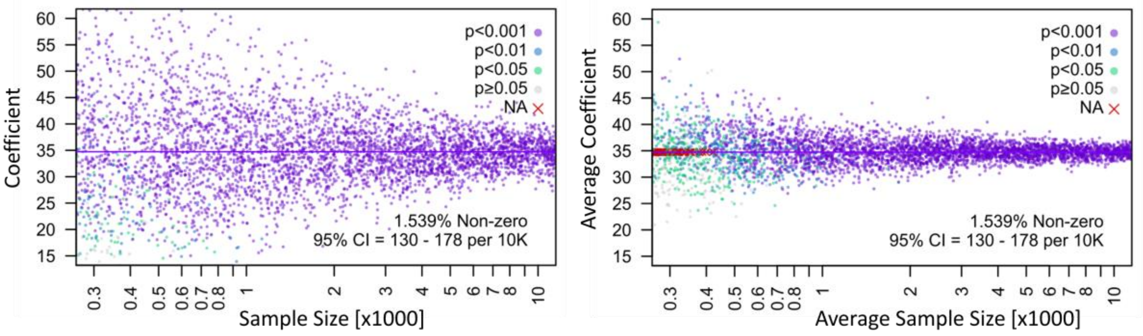

Figure 5-15 — Coefficients for one run (left) and average of ten runs (right) for $\hat{L}IFT$ versus sample size (N_{10}). $\{{}^{T}\hat{D}WL_i : i \in \Psi_{\psi_2(m)}(I)\}$ linear model with inputs from Table 5-2.

Location Variables

Each unique location variable shows a different level of significance and needed sample size for consistent results. In Figure 5-16, only timepoints (top-left) show consistently significant results for the smaller sample sizes. For the remaining location variables (${}^{L}SIG$, ${}^{L}TC$, and ${}^{L}MALL$), significant results occur along the upper edge of the distributions, but each approaches their respective coefficient line from the complete model. The similarities between those plots are not an effect of the number of non-zero datapoints, which vary wildly; rather, it is an effect of influence. While just 469-556 non-zero data points are needed in a $m = 10{,}000$ samples for transit centers, more than 5,000 non-zero data points are need for traffic signals to produce similar results.

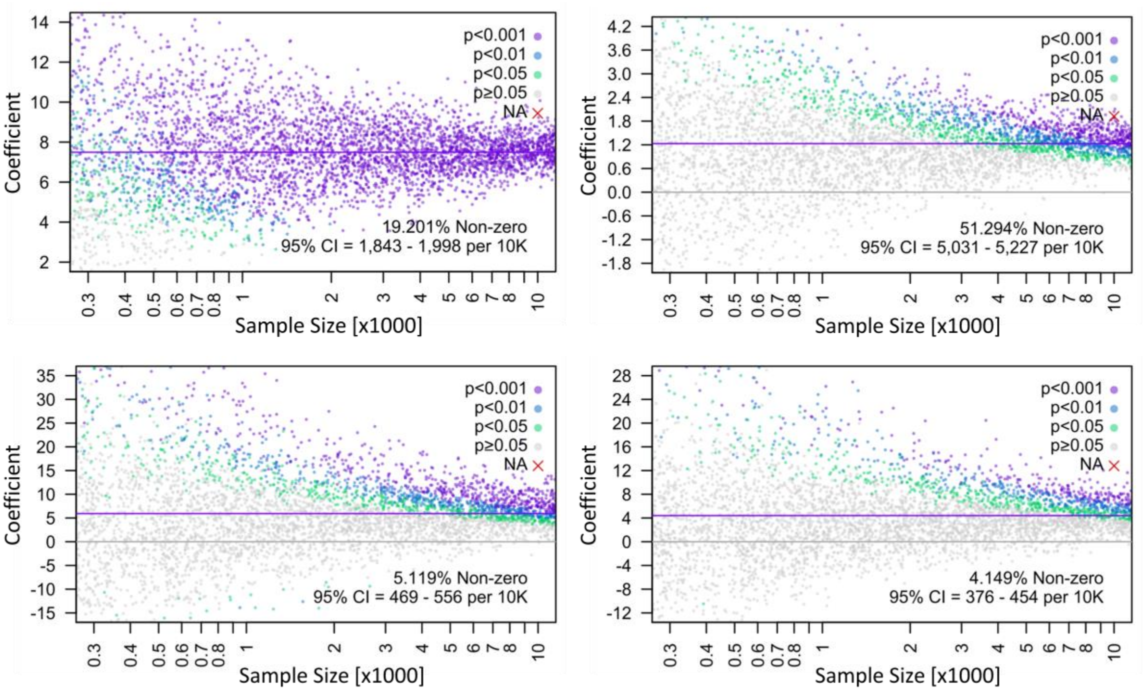

Figure 5-16 — Coefficients for ${}^{L}TP$ (top-left), ${}^{L}SIG$ (top-right), ${}^{L}TC$ (bottom-left), and ${}^{L}MALL$ (bottom-right) versus sample size (N_{10}). $\left\{{}^{T}\widehat{D}WL_i : i \in \Psi_{\psi_2(m)}(I)\right\}$ linear model with inputs from Table 5-2.

Vehicle Interactions

Figure 5-17 and Figure 5-18 both show vehicle interactions. The former includes the same variables as the model from Table 5-2. The latter includes variables dropped from that model and helps highlight why. For each plot in Figure 5-17, a notable feature is the number of missing results, which are an effect of the low percentage of non-zero observations. For leading, tailing, and waiting vehicles, the results are increasingly significant, positive and distributed around their respective coefficient lines. However, for same route interactions (bottom right), the results continue to bounce around zero and does not appear to be obviously approaching a specific value.

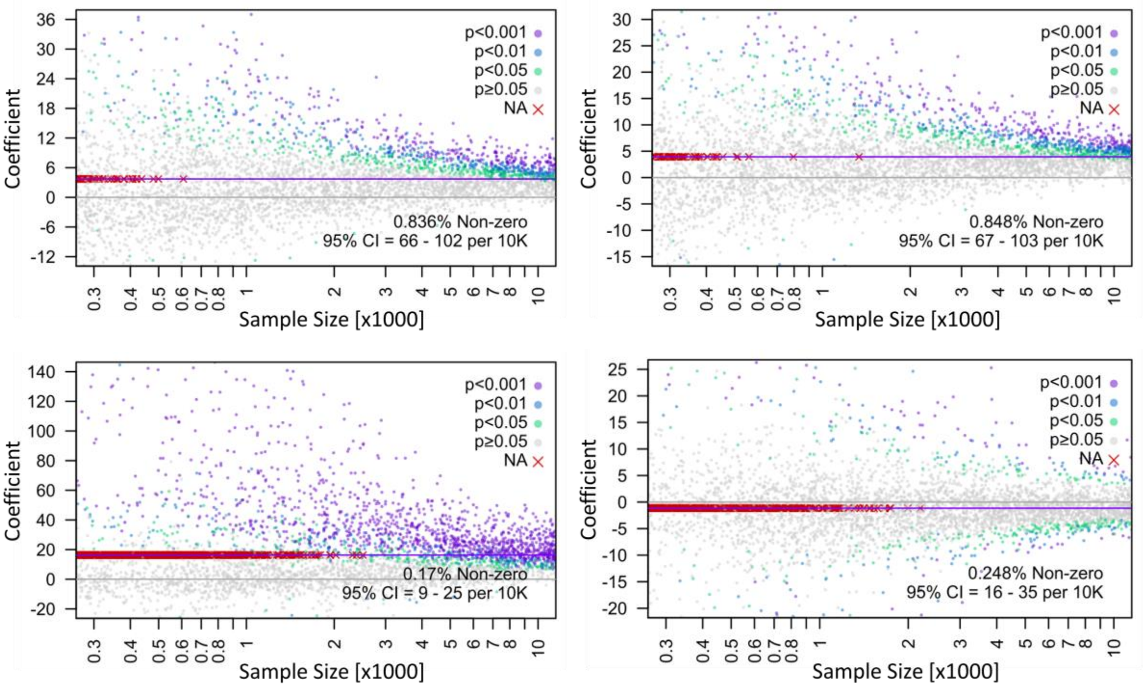

Figure 5-17 — Coefficients for $^{I}LEAD$ (top-left), $^{I}TAIL$ (top-right), $^{I}WAIT$ (bottom-left), and ^{Is}INT (bottom-right) versus sample size (N_{10}). $\left\{^{T}\widehat{D}WL_i : i \in \Psi_{\psi_2(m)}(I)\right\}$ linear model with inputs from Table 5-2.

The left plot from Figure 5-18 is also for same route interactions, as the sample size increases, the results start showing consistent results approaching a specific negative value. While interactions between vehicles of different routes may be useful to include in model for samples sizes around $m = 10{,}000$, same route interactions may prove problematic until around 100,000 observations. Jumping interactions (right) were inconsistent even at higher sample sizes are were dropped from the final models. For the 5,000 runs shown only a few are significant and while values are converging, they are not obviously converging to a non-zero number.

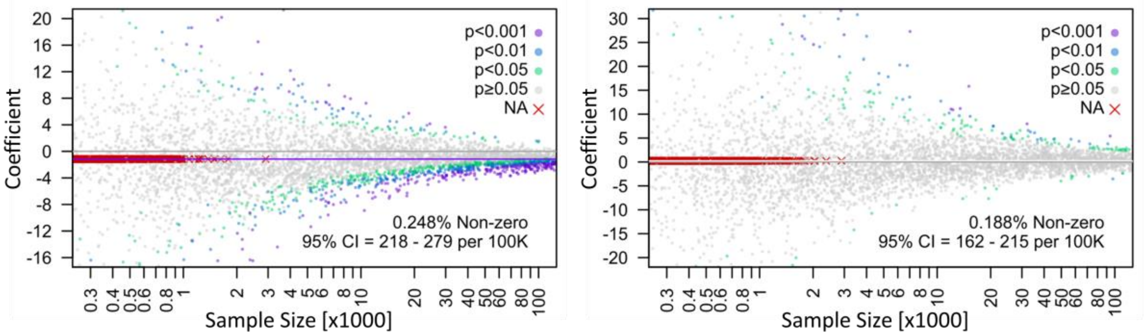

Figure 5-18 — Coefficients for ${}^{Is}INT$ (left) and ${}^{I}JUMP$ (right) versus sample size (N_{100}). $\{{}^{T}\widehat{D}WL_i : i \in \Psi_{\psi_2(m)}(I)\}$ linear model with inputs from Table 5-2.

5.4. Aggregated Bus-Bay Service

The variables included in the aggregated models are based on the ELD, but can provide additional clarity. These models cannot be directly focused on a specific stop type as multiple stops are included in each segment. As such, models for are run for all segments $\{{}^{\Sigma L}VAR_t : t \in \mathbb{t}\}$, and for timepoint-segment predicated on ϕ (i.e. $\{{}^{\Sigma L}VAR_t : (t \in \mathbb{t}) \wedge \phi\}$), where ϕ is defined in equation (5.4.1).

$$(5.4.1) \qquad \phi = \left\{ \begin{array}{r} \phi_{TC} \coloneqq ({}^{\Sigma L}TC_t > 0) \\ \phi_M \coloneqq ({}^{\Sigma L}MALL_t > 0) \wedge ({}^{\Sigma L}TC_t = 0) \\ = {}^{\Sigma L}MALL_t > 0 \wedge \neg\phi_{TC} \\ \phi_A \coloneqq ({}^{\Sigma L}MALL_t > 0) \wedge ({}^{\Sigma L}TC_t = 0) \\ = \neg\phi_M \wedge \neg\phi_{TC} \end{array} \right\}$$

Each timepoint-segment will belong to only one of the three divisions with no overlapping timepoint-segments. These divisions allow for segments with transit centers, segments with stops on the mall, and all other locations to be compared.

5.4.1. Door Open Duration

There are approximately 4.5 million data points that represent all 45.7 million service events. Table 5-9 shows the results of linear regression for all days of the week and times of day. The models for ${}^{\Sigma T}\widehat{D}WL_{t}$ capture much more of the variability in the data, which is an expected outcome of data aggregation.

Table 5-9 — $\left\{{}^{\Sigma T}\widehat{D}WL_{t} : t \in \mathbb{t}\right\}$ aggregated linear regression model.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	-3.94	0.0985		
Number of Vehicles	${}^{\Sigma}VEH$	5.49	0.0460	7.43%	10.13%
Total Passenger Movements	${}^{\Sigma}\hat{O}NS$	3.79	0.0067	13.97%	19.06%
	${}^{\Sigma}\hat{O}FFS$	1.52	0.0068	8.21%	11.20%
	$({}^{\Sigma}\hat{O}NS)^2$	-0.004	0.0000	5.78%	7.89%
	$({}^{\Sigma}\hat{O}FFS)^2$	-0.001	0.0000	3.48%	4.75%
	${}^{\Sigma}\hat{L}IFT$	40.20	0.1017	3.62%	4.94%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	15.53	0.0550	0.94%	1.28%
	${}^{\Sigma L}MALL$	10.17	0.0309	1.16%	1.58%
	${}^{\Sigma L}NEAR$	5.29	0.0184	6.34%	8.66%
	${}^{\Sigma L}FAR$	5.20	0.0265	4.39%	5.99%
	${}^{\Sigma L}OPP$	6.32	0.0375	1.67%	2.28%
	${}^{\Sigma L}AT$	7.59	0.0442	1.05%	1.43%
Total Serviced Bus Stop Locations Near Traffic Signals (ΣLs)	${}^{\Sigma Ls}TC$	3.93	0.0908	0.51%	0.70%
	${}^{\Sigma Ls}NEAR$	1.32	0.0234	5.27%	7.20%
	${}^{\Sigma Ls}FAR$	1.48	0.0330	3.85%	5.25%
	${}^{\Sigma Ls}OPP$	0.91	0.0716	0.64%	0.88%
	${}^{\Sigma Ls}AT$	5.52	0.0741	0.81%	1.10%
High-Frequency RTE	$FREQ$	2.62	0.1099	1.99%	2.72%
Weekdays	W_1^{AM}	-21.80	0.1629	0.10%	0.14%
	W_1^{PM}	-6.52	0.1391	0.22%	0.30%
Weekends	W_0^{P}	7.31	0.1676	0.05%	0.06%
Total Vehicle Interactions at Bus Stops Between Different RTE (ΣId)	${}^{\Sigma Id}LEAD$	2.59	0.1258	0.83%	1.13%
	${}^{\Sigma Id}WAIT$	24.73	0.3380	0.36%	0.50%
	${}^{\Sigma Id}JUMP$	-1.95	0.3272	0.16%	0.22%
Total Vehicle Interactions at Bus Stops Within the Same RTE (ΣIs)	${}^{\Sigma Is}INT$	-2.56	0.1378	0.45%	0.62%
	$n = 4{,}525{,}801$ Adjusted R-Squared = 73.27% p-value $\ll 0.001$ or all variables				

The coefficients for passenger movements are larger than for the event-level data, but have similar relative magnitudes. The model's passenger movement variables (i.e. ${}^{\Sigma}\hat{O}NS$, ${}^{\Sigma}\hat{O}FFS$, $({}^{\Sigma}\hat{O}NS)^2$, $({}^{\Sigma}\hat{O}FFS)^2$, and ${}^{\Sigma}\hat{L}IFT$) account for just under half of the adjusted R-squared value. About a third of the variation is accounted for by locations and are divided between number of services at each location type $({}^{\Sigma L}VAR)$ and the total number of signalized services $({}^{\Sigma Ls}VAR)$. ${}^{\Sigma L}VAR$ and ${}^{\Sigma Ls}VAR$ form a pair, such that ${}^{\Sigma Ls}VAR$ is an additional time when the location is signalized. When vehicles are from different routes, leading $({}^{\Sigma Id}LEAD)$ and waiting $({}^{\Sigma Id}WAIT)$ vehicles both increase ${}^{\Sigma T}\widehat{D}WL$, but the increase for waiting vehicles is much larger. Tailing $({}^{\Sigma Id}TAIL)$ and jumping $({}^{\Sigma Id}JUMP)$ both decrease total door open duration. Overall, interactions from the same route $({}^{\Sigma Is}INT)$ decrease ${}^{\Sigma T}\widehat{D}WL$ by a few seconds. The effect may not apply to that specific vehicle, but has an effect on the system as a whole.

Like with ELD, these models may be specific to different times of day and locations. Table 5-10 shows the model explanatory power after separating by location and date and times. At first look, it appears that some models are extremely good are predicting overall variability. During the PM-Peak for segments with stops on the transit mall, 93% of the variability in the data can be accounted for. Segments with transit centers also appear to be improved; yet much of the improvement may be the inclusion of the stops surrounding those transit centers.

Table 5-10 — Aggregated door open duration linear regression using temporally and location specific models. $\forall {}^{\Sigma T}\widehat{D}WL_{t} \in \left\{{}^{\Sigma T}\widehat{D}WL_{t} : \phi \wedge t \in \mathbb{t}\right\}$).

	Hours	**All Segments**	$\boldsymbol{{}^{\Sigma L}MALL_{t} > 1} \wedge {}^{\Sigma L}\boldsymbol{TC_{t}} = 0$	$\boldsymbol{\phi :=}$ $\boldsymbol{{}^{\Sigma L}TC_{t} = 1}$	$\boldsymbol{{}^{\Sigma L}MALL_{t} +}$ $\boldsymbol{{}^{\Sigma L}TC_{t} = 0}$
All Days	*All*	73.27%	89.89%	54.04%	77.11%
Weekday	*All*	73.75%	90.62%	54.58%	76.88%
	AM-Peak	79.06%	92.14%	53.31%	84.54%
	PM-Peak	82.16%	93.08%	69.67%	82.48%
	Peak (AM & PM)	80.73%	92.39%	63.27%	82.68%
	Off-Peak	68.32%	86.82%	49.80%	72.56%
Weekend	*All*	72.37%	80.70%	51.34%	79.16%
	Peak	73.14%	78.53%	53.45%	78.99%
	Off-Peak	67.74%	78.47%	44.94%	76.74%

Like previously discussed, the model’s usefulness is both related to the explanatory power and to the number of segments, events, time, or distances represented by each. Scaling each model to the percent of total segments, total door open time, or total distance traveled, using multiple models achieves less than 0.5% improvement. Scaled to total service events, a 0.97% improvement may be found by using three models for transit center segments (i.e. AM peak, PM peak, and off-peak), one model for the transit mall segments at all times of day, and one model for the remaining segments separated by weekdays and weekends. The added complexity of five additional models is likely not useful for less than 1% absolute improvement to system predictive power. Figure 5-19, like all stacked bar graphs for the aggregated data, will scale based on total service events. Overall, using multiple models has limited benefits.

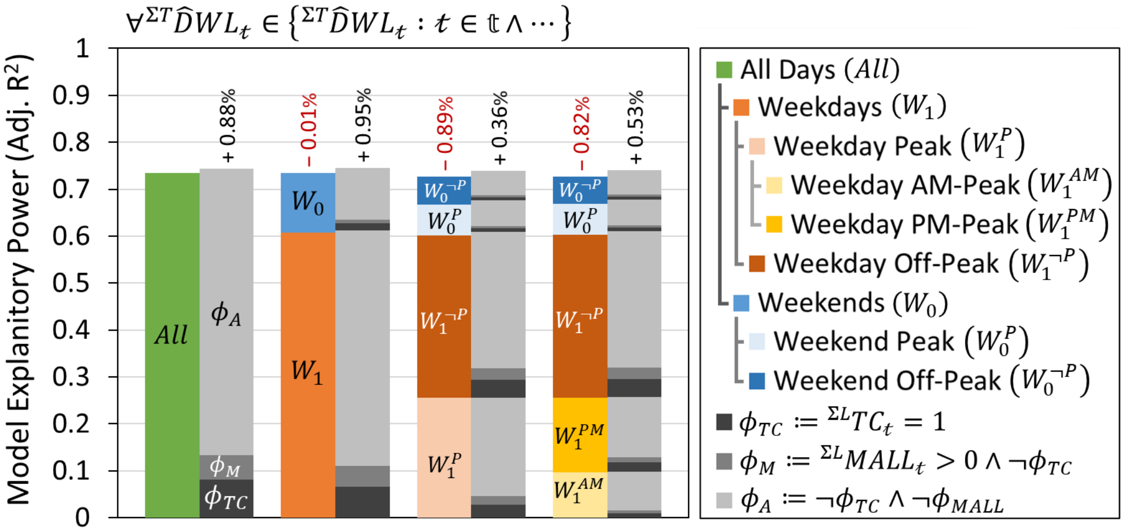

Figure 5-19 — Multiple linear regression models predicting aggregated door open duration (i.e. $\forall^{\Sigma T}\widehat{D}WL_t \in \{^{\Sigma T}\widehat{D}WL_t : \phi \wedge t \in \mathbb{t}\}$). Models are combined and scaled based on the number of bus service events.

Economies of Scale

The economies of scale (Figure 5-20) for the aggregated $^{\Sigma T}\widehat{D}WL_t$ linear regression models are not readily observable for individual stops. While significant, the small negative coefficients do not result in notable time savings for the vast majority of stops.

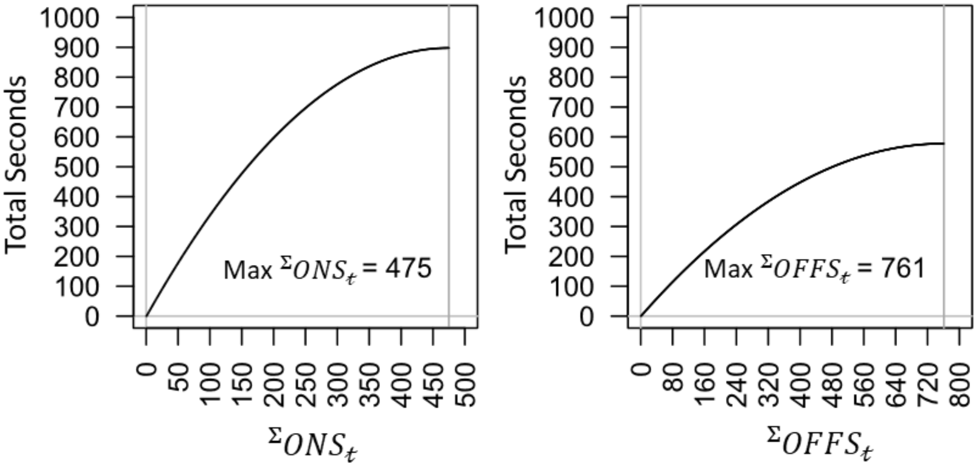

Figure 5-20 — Economies of scale for passenger movement coefficients from $\{^{\Sigma T}\widehat{D}WL_t : t \in \mathbb{t}\}$ aggregated linear regression model in Table 5-9.

As provided in Table 5-9, the squared passenger variables are the square of the sum (i.e. $({}^{\Sigma}\hat{O}NS_{t})^{2}$) over each timepoint segment. Previous research has shown that at the stop level, square terms matter for passenger movements; yet, aggregated, the effect is not as clear. As such, an alternate independent variable could the sum of the square (i.e. ${}^{\Sigma}\hat{O}NS_{t}^{2}$) (rather than the square of the sum). In that form, stop level efficiencies may be observable at an aggregated level. Figure 5-21 plots the sum of the square, as a density, and the square of the sum, as a line, versus total boardings. The vertical lines denote percentiles for the entire data set.

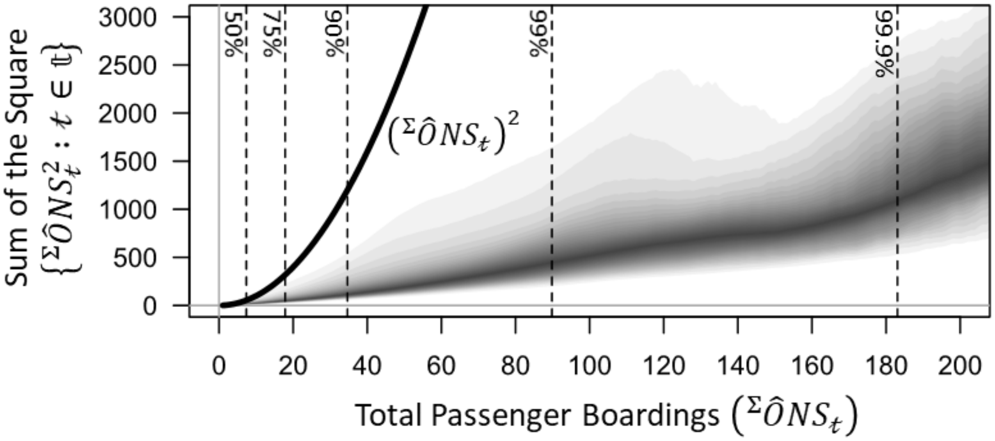

Figure 5-21 — Sum of the square and square of the sum for passenger boardings.

Using this graphic as a reference for the model in Figure 5-20, about 99% of the values will be less than 90; the savings for a segment with 90 passenger boardings are likely to be about 30 seconds. For the sum of the square (i.e. ${}^{\Sigma}\hat{O}NS_{t}^{2}$) , the savings would be higher due to a larger negative coefficient (-0.208), but is not as easily estimated for hypothetical data and does not necessarily make sense. For an example of 90 boardings in a TPS, the sum of the square averages as 537, but there is a large range of potential values; the confidence interval from the 5th to 95th percentile ranges from 236 to 1052, thus representing a range of potential time savings from 49 to 219 seconds. Yet, if the 99th percentile of the sum of the square occurs, the savings are greater than the total boarding time. This relationship is true for most boardings greater than 50.

While the square of the sum (i.e. $({}^{\Sigma}\hat{O}NS_t)^2$) will continue to be used, their overall effect should be evaluated cautiously and consider typical behaviors. For example, passenger boardings per vehicle typically increase as the number of vehicles within a timepoint segment increases (Figure 5-22) and during the peak period, as discussed in section 4.3.2 and further demonstrated using violin plots in Figure C-1 and Figure C-2.

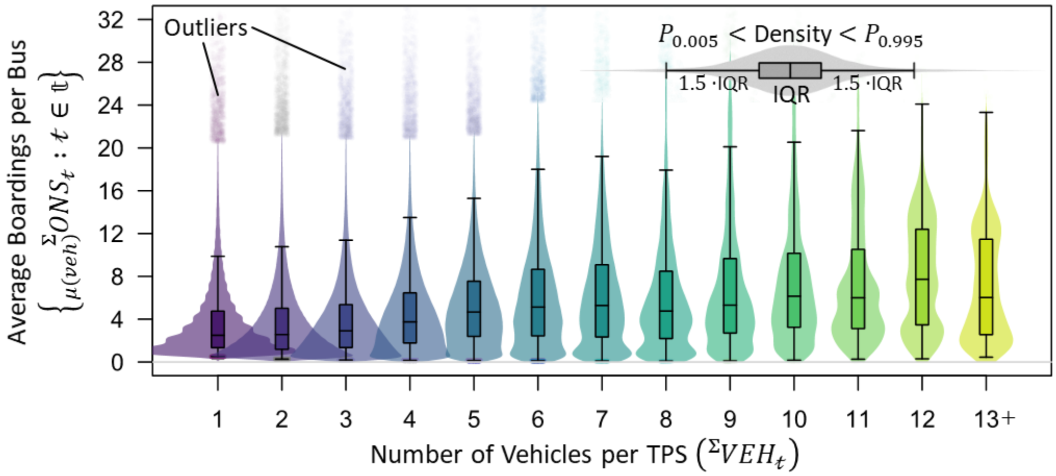

Figure 5-22 — Violin and box-plots for all TPS. Average boardings per vehicle, $\{_{\mu(veh)}{}^{\Sigma}\hat{O}NS_t : t \in \mathbb{t}\}$, given number of vehicles per TPS $({}^{\Sigma}VEH_t)$.

Log-linear Regressions

Log-linear models do not improve the adjusted R-squared or performance at the system level. Table 5-11 shows the adjusted R-squared for these models and Figure 5-23 graphs combined effectiveness of multiple models. A key difference between linear and log-linear models is the reduced variability between locations and times.

Table 5-11 — Aggregated door open duration log-linear regression for temporally and location specific models. $\forall \ln\left[{}^{\Sigma T}\widehat{D}WL_{t}\right] \in \left\{\ln\left[{}^{\Sigma T}\widehat{D}WL_{t}\right] : \phi \wedge t \in \mathfrak{t}\right\}$.

	Hours	All Segments	$\boldsymbol{{}^{\Sigma L}MALL_t > 1}$ $\wedge\, \boldsymbol{{}^{\Sigma L}TC_t} = 0$	$\boldsymbol{\phi :=}$ $\boldsymbol{{}^{\Sigma L}TC_t = 1}$	$\boldsymbol{{}^{\Sigma L}MALL_t +}$ $\boldsymbol{{}^{\Sigma L}TC_t = 0}$
All Days	*All*	61.93%	68.81%	57.58%	62.67%
Weekday	*All*	62.18%	69.85%	57.39%	62.89%
	AM-Peak	63.97%	78.94%	59.47%	64.71%
	PM-Peak	63.85%	74.32%	60.52%	64.25%
	Peak (AM & PM)	63.31%	73.28%	59.05%	63.66%
	Off-Peak	63.43%	69.98%	57.86%	64.27%
Weekend	*All*	63.32%	69.05%	60.55%	64.30%
	Peak	65.10%	72.82%	62.93%	66.20%
	Off-Peak	61.99%	66.52%	58.13%	63.07%

Comparing Figure 5-19 and Figure 5-23 visually, it is clear the log-linear models have reduced explanatory power; but, that both model types have consistent overall performance when multiple models are used for different dates, times and locations.

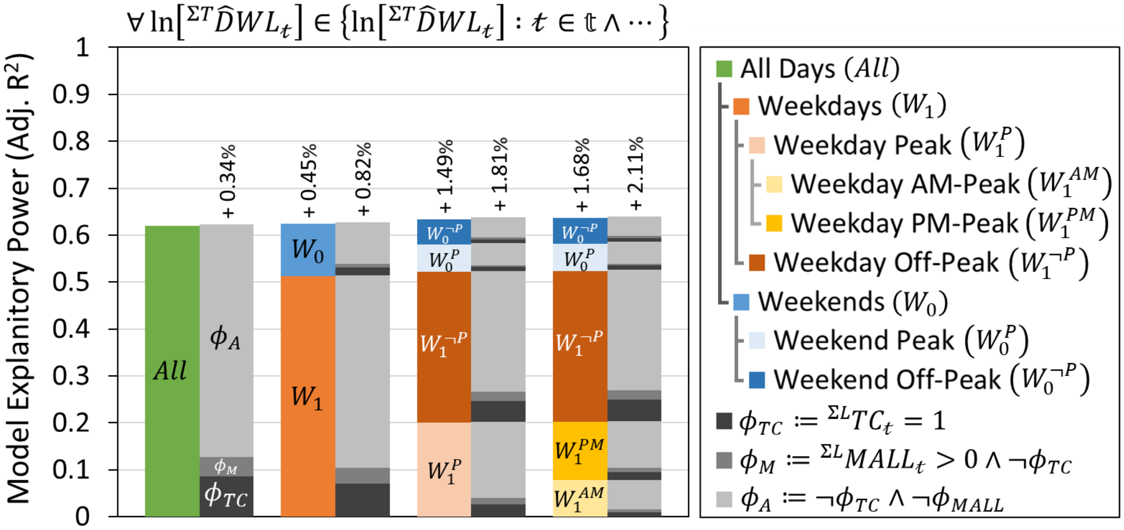

Figure 5-23 — Multiple log-linear regression models predicting aggregated door open duration (i.e. $\forall \ln\left[{}^{\Sigma T}\widehat{D}WL_{t}\right] \in \left\{\ln\left[{}^{\Sigma T}\widehat{D}WL_{t}\right] : \phi \wedge t \in \mathfrak{t}\right\}$). Models are combined and scaled based on the number of bus stops represented.

Composite Variables

Using the results from the discusses collection of models, several alternative models were run where $FREQ$, W_1^{AM}, W_1^{PM}, and W_0^P were replaced by the composite variables created by multiplying by ${}^{\Sigma}VEH_t$ and ${}^{\Sigma}MILES_t$, respectively. This small change to the model formulation ensures that all variables in the aggregated model are a summation of variables that could appear in models at the event level. Stated another way, the composite variables ensure that there are no binary variables in the aggregated models. The aggregated regression model for door open duration, shown in Table 5-12, uses the same set of variables as the model from Table 5-9, with the four exceptions (stated above).

There are minimal differences between the coefficients and contributions of the independent variables, except for the intercept and changed inputs. For the intercept and changed independent variables, the coefficients, contributions, and changes in contribution are given in Table 5-13. The heading "Binary" represents models that do not include composite variables. All of the composite variables have increased contributions over the binary versions, but the adjusted R-squared of the models increased by just 0.12%. Not shown in Table 5-13 is that the contribution of nearly all other variables decreased. The change was small for each variable.

For each aggregated model dependent variable, two alternate formulations were run with composite variables for ${}^{\Sigma}VEH_t$ and ${}^{\Sigma}MILES_t$. Moving forward, the complete models will not be included in the body of this dissertation. Instead only the summary tables will be included in the body, but will reference complete models included in Appendix C. Table 5-14 is the summary for door open duration given ${}^{\Sigma}MILES_t$ composite variables.

Table 5-12 — $\{{}^{\Sigma T}\widehat{D}WL_t : t \in \mathbb{t}\}$ aggregated linear regression model using composite frequency and time variables based on ${}^{\Sigma}VEH_t$.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	-6.04	0.1063		
Number of Vehicles	${}^{\Sigma}VEH$	6.48	0.0559	6.72%	9.16%
Total Passenger Movements	${}^{\Sigma}\hat{O}NS$	3.78	0.0067	13.35%	18.19%
	${}^{\Sigma}\hat{O}FFS$	1.46	0.0067	7.71%	10.51%
	$({}^{\Sigma}\hat{O}NS)^2$	-0.003	0.0000	5.47%	7.46%
	$({}^{\Sigma}\hat{O}FFS)^2$	0.000	0.0000	3.23%	4.41%
	${}^{\Sigma}\hat{L}IFT$	39.22	0.1017	3.47%	4.72%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	15.53	0.0549	0.89%	1.21%
	${}^{\Sigma L}MALL$	10.11	0.0308	1.11%	1.51%
	${}^{\Sigma L}NEAR$	5.39	0.0184	6.00%	8.17%
	${}^{\Sigma L}FAR$	5.21	0.0266	4.14%	5.64%
	${}^{\Sigma L}OPP$	6.30	0.0374	1.60%	2.18%
	${}^{\Sigma L}AT$	7.80	0.0442	0.99%	1.35%
Total Serviced Bus Stop Locations Near Traffic Signals (ΣLs)	${}^{\Sigma Ls}TC$	3.98	0.0906	0.49%	0.67%
	${}^{\Sigma Ls}NEAR$	1.19	0.0234	4.96%	6.76%
	${}^{\Sigma Ls}FAR$	1.45	0.0329	3.61%	4.92%
	${}^{\Sigma Ls}OPP$	0.94	0.0714	0.61%	0.83%
	${}^{\Sigma Ls}AT$	5.15	0.0739	0.77%	1.05%
High-Frequency RTE	${}^{\Sigma}VEH \times FREQ$	1.16	0.0342	4.80%	6.55%
Weekdays	${}^{\Sigma}VEH \times W_1^{AM}$	-7.86	0.0431	0.41%	0.55%
	${}^{\Sigma}VEH \times W_1^{PM}$	-2.99	0.0378	1.13%	1.54%
Weekends	${}^{\Sigma}VEH \times W_0^{P}$	2.43	0.0520	0.23%	0.31%
Total Vehicle Interactions at Bus Stops Between Different RTE (ΣId)	${}^{\Sigma Id}LEAD$	3.39	0.1257	0.78%	1.06%
	${}^{\Sigma Id}WAIT$	25.40	0.3373	0.35%	0.47%
	${}^{\Sigma Id}JUMP$	-1.04	0.3265	0.15%	0.21%
Total Vehicle Interactions at Bus Stops Within the Same RTE (ΣIs)	${}^{\Sigma Is}INT$	-2.21	0.1378	0.41%	0.56%

$n = 4{,}525{,}801$ Adjusted R-Squared = 73.39%
p-value $\ll 0.001$ or all variables

Table 5-13 — Summary table given ${}^{\Sigma}VEH_t$ composite variable for $\{{}^{\Sigma T}\widehat{D}WL_t : t \in \mathbb{t}\}$ aggregated linear models shown in Table 5-9 and Table 5-12

		Coefficient		Contribution		
Variable Type	**Variable**	**Binary**	**× ${}^{\Sigma}VEH$**	**Binary**	**× ${}^{\Sigma}VEH$**	**Change**
Calculated Intercept	Intercept	-3.94	-6.04			
Number of Vehicles	${}^{\Sigma}VEH$	5.49	6.48	7.43%	6.72%	-0.70%
High-Frequency RTE	$FREQ$	2.62	1.16	1.99%	4.80%	+2.81%
Weekdays	W_1^{AM}	-21.80	-7.86	0.10%	0.41%	+0.30%
	W_1^{PM}	-6.52	-2.99	0.22%	1.13%	+0.91%
Weekends	W_0^{P}	7.31	2.43	0.05%	0.23%	+0.18%

Table 5-14 — Summary table given ${}^{\Sigma}MILES_t$ composite variable for $\{{}^{\Sigma T}\widehat{D}WL_t : t \in \mathbb{t}\}$ aggregated linear models shown in Table 5-9 and Table C-1.

		Coefficient		Contribution		
Variable Type	**Variable**	**Binary**	**× ${}^{\Sigma}MILES$**	**Binary**	**× ${}^{\Sigma}MILES$**	**Change**
Calculated Intercept	Intercept	-3.94	-3.81			
Number of Vehicles	${}^{\Sigma}VEH$	5.49	5.07	7.43%	6.78%	-0.65%
High-Frequency RTE	$FREQ$	2.62	1.64	1.99%	5.35%	3.36%
Weekdays	W_1^{AM}	-21.80	-3.44	0.10%	0.37%	0.27%
	W_1^{PM}	-6.52	-0.96	0.22%	1.05%	0.83%
Weekends	W_0^{P}	7.31	1.91	0.05%	0.31%	0.27%

5.4.2. Bus-Bay Stop Duration

Tests, similar to aggregated door open duration and using the same variables, were also conducted for a collection of models to predict total time spent stopped at bus stops. Table 5-15 shows the linear model for ${}^{\Sigma T}\widehat{B}AY_t$. There are similar differences between the ${}^{\Sigma T}\widehat{B}AY_t$ and the ${}^{\Sigma T}\widehat{D}WL_t$ models as were seen between ${}^{T}\widehat{B}AY_i$ and ${}^{T}\widehat{D}WL_i$. Passenger movements now contribute abut two fifths of the R-Squared and location variables have increased to 40%, but it is still evenly divided between the number of services at each location type and the number of those services at locations with traffic signals.

Table 5-15 — $\{{}^{T}\widehat{B}AY_i : \mathcal{t} \in \mathbb{t}\}$ aggregated linear regression model.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	-21.24	0.1777		
Number of Vehicles	${}^{\Sigma}VEH$	20.61	0.0831	9.16%	11.69%
Total Passenger Movements	${}^{\Sigma}\hat{O}NS$	5.55	0.0121	12.65%	16.15%
	${}^{\Sigma}\hat{O}FFS$	2.12	0.0122	9.04%	11.54%
	$({}^{\Sigma}\hat{O}NS)^2$	-0.008	0.0001	5.23%	6.68%
	$({}^{\Sigma}\hat{O}FFS)^2$	-0.001	0.0001	3.89%	4.96%
	${}^{\Sigma}\hat{L}IFT$	43.06	0.1834	2.38%	3.04%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	34.21	0.0991	0.90%	1.15%
	${}^{\Sigma L}MALL$	31.31	0.0593	1.89%	2.42%
	${}^{\Sigma L}NEAR$	14.49	0.0332	8.21%	10.48%
	${}^{\Sigma L}FAR$	12.31	0.0478	4.09%	5.23%
	${}^{\Sigma L}OPP$	11.57	0.0677	1.51%	1.93%
	${}^{\Sigma L}AT$	17.15	0.0798	0.83%	1.06%
Total Serviced Bus Stop Locations Near Traffic Signals (ΣLs)	${}^{\Sigma Ls}TC$	10.90	0.1640	0.61%	0.78%
	${}^{\Sigma Ls}NEAR$	8.73	0.0422	7.35%	9.39%
	${}^{\Sigma Ls}FAR$	1.47	0.0594	3.74%	4.78%
	${}^{\Sigma Ls}OPP$	6.33	0.1290	0.66%	0.85%
	${}^{\Sigma Ls}AT$	2.44	0.1335	0.62%	0.79%
High-Frequency RTE	$FREQ$	-1.62	0.1982	2.08%	2.66%
Weekdays	W_1^{AM}	-19.44	0.2938	0.07%	0.09%
	W_1^{PM}	9.03	0.2510	0.32%	0.40%
Weekends	W_0^{P}	12.84	0.3022	0.03%	0.04%
Total Vehicle Interactions at Bus Stops Between Different RTE (ΣId)	${}^{\Sigma Id}LEAD$	10.56	0.2292	0.97%	1.24%
	${}^{\Sigma Id}TAIL$	10.92	0.2191	0.87%	1.11%
	${}^{\Sigma Id}WAIT$	52.34	0.6119	0.40%	0.51%
	${}^{\Sigma Id}JUMP$	3.68	0.5949	0.21%	0.27%
Total Vehicle Interactions at Bus Stops Within the Same RTE (ΣIs)	${}^{\Sigma Is}INT$	9.78	0.2484	0.61%	0.78%
	$n = 4{,}525{,}801$		Adjusted R-Squared $= 78.33\%$		
	p-value $\ll 0.001$ or all variables				

There are similar differences between the ${}^{\Sigma T}\widehat{B}AY_{\mathcal{t}}$ and the ${}^{\Sigma T}\widehat{D}WL_{\mathcal{t}}$ models as were seen between ${}^{T}\widehat{B}AY_i$ and ${}^{T}\widehat{D}WL_i$. Passenger movements now contribute abut two fifths of the R-Squared and location variables have increased to 40%, but it is still evenly divided between the number of services at each location type and the number of those services at locations with traffic signals.

Breaking down the model by location and times (Table 5-16) and scaling the resulting models (Figure 5-24) results in a maximum absolute increase of 0.46% using at least 6 models. A simple division by weekdays and weekends results in a small improvement of 0.12%. Overall, the small potential benefit of using multiple models is not likely to provide more usefulness than the added complexity of multiple models.

Table 5-16 — Aggregated bus-bay stop duration linear regression using temporally and location specific models. $\forall {}^{\Sigma T}\widehat{B}AY_t \in \left\{{}^{\Sigma T}\widehat{B}AY_t : \phi \wedge t \in \mathbb{t}\right\}$).

	Hours	All Segments	$^{\Sigma L}\boldsymbol{MALL_t > 1}$ $\wedge {}^{\Sigma L}\boldsymbol{TC_t} = 0$	$\boldsymbol{\phi \coloneqq}$ $^{\Sigma L}\boldsymbol{TC_t = 1}$	$^{\Sigma L}\boldsymbol{MALL_t} +$ $^{\Sigma L}\boldsymbol{TC_t = 0}$
All Days	*All*	78.33	94.75	58.40	81.28
Weekday	*All*	78.00	95.03	57.35	80.85
	AM-Peak	78.69	96.04	55.54	81.82
	PM-Peak	83.52	95.83	72.82	82.93
	Peak (AM & PM)	81.61	95.30	65.25	82.30
	Off-Peak	73.79	93.17	51.62	78.79
Weekend	*All*	80.70	90.54	64.65	84.30
	Peak	81.46	89.20	63.59	85.89
	Off-Peak	76.84	89.72	62.22	79.56

However, there are some potential benefits when examining stops on the transit mall without considering other locations. Finally, for the bus-bay durations, the use of log-linear regression (Figure 5-25) for did not improve performance, as compared to the linear models. Aggregation normalizes the data; as such, log-linear models will not be used for other aggregated independent variable modeling.

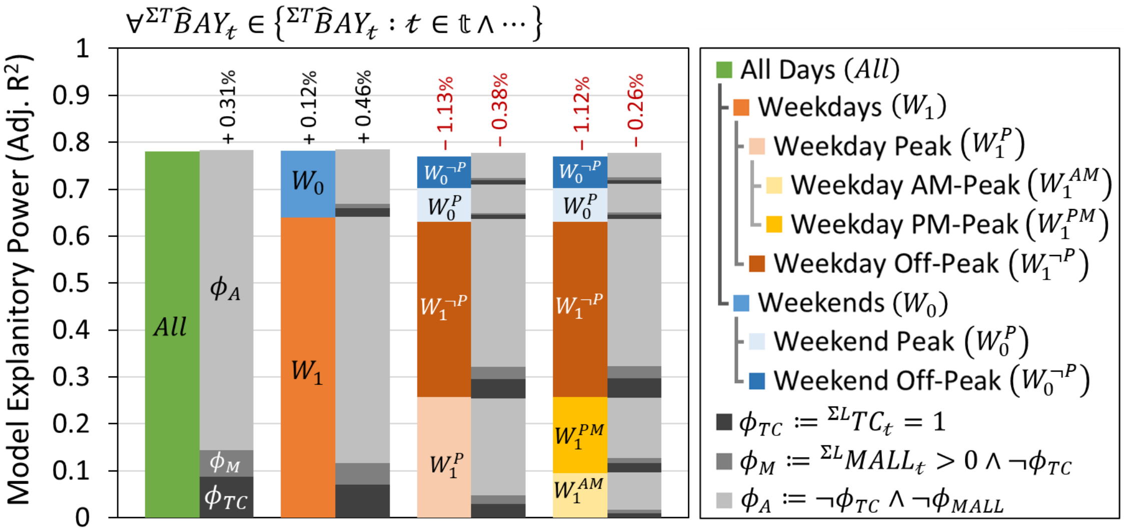

Figure 5-24 — Multiple linear regression models predicting aggregated bus-bay stop duration (i.e. $\forall {}^{\Sigma T}\widehat{B}AY_t \in \{{}^{\Sigma T}\widehat{B}AY_t : \phi \wedge t \in \mathfrak{t}\}$). Models are combined and scaled based on the number of bus stops represented.

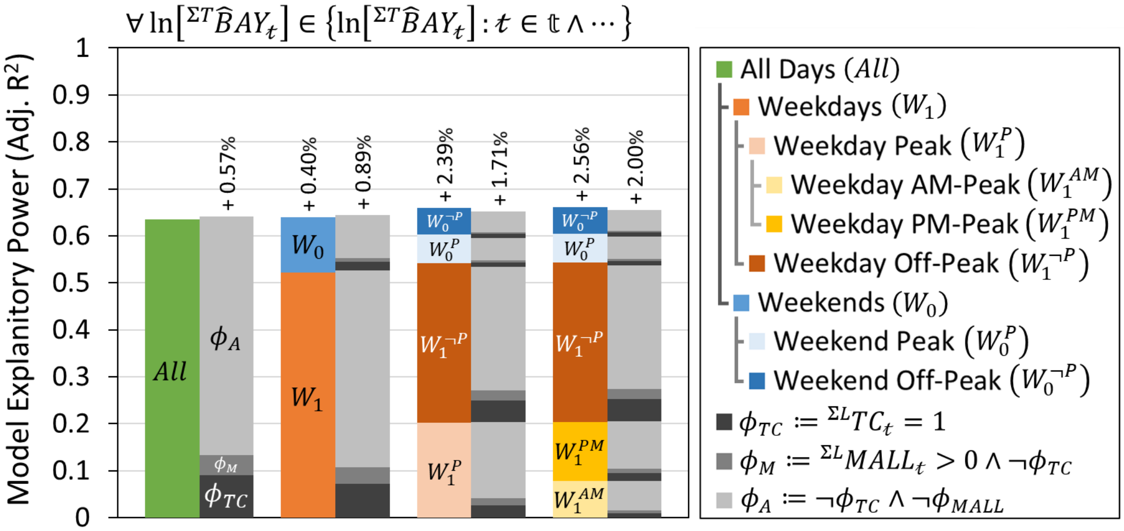

Figure 5-25 — Multiple log-linear regression models predicting aggregated bus-bay stop duration (i.e. $\forall \ln[{}^{\Sigma T}\widehat{B}AY_t] \in \{\ln[{}^{\Sigma T}\widehat{B}AY_t] : \phi \wedge t \in \mathfrak{t}\}$). Models are combined and scaled based on the number of bus stops represented.

The economies of scale for aggregated bus-bay stop duration (Figure 5-26) are similar to aggregated door open duration. As such, the previous discusses explanations are assumed to apply.

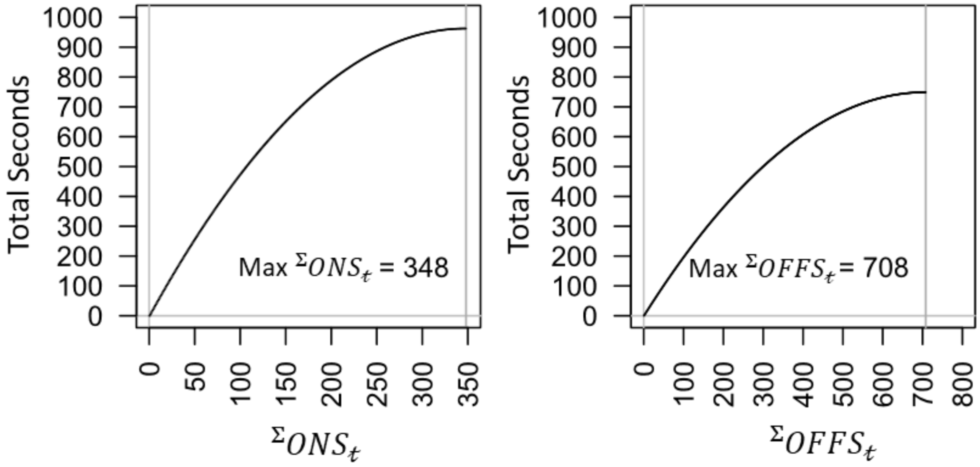

Figure 5-26 — Economies of scale for passenger movement coefficients from $\left\{{}^{\Sigma T}\widehat{B}AY_t : t \in \mathbb{t}\right\}$ aggregated linear regression model in Table 5-15.

Composite Variables

The summaries from Table 5-17 and Table 5-18 show similar differences to the models for door open duration. In both cases, the composite variables have increased contributions to the model explanatory power, while other variables have decreased. Overall, the differences to the adjusted R-squared are just 0.03% in both cases. While ${}^{\Sigma}MILES$ is not an independent variable in either the models for door open duration or bus-bay duration, the associated composite variables are important when comparing and summing model coefficients.

Table 5-17 — Summary table given ${}^{\Sigma}VEH_{t}$ composite variable for $\{{}^{T}\widehat{B}AY_{i} : t \in \mathbb{t}\}$ aggregated linear models shown in Table 5-15 and Table C-2.

		Coefficient		**Contribution**		
Variable Type	**Variable**	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{VEH}$	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{VEH}$	**Change**
Calculated Intercept	Intercept	-21.24	-20.32			
Number of Vehicles	${}^{\Sigma}VEH$	20.61	19.92	9.16%	8.56%	-0.60%
High-Frequency RTE	$FREQ$	-1.62	0.70	2.08%	5.54%	+3.46%
Weekdays	W_{1}^{AM}	-19.44	-6.39	0.07%	0.52%	+0.45%
	W_{1}^{PM}	9.03	3.11	0.32%	1.59%	+1.28%
Weekends	W_{0}^{P}	12.84	4.72	0.03%	0.21%	+0.18%

Table 5-18 — Summary table given ${}^{\Sigma}MILES_{t}$ composite variable for $\{{}^{T}\widehat{B}AY_{i} : t \in \mathbb{t}\}$ aggregated linear models shown in Table 5-15 and Table C-3.

		Coefficient		**Contribution**		
Variable Type	**Variable**	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{MILES}$	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{MILES}$	**Change**
Calculated Intercept	Intercept	-21.24	-19.96			
Number of Vehicles	${}^{\Sigma}VEH$	20.61	20.10	9.16%	8.78%	-0.38%
High-Frequency RTE	$FREQ$	-1.62	0.84	2.08%	5.71%	+3.63%
Weekdays	W_{1}^{AM}	-19.44	-3.38	0.07%	0.49%	+0.42%
	W_{1}^{PM}	9.03	1.28	0.32%	1.41%	+1.09%
Weekends	W_{0}^{P}	12.84	3.11	0.03%	0.27%	+0.24%

5.5. Aggregated Inter-Stop Duration

Stop service durations are only part of understanding transit performance, another primary component is the time spent between bus stops. For the aggregated dataset, these times may be separated between the amount of time moving (i.e. ${}^{\Sigma T}\widetilde{M}OVE_{t}$) and the amount of time stopped (i.e. ${}^{\Sigma T}\widetilde{D}STB_{t}$).

5.5.1. Disturbance Duration

Models for the total disturbance duration $\left({}^{\Sigma T}\widetilde{D}STB_{t}\right)$ are focused on the features of a TPS not defined by passenger movements. Rather, models are based on the number of vehicles, distance traveled, location variables and vehicle interactions. Table 5-19 shows the result of a linear model to predict $^{\Sigma T}\widetilde{D}STB_t$ using TPS for all locations, days and times. The total distance traveled $(^{\Sigma}MILES_t)$ is included as it is relevant to the number of disturbance stops and to moving and total travel times.

Table 5-19 — $\{^{\Sigma T}\widetilde{D}STB_t \,:\, t \in \mathbb{t}\}$ aggregated linear regression model.

Variable Type	**Variable**	**Coefficient**	**Std Error**	**Contrib.**	**Rel-Imp**
Calculated Intercept	Intercept	0.53	0.1639		
Number of Vehicles	$^{\Sigma}VEH$	30.97	0.0775	7.56%	27.10%
Total Distance in Miles	$^{\Sigma}MILES$	5.39	0.0335	3.75%	13.43%
Total Serviced Bus Stop Locations (ΣL)	$^{\Sigma L}TC$	9.17	0.0798	0.68%	2.44%
	$^{\Sigma L}MALL$	2.02	0.0438	0.62%	2.21%
	$^{\Sigma L}NEAR$	-0.30	0.0216	1.11%	3.97%
	$^{\Sigma L}FAR$	0.22	0.0301	1.47%	5.26%
	$^{\Sigma L}OPP$	1.41	0.0546	0.31%	1.11%
	$^{\Sigma L}AT$	2.59	0.0648	0.57%	2.03%
Total Non-Serviced Bus Stop Locations $\binom{\Sigma L}{thru}$	$^{\Sigma L}_{thru}TC$	-1.36	0.2164	0.05%	0.16%
	$^{\Sigma L}_{thru}NEAR$	-1.61	0.0132	0.60%	2.16%
	$^{\Sigma L}_{thru}FAR$	-3.09	0.0233	0.38%	1.37%
	$^{\Sigma L}_{thru}OPP$	-0.94	0.0299	0.15%	0.52%
	$^{\Sigma L}_{thru}AT$	0.41	0.0398	0.21%	0.75%
Total Scheduled Bus Stop Locations near Traffic Signals $\binom{\Sigma Ls}{skd}$	$^{\Sigma Ls}_{skd}TC$	6.23	0.1319	0.43%	1.52%
	$^{\Sigma Ls}_{skd}NEAR$	1.35	0.0198	1.11%	3.96%
	$^{\Sigma Ls}_{skd}FAR$	5.16	0.0298	1.95%	6.97%
	$^{\Sigma Ls}_{skd}OPP$	-3.47	0.0607	0.11%	0.38%
	$^{\Sigma Ls}_{skd}AT$	4.19	0.0730	0.46%	1.66%
High-Frequency RTE	$FREQ$	-28.21	0.1752	0.60%	2.14%
Weekdays	W_1^{AM}	13.57	0.2525	0.18%	0.64%
	W_1^{PM}	63.20	0.2159	2.59%	9.26%
Weekends	W_0^{P}	11.23	0.2596	0.02%	0.07%
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	$^{\Sigma Id}INT$	14.30	0.1014	1.98%	7.08%
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	$^{\Sigma Is}INT$	28.77	0.1923	1.06%	3.78%
	$n = 3{,}684{,}302$				Adjusted R-Squared = 27.91%
	p-value $\ll 0.001$ or all variables				

While the model explanatory power remains relatively low at 27.91%, these two variables account for about 41% of the contribution to the R-squared. On average, each vehicle will add 33 seconds to stopped time between bus stops and each mile of the total distance traveled by all vehicles will add another five.

For locations, three categories of location variables were included, the number of services of each type $({}^{\Sigma L}VAR)$, the number of thru events of each types $\left({}_{thru}^{\Sigma L}VAR\right)$ and the number of scheduled stops $\left({}_{skd}^{\Sigma Ls}var\right)$ near signalized intersections. Except for non-serviced non-served at locations $\left({}_{thru}^{\Sigma L}AT\right)$, non-serviced stops $\left({}_{thru}^{\Sigma L}VAR\right)$ each decrease the duration of disturbance stops. Stops near signalized intersection, including stops on the mall account for nearly 20% of the contribution to the adjusted R-squared and the remaining locations account for another 23%.

Finally, the total number of interactions from vehicles of the same route $({}^{\Sigma Is}INT)$ and from vehicles of different routes $({}^{\Sigma Id}INT)$ were included. Together, they account for 11% of the model's explanatory power. Each ${}^{\Sigma Id}INT$ and ${}^{\Sigma Is}INT$ increase the average disturbance duration by 16 and 30 seconds, respectively. It is notable that interactions from the same route $({}^{\Sigma Is}INT)$ result in an average of twice as much time stopped outside of bus stops as vehicles from different routes $({}^{\Sigma Id}INT)$ within the TPS.

Like with service duration modeling, multiple models were created to represent specific location types separated by time of day, model explanatory power is shown in Table 5-20. Models by time of day show the same location and time-based patterns as the previous models for service durations. Figure 5-27 shows the overall explanatory power when using multiple combinations of models. Using the best combination of (three) models, an overall gain of 0.21% could be achieved. However, most combinations reduced overall effectiveness. The limited model explanatory power for ${}^{\Sigma T}\widetilde{D}STB_t$ is not unexpected. These models do not consider the number of intersections between stops or other factors known to contribute to delays between stops.

Table 5-20 — Aggregated disturbance duration linear regression using temporally and location specific models. $\forall {}^{\Sigma T}\widetilde{D}STB_t \in \{ {}^{\Sigma T}\widetilde{D}STB_t : \phi \wedge t \in \mathbb{t} \}$.

	Hours	**All Segments**	$\boldsymbol{{}^{\Sigma L}MALL_t > 1}$ $\wedge\, \boldsymbol{{}^{\Sigma L}TC_t} = 0$	$\boldsymbol{\phi \coloneqq}$ $\boldsymbol{{}^{\Sigma L}TC_t = 1}$	$\boldsymbol{{}^{\Sigma L}MALL_t +}$ $\boldsymbol{{}^{\Sigma L}TC_t = 0}$
All Days	*All*	27.91%	53.60%	34.93%	24.95%
Weekday	*All*	27.79%	52.92%	34.92%	24.87%
	AM-Peak	26.34%	63.38%	36.54%	23.41%
	PM-Peak	34.30%	50.59%	43.05%	31.69%
	Peak (AM & PM)	30.83%	50.09%	38.96%	28.11%
	Off-Peak	20.70%	46.25%	27.63%	18.17%
Weekend	*All*	25.54%	50.53%	34.63%	22.79%
	Peak	19.02%	52.65%	26.19%	15.38%
	Off-Peak	27.91%	53.60%	34.93%	24.95%

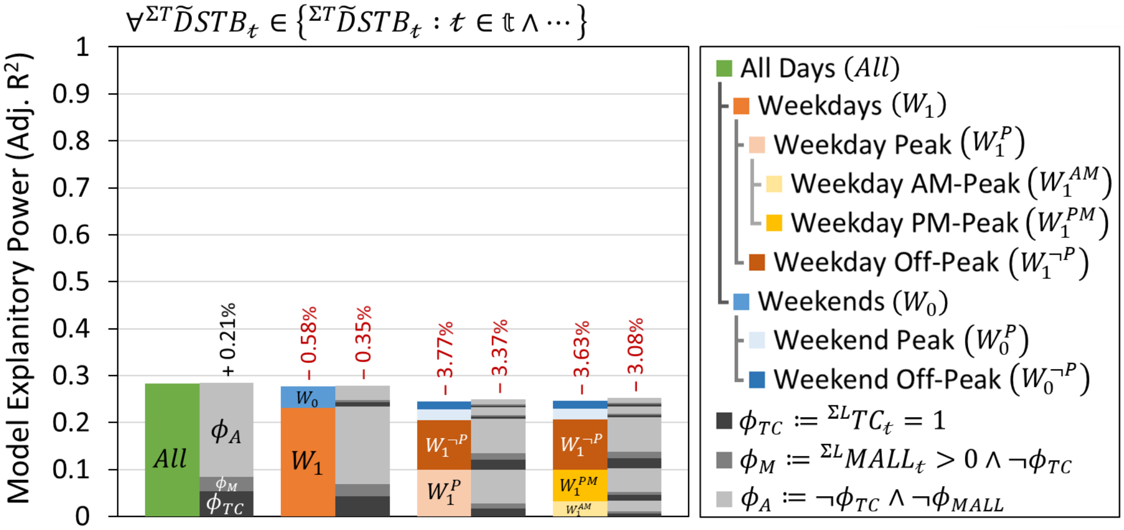

Figure 5-27 — Multiple linear regression models predicting aggregated disturbance duration (i.e. $\forall {}^{\Sigma T}\widetilde{D}STB_t \in \{ {}^{\Sigma T}\widetilde{D}STB_t : \phi \wedge t \in \mathbb{t} \}$). Models are combined and scaled based on the number of bus stops represented.

Composite Variables

Using composite variables, the models saw increases in the adjusted R-squared of 3.1% in both ${}^{\Sigma}VEH_{t}$ (Table 5-21) and the ${}^{\Sigma}MILES_{t}$ (Table 5-22) cases. In both cases, the contribution from the PM weekday peak shows a large increase.

Table 5-21 — Summary table given ${}^{\Sigma}VEH_{t}$ composite variable for $\left\{{}^{\Sigma T}\widetilde{D}STB_{t} : t \in \mathbb{t}\right\}$ aggregated linear models shown in Table 5-19 and Table C-4.

		Coefficient		**Contribution**		
Variable Type	**Variable**	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{VEH}$	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{VEH}$	**Change**
Calculated Intercept	Intercept	0.53	2.43			
Number of Vehicles	${}^{\Sigma}VEH$	30.97	30.97	7.56%	7.10%	-0.46%
Total Distance in Miles	${}^{\Sigma}MILES$	5.39	5.23	3.75%	3.68%	-0.07%
High-Frequency RTE	$FREQ$	-28.21	-9.05	0.60%	1.79%	+1.20%
Weekdays	W_1^{AM}	13.57	3.77	0.18%	0.52%	+0.35%
	W_1^{PM}	63.20	18.86	2.59%	5.11%	+2.53%
Weekends	W_0^{P}	11.23	4.44	0.02%	0.07%	+0.05%

Table 5-22 — Summary table given ${}^{\Sigma}MILES_{t}$ composite variable for $\left\{{}^{\Sigma T}\widetilde{D}STB_{t} : t \in \mathbb{t}\right\}$ aggregated linear models shown in Table 5-19 and Table C-5.

		Coefficient		**Contribution**		
Variable Type	**Variable**	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{MILES}$	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{MILES}$	**Change**
Calculated Intercept	Intercept	0.53	4.57			
Number of Vehicles	${}^{\Sigma}VEH$	30.97	29.69	7.56%	7.81%	+0.25%
Total Distance in Miles	${}^{\Sigma}MILES$	5.39	3.64	3.75%	3.44%	-0.31%
High-Frequency RTE	$FREQ$	-28.21	-4.84	0.60%	1.47%	+0.88%
Weekdays	W_1^{AM}	13.57	2.98	0.18%	0.60%	+0.42%
	W_1^{PM}	63.20	10.50	2.59%	4.73%	+2.15%
Weekends	W_0^{P}	11.23	2.48	0.02%	0.06%	+0.04%

5.5.2. Moving Duration

Models for the moving duration $\left({}^{\Sigma T}\widetilde{M}OVE_{t}\right)$, used the same initial variables as for disturbance duration, but captured just over 90% of the variability in the dependent variable. Table 5-23 shows the linear model for all locations and times.

Table 5-23 — $\left\{{}^{\Sigma T}\widetilde{M}OVE_{t} : t \in \mathbb{t}\right\}$ aggregated linear regression model.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	-19.75	0.1913		
Number of Vehicles	${}^{\Sigma}VEH$	57.08	0.0957	14.82%	16.29%
Total Distance in Miles	${}^{\Sigma}MILES$	91.11	0.0438	27.76%	30.51%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	57.98	0.1008	1.32%	1.45%
	${}^{\Sigma L}MALL$	31.880	0.0585	0.93%	1.02%
	${}^{\Sigma L}NEAR$	16.859	0.0281	7.13%	7.84%
	${}^{\Sigma L}FAR$	13.14	0.0395	5.36%	5.89%
	${}^{\Sigma L}OPP$	7.87	0.0709	2.19%	2.40%
	${}^{\Sigma L}AT$	29.86	0.0836	2.08%	2.29%
Total Non-Serviced Bus Stop Locations $\left({}^{\Sigma L}_{thru}\right)$	${}^{\Sigma L}_{thru}NEAR$	7.99	0.0172	5.51%	6.05%
	${}^{\Sigma L}_{thru}FAR$	3.59	0.0300	3.44%	3.78%
	${}^{\Sigma L}_{thru}OPP$	1.97	0.0381	1.75%	1.93%
	${}^{\Sigma L}_{thru}AT$	3.91	0.0512	1.31%	1.44%
Total Scheduled Bus Stop Locations near Traffic Signals $\left({}^{\Sigma Ls}_{skd}\right)$	${}^{\Sigma Ls}_{skd}TC$	15.33	0.1742	0.88%	0.97%
	${}^{\Sigma Ls}_{skd}NEAR$	-2.08	0.0259	5.26%	5.78%
	${}^{\Sigma Ls}_{skd}FAR$	0.18	0.0390	4.61%	5.07%
	${}^{\Sigma Ls}_{skd}OPP$	0.51	0.0795	0.74%	0.82%
	${}^{\Sigma Ls}_{skd}AT$	4.17	0.0967	0.98%	1.07%
High-Frequency RTE	$FREQ$	-30.28	0.2135	2.07%	2.28%
Weekdays	W_{1}^{AM}	9.78	0.3151	0.22%	0.24%
	W_{1}^{PM}	47.27	0.2699	0.51%	0.56%
Weekends	W_{0}^{P}	7.41	0.3244	0.01%	0.01%
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	${}^{\Sigma Id}INT$	18.83	0.1357	1.35%	1.49%
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	${}^{\Sigma Is}INT$	22.20	0.2587	0.76%	0.83%
	$n = 4{,}524{,}128$			Adjusted R-Squared = 91.00%	
	p-value $\ll 0.001$ or all variables				

As with disturbance durations, ${}^{\Sigma}VEH$ and ${}^{\Sigma}MILES$ account for about a large percentage (45%) of the overall contribution to the adjusted R-squared. Unlike disturbances, the emphasis of different location variables has changed. Service stops of all types account for 23% of the R-squared with ${}^{\Sigma L}NEAR$ and ${}^{\Sigma L}FAR$ capturing two-thirds of that amount. For locations, non-serviced stops add less than serviced stops. Interactions also have a reduced impact on the model. Each interaction adds to the moving time, but only account for 2% of the R-squared. Additionally, the coefficients of the interactions are more similar between same route and different route interactions.

As an additionally consideration, the coefficients ${}^{\Sigma}VEH$ and ${}^{\Sigma}MILES$ also provide a means to briefly check that the model make intuitive sense. If each mile traveled adds 90 seconds, this implies a speed of 40 miles per hour, which is fast for a bus. But, adding the average time per bus of 64 seconds implies an average speed of 23 miles per hour, which would be a reasonable free flow speed for vehicles without intersections or bus stops. Yet, buses encounter both bus stops and intersections. Table 5-24 shows the coefficients from the model in Table 5-23, the average values for model inputs calculated from all TPS, and the calculated moving duration from each variable of the model.

The total moving duration is 755 seconds and corresponds to 4.23 miles traveled. This implies an average moving speed of about 20 miles per hour within all TPS. This is an overestimate, yet, calculating the average speed using averages from all TPS is not how this model should be applied in practice. When calculated directly, the average moving speed is 17.85 mph for all timepoint segments.

Table 5-24 — Average values for ${}^{\Sigma T}\widetilde{MOVE}_{t}$ model inputs for all TPS and calculated average moving duration.

Variable Type	**Variable**	**Coefficient**	**Avg for all TPS**	**Product**
Calculated Intercept	Intercept	-19.75		-19.75
Number of Vehicles	${}^{\Sigma}VEH$	57.08	2.815	160.66
Total Distance in Miles	${}^{\Sigma}MILES$	91.11	4.412	401.99
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	57.98	0.405	23.50
	${}^{\Sigma L}MALL$	31.88	0.392	12.50
	${}^{\Sigma L}NEAR$	16.86	4.762	80.29
	${}^{\Sigma L}FAR$	13.14	2.941	38.64
	${}^{\Sigma L}OPP$	7.87	0.842	6.63
	${}^{\Sigma L}AT$	29.86	0.624	18.64
Total Non-Serviced Bus Stop Locations $\binom{\Sigma L}{thru}$	${}^{\Sigma L}_{thru}NEAR$	7.99	6.827	54.56
	${}^{\Sigma L}_{thru}FAR$	3.59	3.592	12.90
	${}^{\Sigma L}_{thru}OPP$	1.97	1.914	3.77
	${}^{\Sigma L}_{thru}AT$	3.91	1.032	4.03
Total Scheduled Bus Stop Locations near Traffic Signals $\binom{\Sigma Ls}{skd}$	${}^{\Sigma Ls}_{skd}TC$	15.33	0.098	1.51
	${}^{\Sigma Ls}_{skd}NEAR$	-2.08	4.795	-9.98
	${}^{\Sigma Ls}_{skd}FAR$	0.18	3.137	0.55
	${}^{\Sigma Ls}_{skd}OPP$	0.51	0.482	0.24
	${}^{\Sigma Ls}_{skd}AT$	4.17	0.238	0.99
High-Frequency RTE	$FREQ$	-30.28	0.404	-12.24
Weekdays	W_1^{AM}	9.78	0.097	0.95
	W_1^{PM}	47.27	0.146	6.89
Weekends	W_0^{P}	7.41	0.085	0.63
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	${}^{\Sigma Id}INT$	18.83	0.185	3.49
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	${}^{\Sigma Is}INT$	22.20	0.025	0.57
		Average Moving Duration		791.96

Lastly, multiple linear model by time of day and location are shown in Table 5-25 and the scaled models are shown in Figure 5-28. The best case for multiple models to represent the network shows limited improvements over the single model for all TPS. Yet, to achieve that 0.66% gain, 15 models are needed.

Table 5-25 — Aggregated moving duration linear regression using temporally and location specific models. $\forall {}^{\Sigma T}\widetilde{M}OVE_t \in \{{}^{\Sigma T}\widetilde{M}OVE_t : \phi \land t \in \mathfrak{t}\})$.

	Hours	All Segments	${}^{\Sigma L}\boldsymbol{MALL_t} > 1 \land {}^{\Sigma L}\boldsymbol{TC_t} = 0$	$\boldsymbol{\phi} :=$ ${}^{\Sigma L}\boldsymbol{TC_t} = 1$	${}^{\Sigma L}\boldsymbol{MALL_t} +$ ${}^{\Sigma L}\boldsymbol{TC_t} = 0$
All Days	*All*	91.00%	95.23%	91.93%	91.05%
Weekday	*All*	90.96%	94.94%	91.71%	91.01%
	AM-Peak	93.52%	95.70%	95.92%	93.20%
	PM-Peak	90.37%	94.94%	92.69%	90.00%
	Peak (AM & PM)	91.25%	93.92%	93.29%	90.98%
	Off-Peak	90.83%	95.75%	90.83%	91.13%
Weekend	*All*	91.47%	98.06%	93.57%	91.30%
	Peak	91.56%	98.16%	93.17%	91.72%
	Off-Peak	90.56%	98.21%	94.47%	90.02%

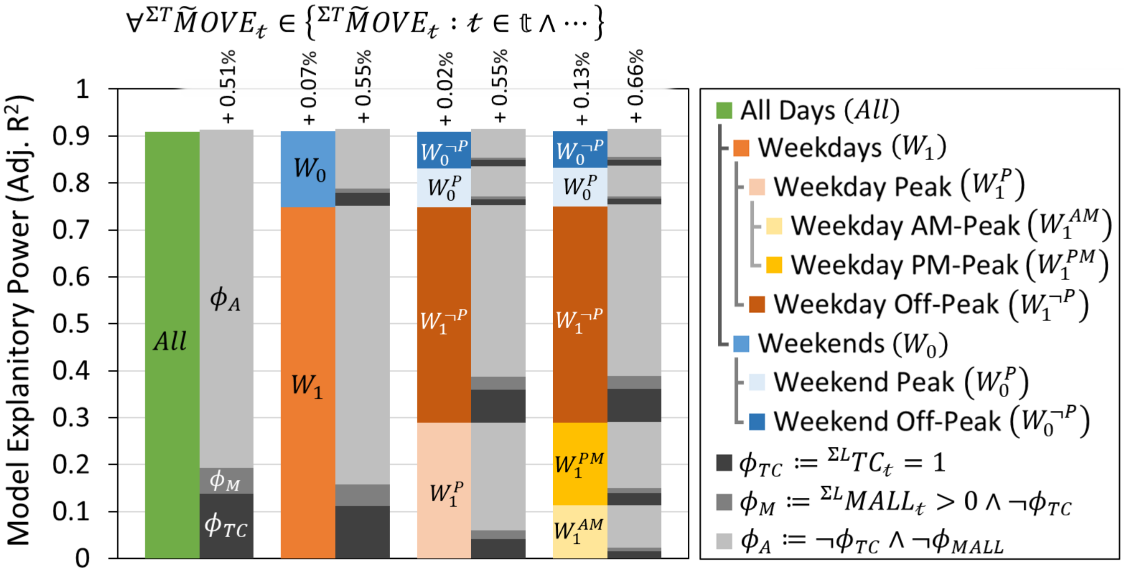

Figure 5-28 — Multiple linear regression models predicting aggregated moving duration (i.e. $\forall {}^{\Sigma T}\widetilde{M}OVE_t \in \{{}^{\Sigma T}\widetilde{M}OVE_t : \phi \land t \in \mathfrak{t}\}$). Models are combined and scaled based on the number of bus stops represented.

Weekend Transit Mall Model

The model for TPS with stops on the transit mall (Table 5-26) appears to capture >98% of variability for weekends. Not all location types are significant and the magnitude and signs of many of the coefficients have changed. ${}^{\Sigma}MILES$, ${}^{\Sigma}VEH$, and ${}^{\Sigma L}MALL$ account for 56% of the total variability. Nearside variables ($^{\Sigma L}NEAR$, $_{thur}^{\Sigma L}NEAR$, and $_{skd}^{\Sigma Ls}NEAR$) account for another 20%. The total distance traveled ($^{\Sigma}MILES$) now has a coefficient of 242, indicating a speed of 15 miles per hour without accounting for any other variables.

Table 5-26 — $\left\{^{\Sigma T}\widetilde{M}OVE_{t} : t \in \left((\mathcal{w} = 1) \cap \mathbb{t}\right) \wedge (^{\Sigma L}MALL_{t} > 0) \wedge (^{\Sigma L}TC_{t} = 0)\right\}$ Aggregated linear regression model for moving time focused on timepoint segments on the transit all and weekends only.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	19.97	0.7208		
Number of Vehicles	$^{\Sigma}VEH$	-1.93	0.5659	13.88%	14.16%
Total Distance in Miles	$^{\Sigma}MILES$	242.07	0.8845	28.37%	28.93%
Total Serviced Bus Stop Locations (ΣL)	$^{\Sigma L}MALL$	6.80	0.2231	12.87%	13.13%
	$^{\Sigma L}NEAR$	-24.10	0.5956	6.46%	6.59%
	$^{\Sigma L}OPP$	-2.97	0.7259	3.16%	3.22%
	$^{\Sigma L}AT$	-2.13	0.7445	0.81%	0.83%
Total Non-Serviced Bus Stop Locations $\binom{\Sigma L}{thru}$	$_{thru}^{\Sigma L}NEAR$	-39.10	0.7110	2.71%	2.76%
	$_{thru}^{\Sigma L}FAR$	-7.06	0.2775	7.30%	7.44%
	$_{thru}^{\Sigma L}OPP$	-7.78	0.2955	4.01%	4.09%
	$_{thru}^{\Sigma L}AT$	-23.19	0.2473	1.64%	1.67%
Total Scheduled Bus Stop Locations near Traffic Signals $\binom{\Sigma Ls}{skd}$	$_{skd}^{\Sigma Ls}NEAR$	-24.17	0.6163	2.98%	3.04%
	$_{skd}^{\Sigma Ls}FAR$	26.95	0.6284	5.80%	5.92%
	$_{skd}^{\Sigma Ls}AT$	14.19	1.0128	0.90%	0.92%
High-Frequency $ROUTE$	$FREQ$	29.99	0.6975	3.69%	3.76%
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	$^{\Sigma Id}INT$	-6.28	0.7983	1.75%	1.79%
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	$^{\Sigma Is}INT$	4.48	0.2606	1.68%	1.71%
	$n = 45{,}221$ p-value $\ll 0.001$ or all variables			Adjusted R-Squared $= 98.06\%$	

This model, while it captures more than 98% of the variability in the data for moving duration, is non-intuitive and only represents 0.89% of the time moving in the network. Despite its good performance (in terms of captured variability), this model has limited practical benefit for understanding the transit system.

Composite Variables

Again, the models using non-binary variables for frequency and peak periods have increased contributions for those inputs, but the overall model explanatory power increases by just 0.06% and 0.19% for ${}^{\Sigma}VEH_t$ (Table 5-27) and ${}^{\Sigma}MILES_t$ (Table 5-28), respectively.

Table 5-27 — Summary table given ${}^{\Sigma}VEH_t$ composite variable for $\left\{{}^{\Sigma T}\widetilde{M}OVE_t : t \in \mathbb{t}\right\}$ aggregated linear models shown in Table 5-23 and Table C-6.

		Coefficient		**Contribution**		
Variable Type	**Variable**	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{VEH}$	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{VEH}$	**Change**
Calculated Intercept	Intercept	-19.75	-21.36			
Number of Vehicles	${}^{\Sigma}VEH$	57.08	57.28	14.82%	13.43%	-1.39%
Total Distance in Miles	${}^{\Sigma}MILES$	91.11	90.89	27.76%	26.73%	-1.03%
High-Frequency RTE	$FREQ$	-30.28	-9.85	2.07%	6.20%	4.13%
Weekdays	W_1^{AM}	9.78	4.99	0.22%	1.06%	0.85%
	W_1^{PM}	47.27	17.52	0.51%	2.21%	1.71%
Weekends	W_0^{P}	7.41	4.87	0.01%	0.19%	0.19%

Table 5-28 — Summary table given ${}^{\Sigma}MILES_t$ composite variable for $\left\{{}^{\Sigma T}\widetilde{M}OVE_t : t \in \mathbb{t}\right\}$ aggregated linear models shown in Table 5-23 and Table C-7.

		Coefficient		**Contribution**		
Variable Type	**Variable**	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{MILES}$	**Binary**	$\times\, {}^{\Sigma}\boldsymbol{MILES}$	**Change**
Calculated Intercept	Intercept	-19.75	-21.69			
Number of Vehicles	${}^{\Sigma}VEH$	57.08	55.09	14.82%	13.18%	-1.64%
Total Distance in Miles	${}^{\Sigma}MILES$	91.11	89.14	27.76%	24.25%	-3.51%
High-Frequency RTE	$FREQ$	-30.28	-8.34	2.07%	8.67%	6.60%
Weekdays	W_1^{AM}	9.78	7.51	0.22%	2.16%	1.94%
	W_1^{PM}	47.27	11.26	0.51%	3.77%	3.26%
Weekends	W_0^{P}	7.41	3.45	0.01%	0.41%	0.41%

5.6. Aggregated Total Travel Time

As a last set of regression models, total travel time (i.e. ${}^{\Sigma T}\tilde{T}RVL_t$) within a TPS was calculated using the sum of ${}^{\Sigma T}\widehat{B}AY_t$, ${}^{\Sigma T}\widetilde{D}STB_t$, and ${}^{\Sigma Tr}\widetilde{M}OVE_t$ as the dependent variable. This model considered passenger activity, locations types for service events, thru events, and traffic signals, and vehicle interaction types for same route and different routes. Table 5-29 shows the linear model for the entire network. Using the same stepwise procedure as previous regressions to create the model, 90% of the variability in total travel time is accounted for.

The magnitude and signs of the coefficients make intuitive sense using the previous models as a guide. Passenger movements account for 27% of the contribution to the R-squared. The economies of scale are still seen for boardings and alightings and their coefficients are the same relative magnitude as the models for ${}^{\Sigma T}\widehat{B}AY_t$. 30% of the contribution to the model explanatory power comes from ${}^{\Sigma}VEH$ and ${}^{\Sigma}MILES$. The coefficients of these two variables are closely related to the sums of the previous model coefficients. For ${}^{\Sigma}VEH$, the sum of the previous coefficients is 118.16, just 8% larger than the ${}^{\Sigma T}\tilde{T}RVL_t$ model at 109.56; for ${}^{\Sigma}MILES$ the sum of the previous coefficients is nearly identical to the coefficient in ${}^{\Sigma T}\tilde{T}RVL_t$ model.

For locations, each serviced stop adds to the total time with transit centers $({}^{\Sigma L}TC)$ and mall stops $({}^{\Sigma L}MALL)$ adding the most. In terms of contributions, nearside location variables collectively $({}^{\Sigma L}NEAR,\ {}^{\Sigma L}_{thru}NEAR,\ {}^{\Sigma Ls}_{skd}NEAR)$ account for 16% of model explanatory power and farside stops collectively $({}^{\Sigma L}FAR,\ {}^{\Sigma L}_{thru}FAR,\ {}^{\Sigma Ls}_{skd}FAR)$ account for 12%. All other locations variables account for less than 10%.

Table 5-29 — $\{{}^{\Sigma T}\tilde{T}RVL_{t} : t \in \mathbb{t}\}$ aggregated linear regression model.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	*Intercept*	-55.82	0.3330		
Number of Vehicles	${}^{\Sigma}VEH$	103.72	0.1647	10.93%	12.11%
Total Distance in Miles	${}^{\Sigma}MILES$	95.95	0.0751	14.22%	15.75%
Total Passenger Movements (Σ)	${}^{\Sigma}\hat{O}NS$	7.00	0.0228	7.77%	8.60%
	${}^{\Sigma}\hat{O}FFS$	4.51	0.0228	7.27%	8.05%
	$({}^{\Sigma}\hat{O}NS)^2$	-0.012	0.0001	3.16%	3.50%
	$({}^{\Sigma}\hat{O}FFS)^2$	-0.006	0.0001	3.00%	3.32%
	${}^{\Sigma}\hat{L}IFT$	39.82	0.3377	1.23%	1.36%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	95.58	0.1838	1.11%	1.23%
	${}^{\Sigma L}MALL$	61.84	0.1097	1.14%	1.27%
	${}^{\Sigma L}NEAR$	29.75	0.0596	6.03%	6.68%
	${}^{\Sigma L}FAR$	21.81	0.0765	4.20%	4.65%
	${}^{\Sigma L}OPP$	19.50	0.1171	1.56%	1.73%
	${}^{\Sigma L}AT$	45.19	0.1505	1.35%	1.49%
Total Non-Serviced Bus Stop Locations $\left({}^{\Sigma L}_{thru}\right)$	${}^{\Sigma L}_{thru}NEAR$	6.70	0.0300	2.91%	3.22%
	${}^{\Sigma L}_{thru}FAR$	1.60	0.0520	1.81%	2.00%
	${}^{\Sigma L}_{thru}OPP$	-0.47	0.0647	0.80%	0.89%
	${}^{\Sigma L}_{thru}AT$	5.84	0.0884	0.62%	0.69%
Total Scheduled Bus Stop Locations near Traffic Signals $\left({}^{\Sigma Ls}_{skd}\right)$	${}^{\Sigma Ls}_{skd}TC$	31.38	0.2978	0.72%	0.80%
	${}^{\Sigma Ls}_{skd}NEAR$	3.43	0.0444	4.41%	4.88%
	${}^{\Sigma Ls}_{skd}FAR$	7.15	0.0664	3.42%	3.79%
	${}^{\Sigma Ls}_{skd}AT$	9.79	0.1667	0.66%	0.73%
High-Frequency $ROUTE$	$FREQ$	-60.02	0.3675	1.54%	1.70%
Weekdays	W_1^{AM}	2.63	0.5414	0.12%	0.14%
	W_1^{PM}	110.02	0.4643	0.51%	0.56%
Weekends	W_0^{P}	30.78	0.5564	0.01%	0.01%
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	${}^{\Sigma Id}LEAD$	39.92	0.4220	3.61%	3.99%
	${}^{\Sigma Id}TAIL$	40.99	0.4036	3.38%	3.74%
	${}^{\Sigma Id}WAIT$	91.90	1.1265	1.28%	1.42%
	${}^{\Sigma Id}JUMP$	56.19	1.0954	0.94%	1.05%
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	${}^{\Sigma Is}INT$	57.12	0.4575	0.60%	0.67%

$n = 4{,}525{,}799$ Adjusted R-Squared $= 90.28\%$

p-value $\ll 0.001$ or all variables

By time of day and location (Table 5-30) similar patterns to the previous models are still observed. Segments with stops on the mall perform better than the system average and segments with transit centers have lower performance. The variations by time-of-day show much smaller variations. Combinations of scaled models are shown in Figure 5-29.

Table 5-30 — Aggregated total service duration linear regression using temporally and location specific models. $\forall^{\Sigma T}\tilde{T}RVL_t \in \{^{\Sigma T}\tilde{T}RVL_t : \phi \wedge t \in \mathfrak{t}\})$.

	Hours	**All Segments**	$^{\Sigma L}\boldsymbol{MALL_t} > \mathbf{1}$ $\wedge\, ^{\Sigma L}\boldsymbol{TC_t} = 0$	$\boldsymbol{\phi} \coloneqq$ $^{\Sigma L}\boldsymbol{TC_t} = \mathbf{1}$	$^{\Sigma L}\boldsymbol{MALL_t} +$ $^{\Sigma L}\boldsymbol{TC_t} = \mathbf{0}$
All Days	*All*	90.12%	95.74%	88.22%	90.33%
Weekday	*All*	89.86%	95.68%	87.82%	90.13%
	AM-Peak	91.22%	96.77%	90.37%	91.25%
	PM-Peak	90.31%	95.74%	90.60%	89.93%
	Peak (AM & PM)	90.21%	95.01%	89.85%	90.03%
	Off-Peak	89.34%	95.93%	86.00%	90.12%
Weekend	*All*	91.26%	97.01%	90.69%	91.69%
	Peak	91.49%	96.84%	90.41%	92.25%
	Off-Peak	90.01%	96.83%	91.12%	90.00%

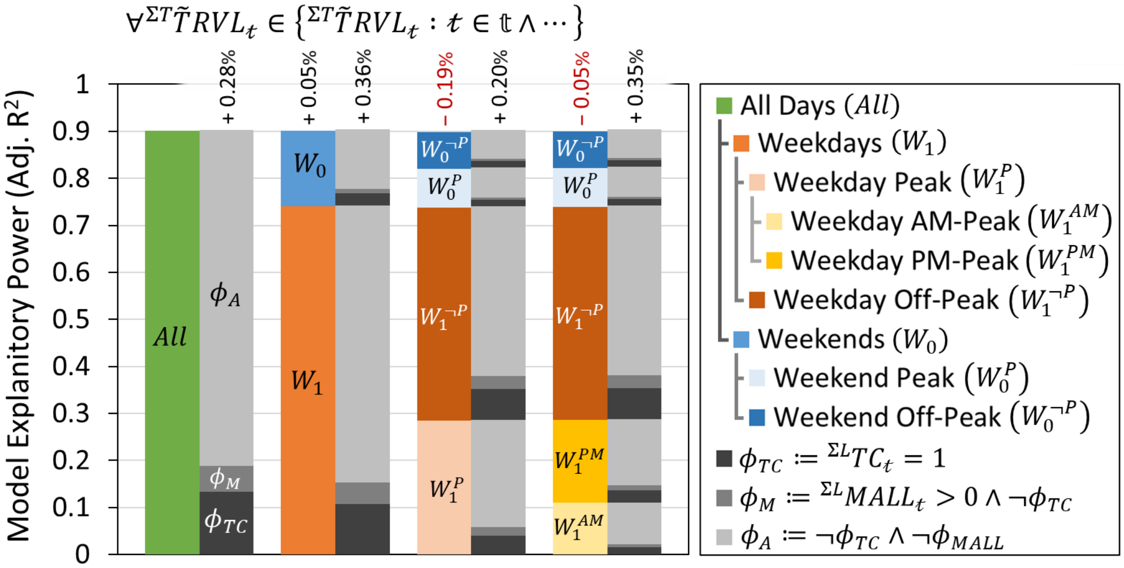

Figure 5-29 — Multiple linear regression models predicting aggregated total service duration (i.e. $\forall^{\Sigma T}\tilde{T}RVL_t \in \{^{\Sigma T}\tilde{T}RVL_t : \phi \wedge t \in \mathfrak{t}\})$. Models are combined and scaled based on the number of bus stops represented.

In summary, the model shown in Table 5-29 used aggregated versions of the key variables identified at the stop event level. The resulting model captures passenger activity, location features, distances traveled, and vehicle interactions to produce a model that accounts for greater than 90% of the variability in the data. This model has intuitive results and does not require a combination of subset models to produce usable information. Furthermore, the combination of multiple models results in just 0.36% improvement in the best case. Given the need for 6 models, in that case, the added complexity is likely not worth the increase.

As total travel time again includes passenger movements, the economies of scale are examined for boardings and alightings (Figure 5-30). Once again, there are some benefits at the timepoint-segment level, but for a typical segment, the total time savings from the square terms will be small.

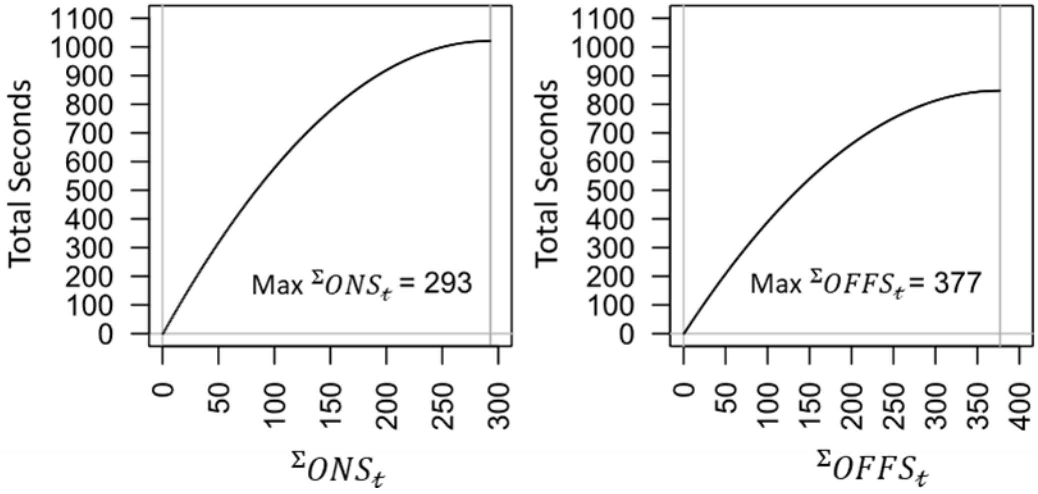

Figure 5-30 — Economies of scale for passenger movement coefficients from $\forall^{\Sigma T}\tilde{T}RVL_t \in \{^{\Sigma T}\tilde{T}RVL_t : t \in \mathbb{t}\}$ linear regression model in Table 5-29.

Composite Variables

Table 5-31 and Table 5-32 show the summary tables for the models with composite variables. Both just a 0.02% increase in the overall explanatory power. But, the increased contribution of the high-frequency routes and peak periods, seen in each of the previous models is again reflected here.

Table 5-31 — Summary table given ${}^{\Sigma}VEH_{t}$ composite variable for $\{{}^{\Sigma T}\tilde{T}RVL_{t} : t \in \mathbb{t}\}$ aggregated linear models shown in Table 5-29 and Table C-8.

		Coefficient		Contribution		
Variable Type	**Variable**	**Binary**	**× $^{\Sigma}VEH$**	**Binary**	**× $^{\Sigma}VEH$**	**Change**
Calculated Intercept	Intercept	-55.82	-57.90			
Number of Vehicles	$^{\Sigma}VEH$	103.72	103.91	10.91%	10.03%	-0.88%
Total Distance in Miles	$^{\Sigma}MILES$	95.95	95.67	14.20%	13.69%	-0.50%
High-Frequency RTE	$FREQ$	-60.02	-18.73	1.53%	4.64%	+3.11%
Weekdays	W_1^{AM}	2.63	2.64	0.12%	0.67%	+0.55%
	W_1^{PM}	110.02	38.43	0.51%	1.93%	+1.42%
Weekends	W_0^{P}	30.78	14.03	0.01%	0.16%	+0.15%

Table 5-32 — Summary table given ${}^{\Sigma}MILES_{t}$ composite variable for $\{{}^{\Sigma T}\tilde{T}RVL_{t} : t \in \mathbb{t}\}$ aggregated linear models shown in Table 5-29 and Table C-9.

		Coefficient		Contribution		
Variable Type	**Variable**	**Binary**	**× $^{\Sigma}MILES$**	**Binary**	**× $^{\Sigma}MILES$**	**Change**
Calculated Intercept	Intercept	-55.82	-55.77			
Number of Vehicles	$^{\Sigma}VEH$	103.72	100.08	10.91%	10.12%	-0.79%
Total Distance in Miles	$^{\Sigma}MILES$	95.95	92.17	14.20%	12.60%	-1.60%
High-Frequency RTE	$FREQ$	-60.02	-12.49	1.53%	5.71%	+4.18%
Weekdays	W_1^{AM}	2.63	7.93	0.12%	1.10%	+0.98%
	W_1^{PM}	110.02	23.09	0.51%	2.53%	+2.03%
Weekends	W_0^{P}	30.78	9.09	0.01%	0.29%	+0.28%

5.6.1. *Sample Sizes*

To evaluate the potential tradeoffs of different sample sizes for the aggregated data, regressions were performed on total travel time using different sample sizes. Figure 5-31 and Figure 5-32 highlight the same type of result; specifically, that a relatively small sample size is needed for consistently significant results, but the coefficients do not reliably converge until much larger sample sizes. For the number of vehicles and total distance traveled (Figure 5-31) that convergence takes place more slowly than for passenger movements (Figure 5-32).

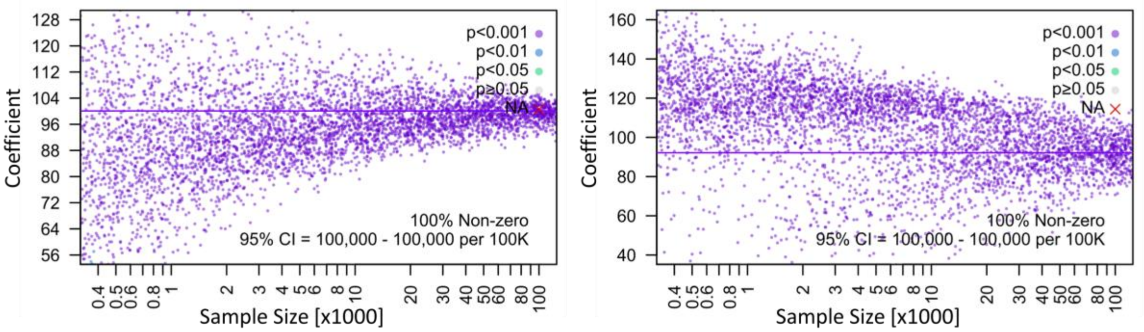

Figure 5-31 — Coefficients for ${}^{\Sigma}VEH$ (left) and ${}^{\Sigma}MILES$ (right) versus sample size (N_{100}). $\left\{{}^{\Sigma T}\widetilde{T}RVL_{t} : t \in \Psi_{\psi_2(m)}(\mathfrak{t})\right\}$ linear model inputs from Table C-9.

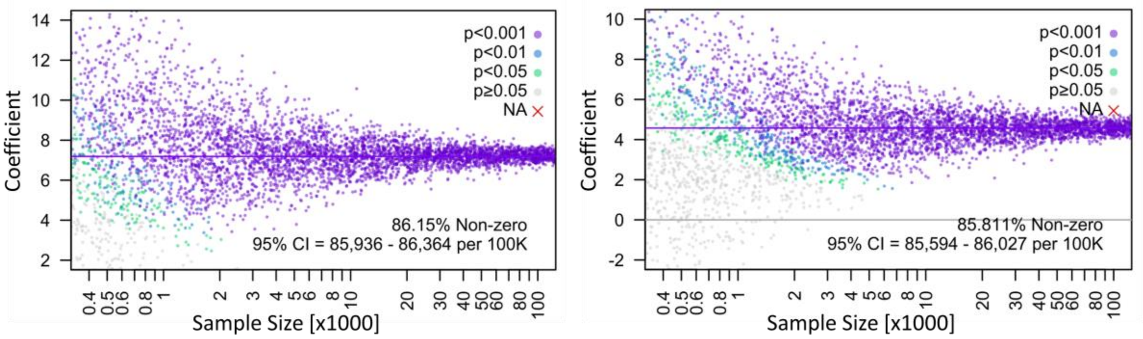

Figure 5-32 — Coefficients for ${}^{\Sigma}ONS$ (left) and ${}^{\Sigma}OFFS$ (right) versus sample size (N_{100}). $\left\{{}^{\Sigma T}\widetilde{T}RVL_{t} : t \in \Psi_{\psi_2(m)}(\mathfrak{t})\right\}$ linear model inputs from Table C-9.

For the number of serviced locations (Figure 5-33), all location types converge at a similar rate; however, $^{\Sigma L}AT$ do not have a consistent sign until $m >$ 5,000 and not consistently significant until $m > 10{,}000$. For samples above this size, these plots highlight that the number of serviced stops of each type are highly significant and likely useful when examined simultaneously.

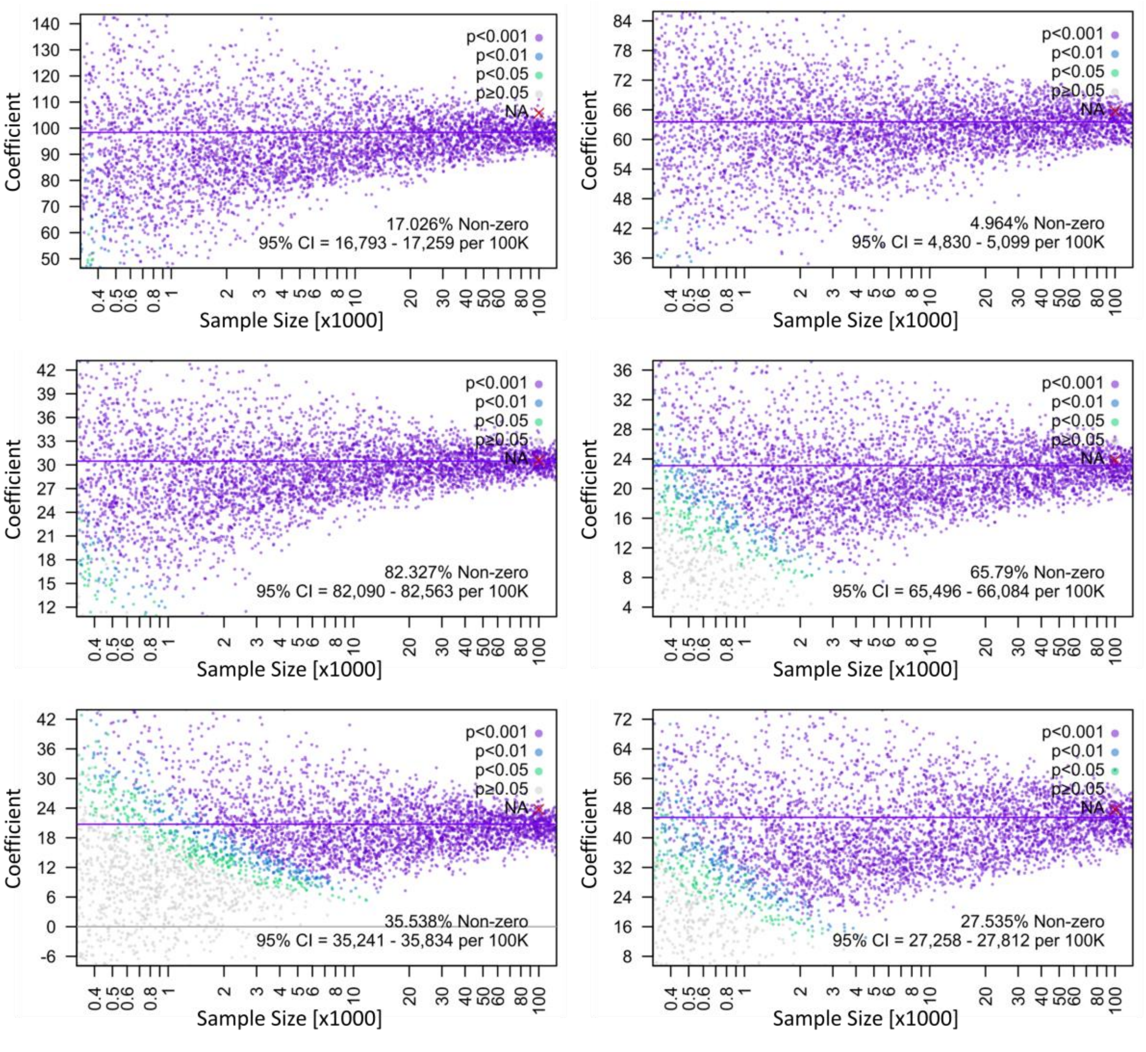

Figure 5-33 — Serviced location coefficients for $^{\Sigma L}TC$ (top-left), $^{\Sigma L}MALL$ (top-right), $^{\Sigma L}NEAR$ (middle-left), $^{\Sigma L}FAR$ (middle-right), $^{\Sigma L}OPP$ (bottom-left), and $^{\Sigma L}AT$ (bottom-right) versus sample size (N_{100}). $\left\{^{\Sigma T}\widetilde{T}RVL_{t} : t \in \Psi_{\psi_2(m)}(\mathbb{t})\right\}$ linear model inputs from Table C-9.

In contrast, different conclusions could be drawn from the non-serviced stop locations. As an example, Figure 5-34 shows that while the coefficients for nearside stops are consistently positive and significant above $m = 8{,}000$, the coefficients for farside stops are not. Given this plot, sample sizes $m < 100{,}000$ have a potential to give results that are inconsistent and dramatically different than models with larger samples.

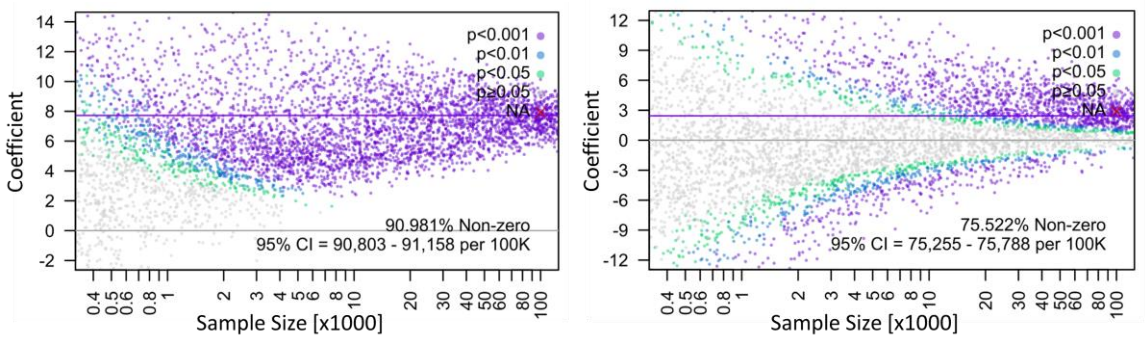

Figure 5-34 — Non-serviced location coefficients for $^{\Sigma L}NEAR$ (left) and $^{\Sigma L}FAR$ (right) versus sample size (N_{100}). $\left\{^{\Sigma T}\widetilde{T}RVL_t : t \in \Psi_{\psi_2(m)}(\mathbb{t})\right\}$ linear model inputs from Table C-9.

Figure 5-35 compares the coefficients from the composite variables for high frequency routes and peak periods. In general, $VAR \times {}^{\Sigma}VEH$ models (right) tend to converge more quickly than $VAR \times {}^{\Sigma}MILES$ models (left). For both composite models, the AM-peak (2nd row) has the potential for inconsistent signs for smaller samples. While the percent of non-zero observations are similar for the AM-peak and weekend peak, the different level of significance may indicate that weekend peak travel is more different from the baseline than the weekday AM-peak.

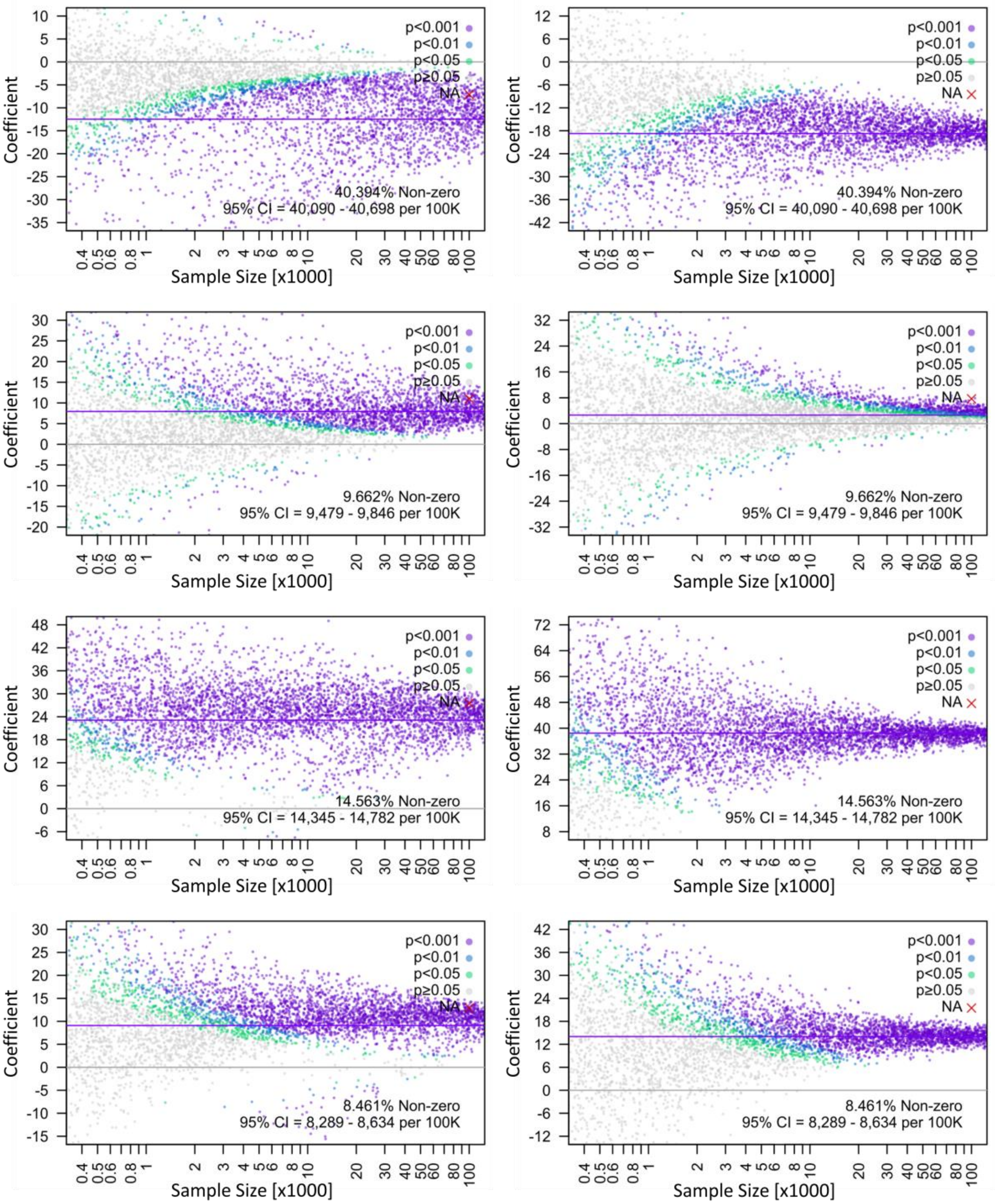

Figure 5-35 — Composite variables comparison for high frequency (top) weekday AM-peak (2nd row), weekday PM-peak (3rd row), and weekend peak (bottom) $VAR \times {}^{\Sigma}MILES$ (left) and $VAR \times {}^{\Sigma}VEH$ (right) versus sample size (N_{100}). $\left\{{}^{\Sigma T}\widetilde{T}RVL_{t} : t \in \Psi_{\psi_{2(m)}}(\mathbb{t})\right\}$ linear model inputs from Table C-9 and Table C-8.

Lastly, the trends for different-route vehicle interactions (Figure 5-36) are similar for each type of interactions and have the potential for consistently signed and significant results at about 10,000 observations. However, jumping interactions continue to have the lowest significance, as was observed across the regression modeling. Given that the percentages of leading/tailing or waiting/jumping interactions are related, the differences in the trends of their coefficients may be attributed to other factors and potentially related back to their level of importance in the final models.

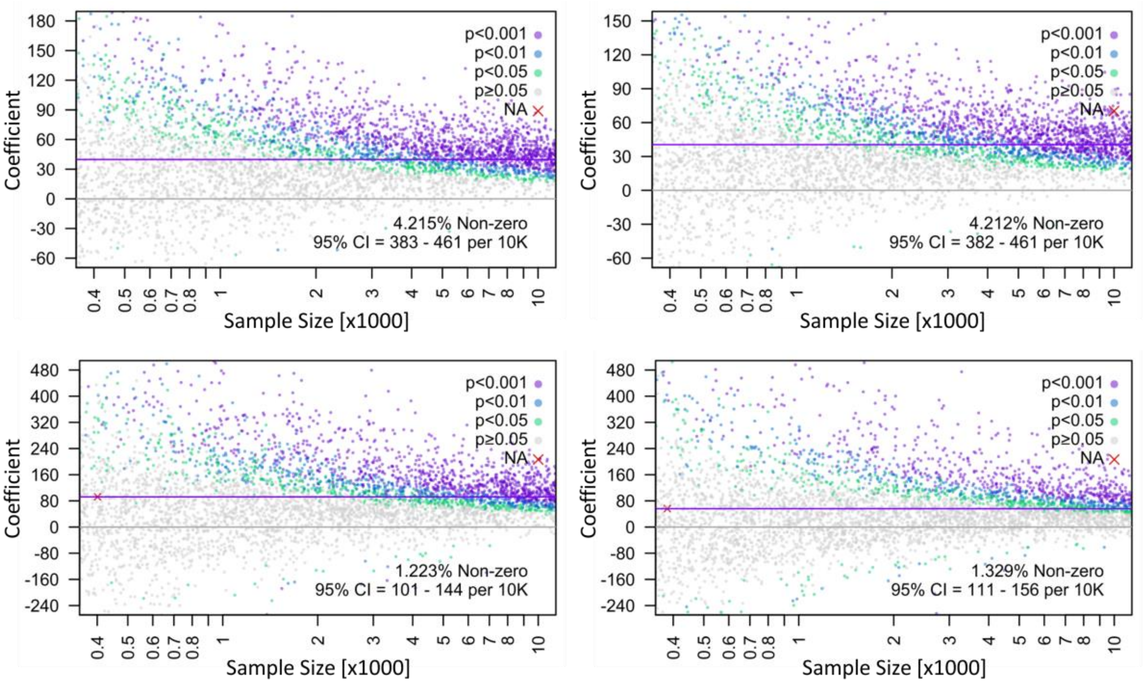

Figure 5-36 — Interaction coefficients for ${}^{\Sigma Id}LEAD$ (top-left), ${}^{\Sigma Id}TAIL$ (top-right), ${}^{\Sigma Id}WAIT$ (bottom-left), and ${}^{\Sigma Id}JUMP$ (bottom-right) versus sample size (N_{100}). $\left\{{}^{\Sigma T}\widetilde{T}RVL_{t} : t \in \Psi_{\psi_2(m)}(\mathfrak{t})\right\}$ linear model inputs from Table C-9.

5.6.2. Comparing Models

The coefficients from the aggregated models for $\left\{{}^{T}\widehat{B}AY_{i} : \mathscr{t} \in \mathbb{t}\right\}$, $\left\{{}^{\Sigma T}\widetilde{D}STB_{\mathscr{t}} : \mathscr{t} \in \mathbb{t}\right\}$ and $\left\{{}^{\Sigma T}\widetilde{M}OVE_{\mathscr{t}} : \mathscr{t} \in \mathbb{t}\right\}$ are summed for the base models (with binary variables) and each of the two composite variable models. These summations are compared to coefficients of each of the total travel time models (Table 5-33). Given some differences in the included variables, coefficients will sometimes be combined. For example, ${}^{\Sigma I}INT + {}^{\Sigma Id}LEAD$ will be used given that each occurrence of ${}^{\Sigma I}INT$ also applies to ${}^{\Sigma Id}LEAD$.

In general, the summed coefficients are similar to the total travel time models. However, there are some exceptions that will be discussed further. First, the squared term of passenger alightings (i.e. $({}^{\Sigma}\hat{O}FFS)^{2}$) was not significant in either of the bus-bay models that used composite variables. As neither moving duration or disturbance duration includes passenger movements, there is no summed coefficient. For the square term of passenger movements, the coefficients are larger for the total travel time model than they are for the summation models. The increase means that benefits will be more noticeable (mathematically) for passenger boardings, given typical usage. Continuing with passenger movements, boardings and alightings have a higher coefficients in the total travel time model. This is likely due these variable's non-inclusion in two of the three other models. Increased passenger movements will likely be correlated with increased moving and disturbance durations, but are not direct contributors. Finally, there is a sign flip for ${}_{thru}^{\Sigma L}OPP$ in two of the three models.

Table 5-33 — Aggregated model comparsions for $\{{}^{\Sigma T}\tilde{T}RVL_{t} : t \in \mathbb{t}\}$ versus the sum of $\{{}^{T}\widehat{B}AY_{i} : t \in \mathbb{t}\}$, $\{{}^{\Sigma T}\widetilde{D}STB_{t} : t \in \mathbb{t}\}$, and $\{{}^{\Sigma T}\widetilde{M}OVE_{t} : t \in \mathbb{t}\}$.

	Binary		**× $^{\Sigma}VEH$**		**× $^{\Sigma}MILES$**	
Variable	**Total**	**Sum**	**Total**	**Sum**	**Total**	**Sum**
Intercept	-55.82	-40.46	-57.90	-39.25	-55.77	-37.08
${}^{\Sigma}VEH$	103.72	108.66	103.91	108.17	100.08	104.88
${}^{\Sigma}MILES$	95.95	96.50	95.67	96.12	92.17	92.78
${}^{\Sigma}\hat{O}NS$	7.00	5.55	7.06	5.54	7.19	5.57
${}^{\Sigma}\hat{O}FFS$	4.51	2.12	4.36	2.04	4.57	2.04
$({}^{\Sigma}\hat{O}NS)^{2}$	-0.012	-0.008	-0.013	-0.008	-0.014	-0.008
$({}^{\Sigma}\hat{O}FFS)^{2}$	-0.006	-0.001	-0.006		-0.009	
${}^{\Sigma}\hat{L}IFT$	39.82	43.06	40.49	42.25	43.03	42.34
${}^{\Sigma L}TC$	95.58	101.36	96.24	101.78	98.50	104.88
${}^{\Sigma L}MALL$	61.84	65.21	62.87	65.82	63.55	66.52
${}^{\Sigma L}NEAR$	29.75	31.05	29.92	31.04	30.47	31.66
${}^{\Sigma L}FAR$	21.81	25.67	22.27	25.99	23.09	26.95
${}^{\Sigma L}OPP$	19.50	20.84	19.56	20.69	20.97	22.33
${}^{\Sigma L}AT$	45.19	49.61	45.12	49.31	45.53	49.88
${}^{\Sigma L}_{thru}NEAR$	6.70	6.38	6.95	6.68	7.71	7.46
${}^{\Sigma L}_{thru}FAR$	1.60	0.50	1.75	0.74	2.43	1.44
${}^{\Sigma L}_{thru}OPP$	-0.47	1.03	-0.34	1.19	0.50	2.15
${}^{\Sigma L}_{thru}AT$	5.84	4.32	5.97	4.62	7.72	6.52
${}^{\Sigma Ls}TC + {}^{\Sigma Ls}_{skd}TC$	31.38	32.46	31.22	32.32	29.26	29.92
${}^{\Sigma Ls}NEAR + {}^{\Sigma Ls}_{skd}NEAR$	3.43	8.01	3.41	7.89	3.66	8.20
${}^{\Sigma Ls}FAR + {}^{\Sigma Ls}_{skd}FAR$	7.15	6.80	6.86	6.48	6.89	6.58
${}^{\Sigma Ls}AT + {}^{\Sigma Ls}_{skd}AT$	9.79	10.80	10.02	10.58	10.86	11.15
$FREQ$	-60.02	-60.11	-18.73	-18.21	-12.49	-12.34
W_{1}^{AM}	2.63	3.92	2.64	2.37	7.93	7.12
W_{1}^{PM}	110.02	119.50	38.43	39.49	23.09	23.05
W_{0}^{P}	30.78	31.48	14.03	14.03	9.09	9.04
${}^{\Sigma I}INT + {}^{\Sigma Id}LEAD$	39.92	43.69	37.51	44.12	39.80	44.29
${}^{\Sigma I}INT + {}^{\Sigma Id}TAIL$	40.99	44.05	38.79	44.31	40.48	44.53
${}^{\Sigma I}INT + {}^{\Sigma Id}WAIT$	91.90	85.47	92.02	85.88	92.25	86.09
${}^{\Sigma I}INT + {}^{\Sigma Id}JUMP$	56.19	36.81	56.34	37.39	56.29	37.80
${}^{\Sigma Is}INT$	57.12	60.74	52.99	55.14	54.19	54.33

Aggregated vs Non-Aggregated

Finally, we may relate the coefficients of the aggregate models back to the non-aggregated models. Not all variables are directly comparable between the models. Given the variable selection loop and model evaluation, signalized locations were not separated by type for the ELD models. As such, the specific signalized locations (i.e, ${}^{\Sigma Ls}VAR$) from the aggregated models are compared to ${}^{L}SIG$. Additionally, the intercepts, ${}^{L}TP$, and ${}^{\Sigma}VEH$ must be evaluated separately, but may be compared. Each TPS has one timepoint location that is served by year vehicle. In the ELD models, that location adds 7.54 seconds. This value is larger than the number of seconds added by each vehicle; yet, some differences may be attributed to transit centers, which are often timepoints, and stops on the downtown transit mall.

For passenger movements, the coefficients are similar between the ELD models and the aggregated models, with the exception of the square term, which has been discussed previously. The aggregated models are summations of the variables included in the ELD, with the exception of high-frequency and peak period binary variables. A more apt comparison is the relationship between the estimated coefficients and the composite variables. For door open duration, the composite variables created by multiplying the total distance traveled (i.e. ${}^{\Sigma}MILES$) results in coefficients that most closely reflect the non-aggregated values. However, these models are comparing stop durations and the total distance traveled is not necessarily a valid estimator for events taking place at bus-bays. That said, total distance traveled is relevant to disturbance times, moving times, and total times, which is why it is considered.

Table 5-34 — Aggregated and non-aggregated door open duration comparsion.

		Aggregated		
Variable	**ELD**	**Binary**	$\times\, {}^{\Sigma}VEH$	$\times\, {}^{\Sigma}MILES$
Intercept	4.91	-3.94	-6.04	-3.81
${}^{L}TP$	7.54			
${}^{\Sigma}VEH$		5.49	6.48	5.07
$\hat{O}NS$ and ${}^{\Sigma}\hat{O}NS$	4.18	3.79	3.78	3.79
$\hat{O}FFS$ and ${}^{\Sigma}\hat{O}FFS$	1.37	1.52	1.46	1.49
$\hat{O}NS^2$ and $({}^{\Sigma}\hat{O}NS)^2$	-0.160	-0.004	-0.003	-0.003
$\hat{O}FFS^2$ and $({}^{\Sigma}\hat{O}FFS)^2$	-0.003	-0.001		-0.001
$\hat{L}IFT$ and ${}^{\Sigma}\hat{L}IFT$	34.69	40.20	39.22	39.52
${}^{L}TC$ and ${}^{\Sigma L}TC$	5.45	15.53	15.53	15.07
${}^{L}MALL$ and ${}^{\Sigma L}MALL$	4.00	10.17	10.11	9.99
${}^{L}AT$ and ${}^{\Sigma L}AT$	1.32	7.59	7.80	7.43
${}^{L}SIG$	1.24			
${}^{\Sigma Ls}TC$		3.93	3.98	4.24
${}^{\Sigma Ls}NEAR$		1.32	1.19	1.23
${}^{\Sigma Ls}FAR$		1.48	1.45	1.58
${}^{\Sigma Ls}OPP$		0.91	0.94	1.30
${}^{\Sigma Ls}AT$		5.52	5.15	5.37
$FREQ$	0.79	2.62	1.16	1.64
W_1^{AM}	-1.66	-21.80	-7.86	-3.44
W_1^{PM}	-0.21	-6.52	-2.99	-0.96
W_0^{P}	0.74	7.31	2.43	1.91
${}^{I}LEAD$ and ${}^{\Sigma Id}LEAD$	3.73	2.59	3.39	2.98
${}^{I}WAIT$ and ${}^{\Sigma Id}WAIT$	16.02	24.73	25.40	25.22
${}^{Is}INT$ and ${}^{\Sigma Is}INT$	-1.12	-2.56	-2.21	-2.51

The similarities that exist for door open duration also exist when comparing bus-bay stop durations at the two levels. In particularly, the passenger movements, and vehicle interaction variables appear to be capturing similar total effects. One exception is the interaction between vehicles of the same route, which experiences a sign flip.

Table 5-35 — Aggregated and non-aggregated bus-bay stop duration comparsion.

		Aggregated		
Variable	**ELD**	**Binary**	$\times\,{}^{\Sigma}\boldsymbol{VEH}$	$\times\,{}^{\Sigma}\boldsymbol{MILES}$
Intercept	15.70	-21.24	-20.32	-19.96
${}^{L}TP$	13.83			
${}^{\Sigma}VEH$		20.61	19.92	20.10
$\hat{O}NS$ and ${}^{\Sigma}\hat{O}NS$	5.04	5.55	5.54	5.57
$\hat{O}FFS$ and ${}^{\Sigma}\hat{O}FFS$	1.71	2.12	2.04	2.04
$\hat{O}NS^2$ and $({}^{\Sigma}\hat{O}NS)^2$	-0.194	-0.008	-0.008	-0.008
$\hat{O}FFS^2$ and $({}^{\Sigma}\hat{O}FFS)^2$	-0.006	-0.001		
$\hat{L}IFT$ and ${}^{\Sigma}\hat{L}IFT$	36.86	43.06	42.25	42.34
${}^{L}TC$ and ${}^{\Sigma L}TC$	6.85	34.21	34.17	33.84
${}^{L}MALL$ and ${}^{\Sigma L}MALL$	10.97	31.31	31.34	31.07
${}^{L}FAR$ and ${}^{\Sigma L}FAR$	-5.20	12.31	12.25	12.02
${}^{L}SIG$	7.82			
${}^{\Sigma Ls}TC$		10.90	10.90	10.89
${}^{\Sigma Ls}NEAR$		8.73	8.56	8.60
${}^{\Sigma Ls}FAR$		1.47	1.50	1.53
${}^{\Sigma Ls}OPP$		6.33	6.45	6.60
${}^{\Sigma Ls}AT$		2.44	2.26	2.09
$FREQ\cdots$	0.79	-1.62	0.70	0.84
$W_1^{AM}\cdots$	-1.23	-19.44	-6.39	-3.38
$W_1^{PM}\cdots$	1.52	9.03	3.11	1.28
$W_0^{P}\cdots$	1.80	12.84	4.72	3.11
${}^{I}LEAD$ and ${}^{\Sigma Id}LEAD$	10.47	10.56	10.98	11.16
${}^{I}TAIL$ and ${}^{\Sigma Id}TAIL$	14.05	10.92	11.18	11.40
${}^{I}WAIT$ and ${}^{\Sigma Id}WAIT$	35.31	52.34	52.74	52.95
${}^{I}JUMP$ and ${}^{\Sigma Id}JUMP$	2.91	3.68	4.26	4.66
${}^{Is}INT$ and ${}^{\Sigma Is}INT$	-1.63	9.78	9.23	9.33

5.7. Conclusion

The service duration models presented in this chapter is focused on two main analysis levels. The first is a largescale microscopic analysis using more traditional event-level data. The second is a mesoscopic analysis using data aggregated at the timepoint-segment level. When directly compared (e.g. $\hat{O}NS$ and ${}^{\Sigma}\hat{O}NS$), the coefficients of the different regression levels are similar, which indicates that the aggregated data is capturing the same types of relationships as the non-aggregated. Yet, the amount of variability captured is very different. Where ELD models may capture 30%-40% of the variability, TPS models capture twice that overall variability.

Because the schedules for fixed route transit systems are generally defined and maintained using timepoints, the TPS data has the potential to be more useful for the planning process by making more accurate predictions of performance over a segment. Previous literature has well established that different stop types have different performance and usage trends; however, for system planning, those differences not the primary focus. Rather, the focus is on on-time performance from timepoint to timepoint.

Regressions at the TPS level are not intended to replace ELD models all together. But, if employed, they can reduce the time and energy needed for system level modeling. For the entire system, modeling with ELD is slower, less accurate, and potentially more expensive due to the data requirements. Yet, for smaller time periods or specific areas, modeling with ELD may still prove useful, but can be guided by the results at the aggregated level.

Lastly, Chapter 5 evaluated the amount of data needed for useful results at both the ELD and TPS levels. In brief, the amount of data needed for consistent and significant results varies variable to variable. However, most variables began showing consistent results by 100,000 observations. Using coefficient versus sample size plots, it may also be possible to evaluate the potential usefulness of a given variable. If results remain inconsistent at 100,000 observations, then such a variable should be given additional scrutiny before relying on its outputs.

CHAPTER 6 — RESULTS: HEADWAYS AND CONGESTION

6.1. Introduction

Regression modeling is a useful way to quantify transit performance. Such models give insights into how different factors influence operations and provide a means to improve scheduling and transit planning. However, even models that can capture a large percentage of variability within the data are not necessarily the correct tool to identify or examine specific areas that require a closer look. Using aggregated data from the timepoint-segment level, headway performance metrics and costs estimates, related to congestion, may be used to identify problematic areas visually and quantitatively. Once identified, microscopic analysis methodologies may be employed in a focused way; thus, reducing the overall computational requirements.

The methodologies and visuals of Chapter 6 are applied broadly to the transit system and to more specific areas. In particular, Route 9, which has been well studied, will be used as a test case. Results that previous research will provide examples for how a mesoscopic analysis can focus research to a specific area and the types of analysis that can be performed with microscopic data sets, if given a reduced scope.

6.2. Headways

As with many variables created using the aggregated methodology, the distribution of headways is dependent on the number of the vehicles in each timepoint-segment, the hour-of-the-day, and the specific location. Figure 6-1 shows the scheduled arrival deviation index (i.e. ${}^{SA}_{idx}H_t$) for segments dependent on the number of vehicles (i.e. ${}^{\Sigma}VEH_t$). Several notable patterns can be observed: first, the violin plots highlight the skew towards smaller scheduled deviation indexes for segments with six or fewer vehicles; second, segments with higher vehicles often have a non-normal and/or bi-modal distribution; and third, as the number of vehicles increases, the scheduled deviation indexes appear to trend towards smaller values. ${}^{SA}_{idx}H_t$, which are calculated based on the service schedule at the first stop of each timepoint-segment, are more consistent than actual arrival times of vehicles.

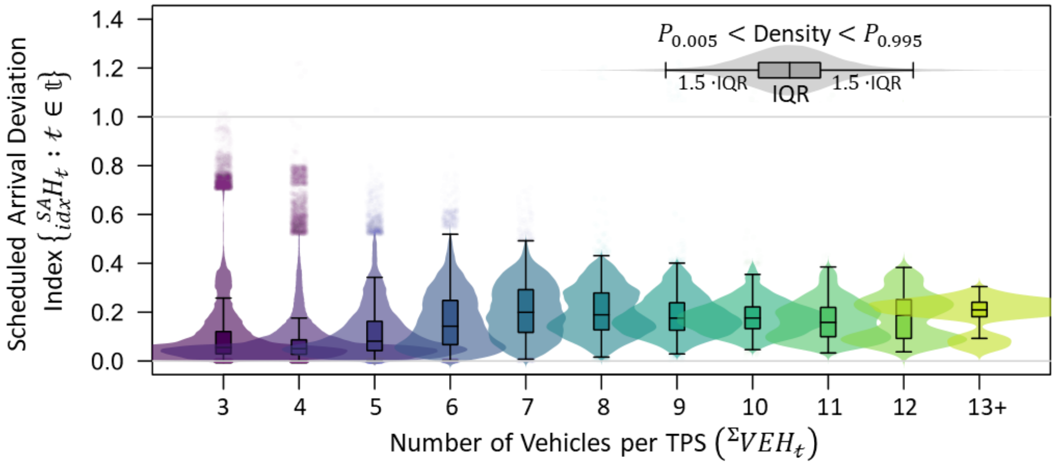

Figure 6-1 — Violin and box-plots for scheduled arrival deviation index (${}^{SA}_{idx}H_t$), given number of vehicles per timepoint-segment (${}^{\Sigma}VEH_t$).

The bimodal distributions are an effect of segments with two groups of scheduled headways. For example, Route 9, inbound to Portland at 7:00 AM, will often have eight scheduled vehicles with uneven headways that range between 5 minutes and 15 minutes. Like Figure 6-1, Figure 6-2 and Figure 6-3 are applicable the entire network and show the observed arrival deviation index and the adjusted deviation index, respectively.

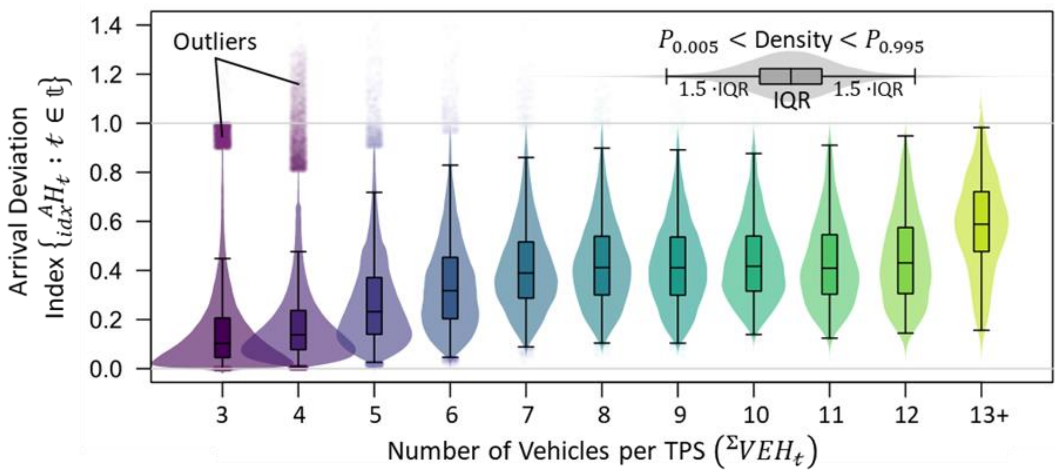

Figure 6-2 — Violin and box-plots for arrival deviation index $({}_{idx}^{A}H_t)$, given number of vehicles per timepoint-segment $({}^{\Sigma}VEH_t)$.

Comparing the scheduled arrival deviation index (i.e. ${}_{idx}^{SA}H_t$) to the "actual" arrival deviation index (i.e. ${}_{idx}^{A}H_t$), there are obvious differences. ${}_{idx}^{A}H_t$ has a more normal distribution and trends upwards as the number of vehicles increase, rather than downwards. Segments with five or fewer vehicles are still skewed towards lower deviation indexes, but also have the highest outliers. When ${}^{\Sigma}VEH_t = 4$, the outliers are at their largest. This is an effect of using three headways in the calculations. If three buses are bunched (e.g. have headways of one minute) and one bus has a headway of 40 minutes, the deviation index will be about 1.2. With more vehicles, the number of such extreme scenarios goes down. Without only three vehicles (i.e. two headways), it is not possible for the deviation index to be above one.

Figure 6-3 is the adjusted deviation index, a ratio of the observed versed scheduled headways. Similar trends to Figure 6-2 remain, but the formulation of the adjusted index allows for negative values. For segments with five or fewer vehicles, a skew towards positive, but smaller, deviation indexes are observed.

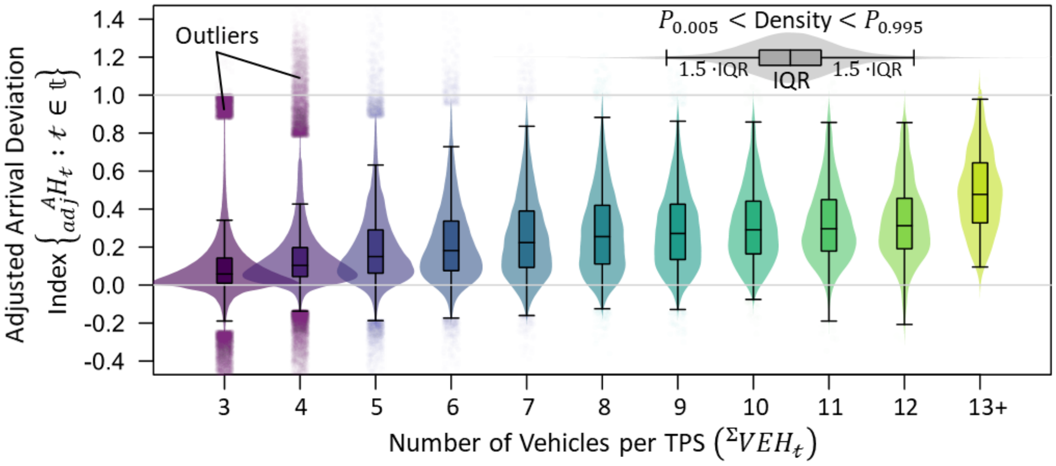

Figure 6-3 — Violin and box-plots for adjusted arrival deviation index (${}_{adj}^{A}H_t$), given number of vehicles per timepoint-segment (${}^{\Sigma}VEH_t$).

Grouping by the number of vehicles provides useful information; but, Figure 4-15 and Figure 4-16 showed that ${}^{\Sigma}VEH_t$ is also time dependent. Figure 6-4 shows the adjusted arrival deviation index, using the time-of-the-day as the x-axis grouping. (Note: a similar color scale is used for time of day plots as for previous plots. There is no relationship.)

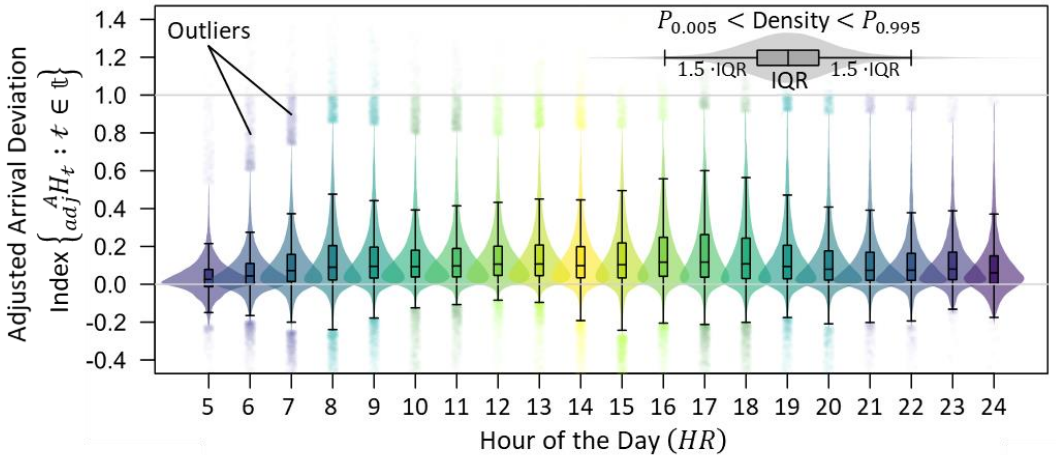

Figure 6-4 — Violin and box-plots for adjusted arrival deviation index (${}_{adj}^{A}H_t$), given hour-of-the-day (HR_t).

For all hours of the day, the adjusted indexes are skewed towards low positive values. The largest ranges are observed during the morning and evening commute hours, which is expected given the number of the trips during those hours. The evening commute appears to have more variability than the morning, which may be related to the number of trips, but also to trip patterns within TriMet’s network.

6.2.1. Inbound vs Outbound

A trip pattern is the set of trips that will be taken by a single vehicle. In a simple case, a bus will complete Route 14 inbound to the city center, then Route 14 outbound. In some ways, this pattern behaves has one long route. More complicated patterns are also used where the Route number changes depending on demand. Typically, trips outbound from the city center are more likely to be continuations of previous inbound trips, due to the distribution of bus depots around the tri-county area. As such, we may expect to see higher deviation indexes for outbound trips than for inbound. Additionally, the differences may be more pronounced during the PM peak, given the demand for trips leaving the downtown city center.

Figure 6-5 and Figure 6-6 divide the network based on their direction (i.e. inbound vs outbound). The results shown in Figure 6-4 lie between these partitions; inbound trips have smaller ranges than outbound trips for nearly all times of day. It is important to note that while $DIR_t = 1$ trips typically mean inbound to city center, that does not apply to all trips. Trips that do not terminate downtown or pass through the urban core, will have different definitions.

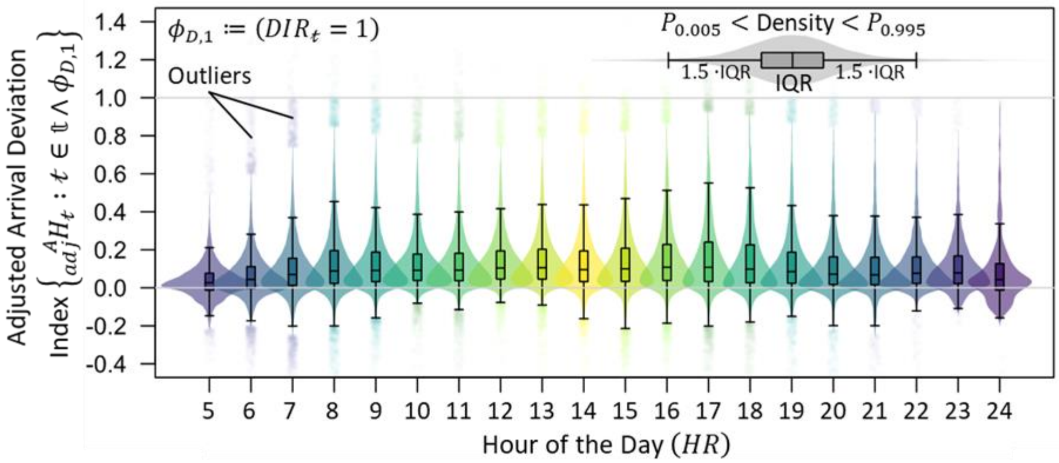

Figure 6-5 — Violin and box-plots for *inbound* transit service. Adjusted arrival deviation index $({}_{adj}^{A}H_t)$, given hour-of-the-day (HR_t).

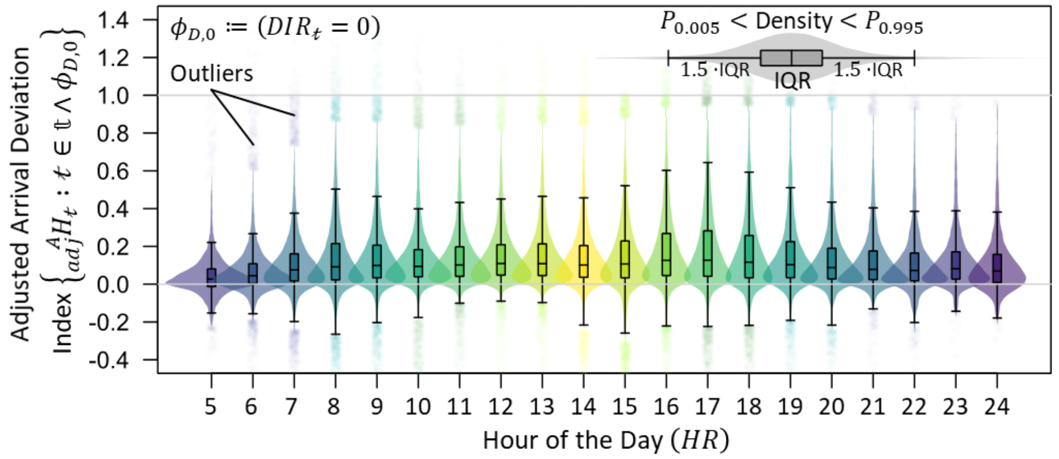

Figure 6-6 — Violin and box-plots for *outbound* transit service. Adjusted arrival deviation index $({}_{adj}^{A}H_t)$, given hour-of-the-day (HR_t).

Figure 6-5 and Figure 6-6 still to all areas of TriMet's network and differences that may be more pronounced for individual route or locations are somewhat hidden. As a case study, additional violin plots will be used to examine Route 9 in detail. But first, attention will be given to how headway performance metrics can be used to compare performance of overlapping service for different routes on the downtown transit mall and bi-directional service from the same route.

6.2.2. Route Comparisons on the Mall

Many routes overlap on the downtown transit mall. Comparing performance can provide insights into behaviors that are route or location specific. For example, the northbound segment with a ${}^{L}TP$ at SW 6th and Alder is the same for routes 8 and 9. Figure 6-7 shows hourly interquartile range (IQR) for headway performance metrics for these two high-frequency routes.

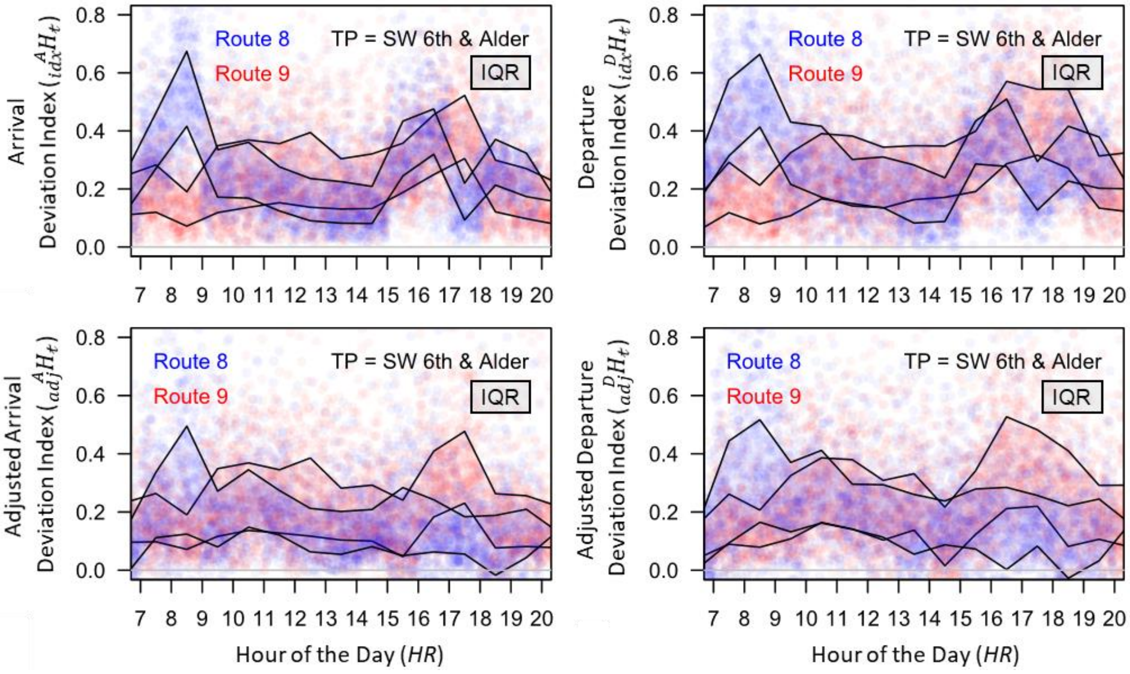

Figure 6-7 — High-frequency overlapping transit service on the downtown transit mall. Arrival (left) and departure (right) headway deviation indexes (top) and adjusted deviation indexes (bottom) for TriMet routes 8 and 9 northbound.

Figure 6-7 uses both the arrival headway entering the timepoint-segment as well as the departure headway leaving. During the AM peak, route 8 has much larger deviation indexes than route 9 at the beginning and end of this segment. When adjusted to their scheduled headways, the confident intervals mostly overlap. Route 8 still shows the potential for higher schedule deviations in the AM peak while route 9 shows potential for higher adjusted deviations during the PM peak. These graphics indicate that there are differences in performance that are route based for the northbound segment. For a southbound segment on the transit mall, Figure 6-8 shows the same performance metrics for the low-frequency routes 17 and 19. Neither of these routes stands out for their AM performance, but route 17 does experience higher deviations during the PM hours.

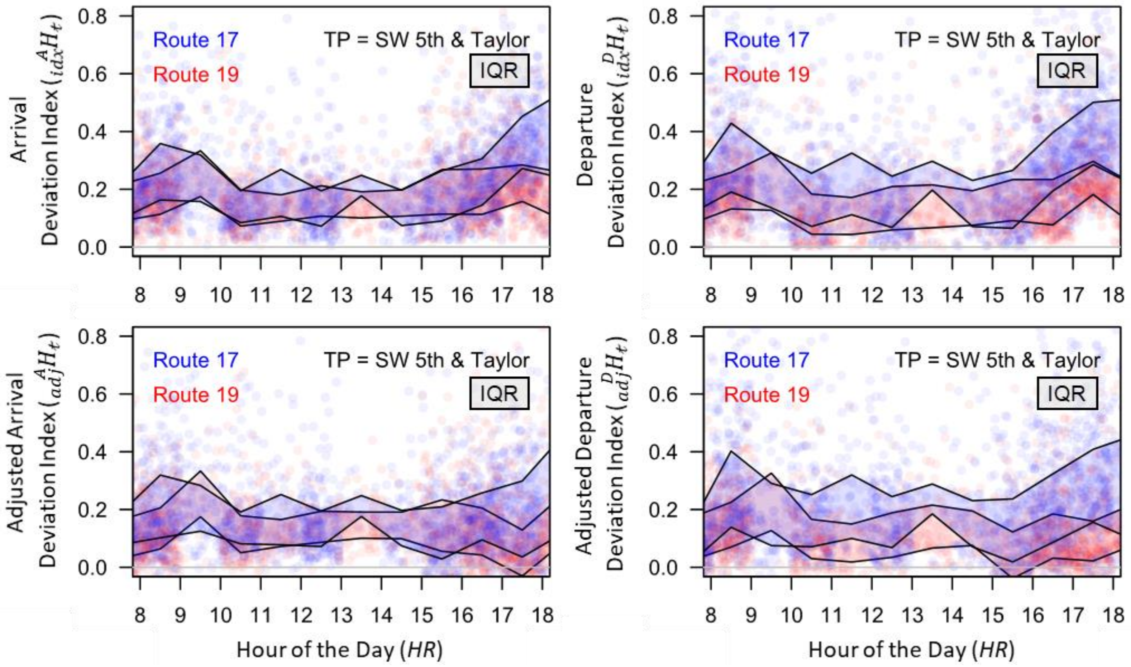

Figure 6-8 — Low-frequency overlapping transit service on the downtown transit mall. Arrival (left) and departure (right) headway deviation indexes (top) and adjusted deviation indexes (bottom) for TriMet routes 17 and 19 southbound.

While the differences in Figure 6-7 are less dramatic than in Figure 6-8, route specific differences can still be observed. In both the northbound and southbound cases, there is also an increased IQR for the departure deviation indexes than for the arrivals. This indicates that headways are disrupted over the evaluated timepoint-segments. The increase in the IQR is larger for route 17 than for 19, which may warrant further investigation. The specific causes of the differences are not the focus of these visuals. Rather, they serve as an investigative tool to determine where further analysis is needed and to prioritize the use of higher resolution and computationally intensive methods.

6.2.3. Direction Comparisons

Figure 6-9, unlike the previous plots, shows two different directions for the same route for overlapping timepoint-segments. Some segments serve parallel trajectories for the different direction (e.g. one-way streets) and are therefore not plotted together. Route 75 was partitioned into 11 timepoint-segments in each direction. Figure 6-9 plots eight in each direction.

The top-left plot represents the first TPS for southbound service and the last TPS for northbound. Conversely, the bottom-right plot represents the last TPS for southbound and first for northbound. This figure shows how the range of adjusted arrival deviation indexes increases along a route, how the two directions show similar but different trends, and how time-of-day impacts performance. By visually comparing adjacent segments, it is possible to identify which segments typically experience the highest and lowest disruptions. For example, the increased IQR for northbound service from TPS 10/11 to TPS 11/11 may indicate some disruptions in that 10th segment. In contrast, the near identical distributions between TPS 6/11 and 7/11 for southbound service may indicate good performance in the 6th segment.

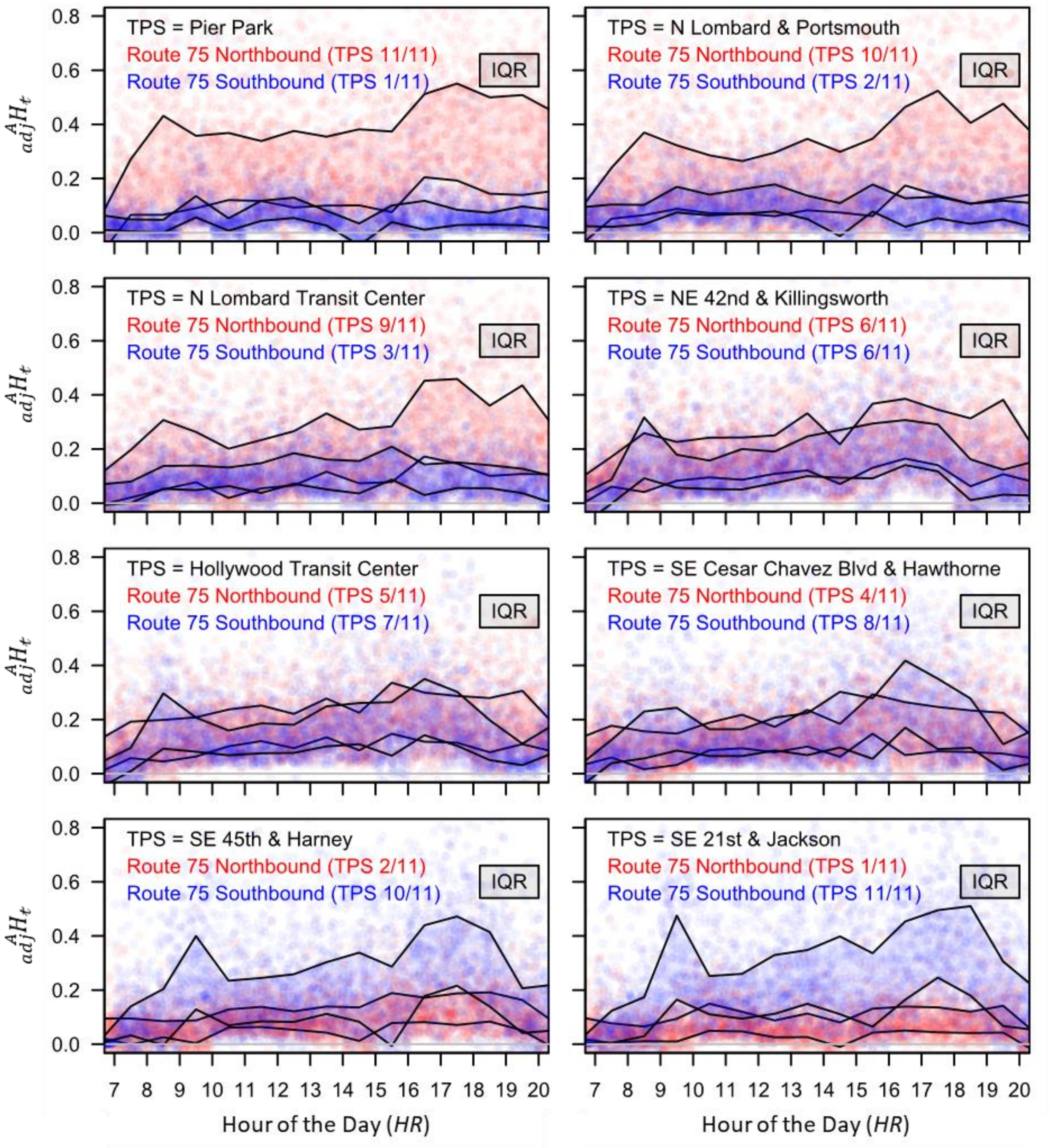

Figure 6-9 — High-frequency northbound and southbound transit service. Adjusted arrival deviation index, $\{{}^{A}_{adj}H_t : t \in \mathfrak{t}\}$, for TriMet's route 75.

Route 75 is a high-frequency route with northbound and southbound service outside of the transit mall. Route 71 (Figure 6-10) is also northbound and southbound outside the mall, but is a low-frequency route. Again, increases are observed along the route in both directions, but a key difference appears based on the time day. The largest increase for northbound service is seen during the morning while much larger increases are observed during PM hours for southbound service. Overall, northbound headway deviation indexes are more consistent across the day than for southbound.

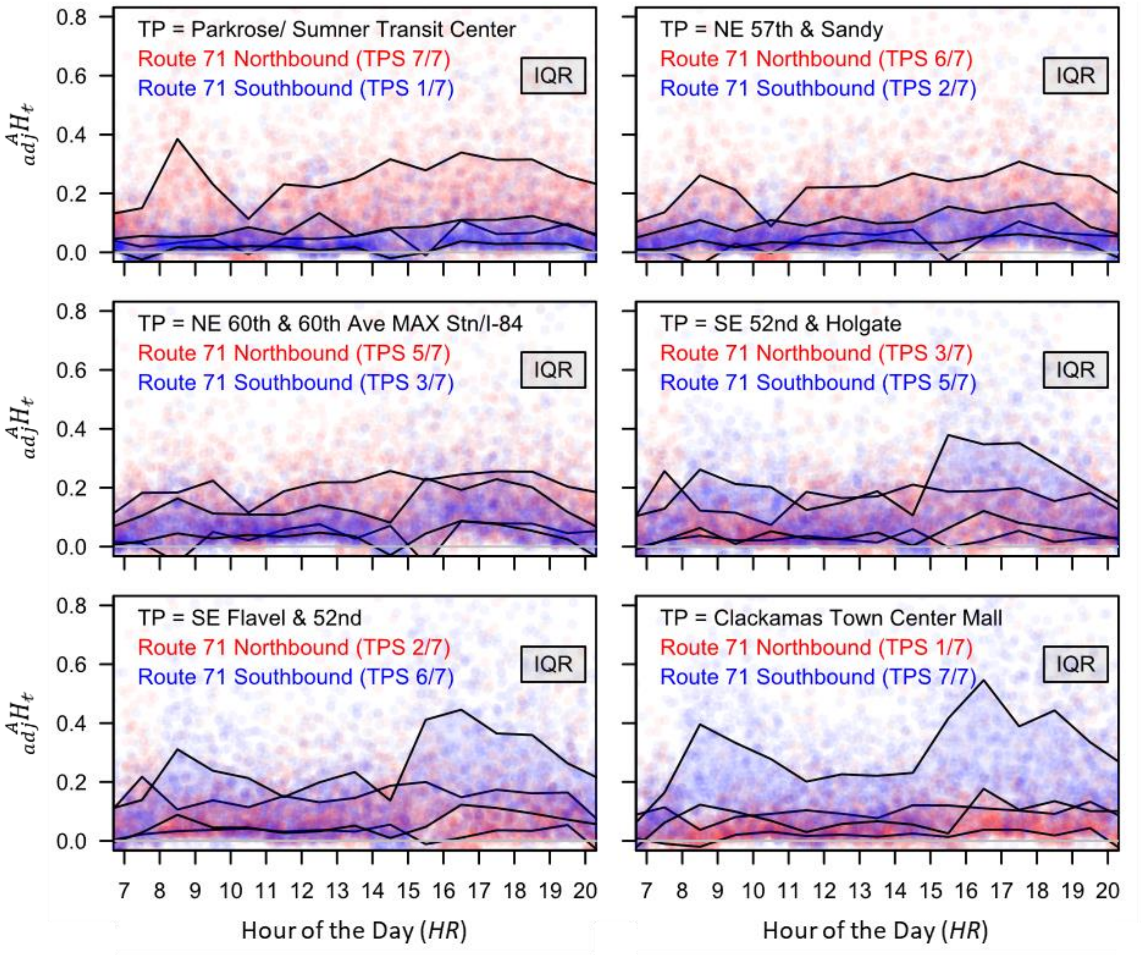

Figure 6-10 — Low-frequency northbound and southbound transit service. Adjusted arrival deviation index, $\{_{adj}^{A}H_{t} : t \in \mathbb{t}\}$, for TriMet's route 71.

Continuing with evaluations outside the transit mall, Figure 6-11 shows both directions of Route 9 along Powell Blvd. Larger deviations during the PM hours are clearly visible for outbound service. Route 9 is a commuter route with high inbound demand during AM hour and high outbound demand during PM hours. Route 9 also terminates in the downtown urban core. As such, there will be expected differences in headway performance. In this case, comparing the overlap does not necessarily provide useful information; yet, the trends along the route (for the same direction) are useful.

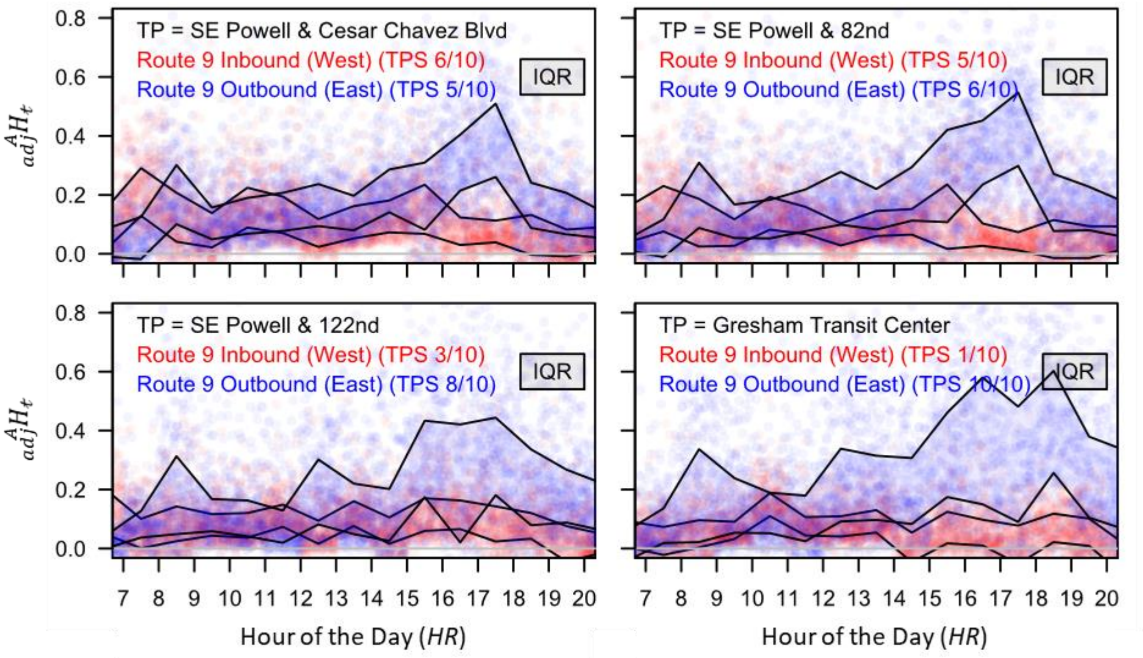

Figure 6-11 — High-frequency inbound (i.e. westbound) and outbound (i.e. eastbound) transit service. Adjusted arrival deviation index, $\{_{adj}^{A}H_{t} : t \in \mathbb{t}\}$, for TriMet's route 9 on SE Powell Blvd.

Route 12 has an interesting map that begins/ends at the Tigard and Parkrose/Sumner Transit Centers, depending on the direction of travel. Route 12 passes through downtown as the middle part of service, but also services many stops to the northeast and southwest of downtown. Figure 6-12 and Figure 6-13 compare the first and last stop of Route 12. In the first set, the timepoint-segments are the same; in the second set, the left plot shows the first TPS and the right shows the last. These two plots highlight that performance in both directions remains similar. Both show increased IQR along their lengths, but neither IQR stand out from the other.

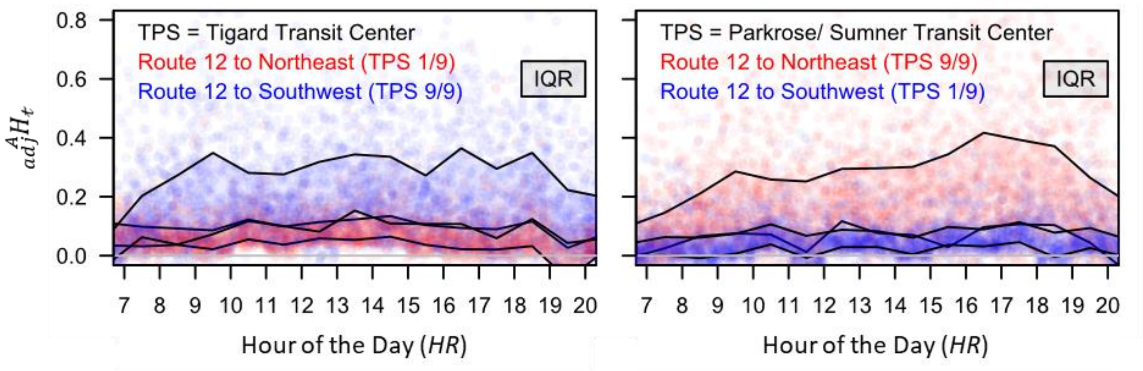

Figure 6-12 — Terminal timepoint-segments for TriMet's high-frequency route 12. Adjusted arrival deviation index, $\left\{{}_{adj}^{A}H_{t} : t \in \mathbb{t}\right\}$, for overlapping segments.

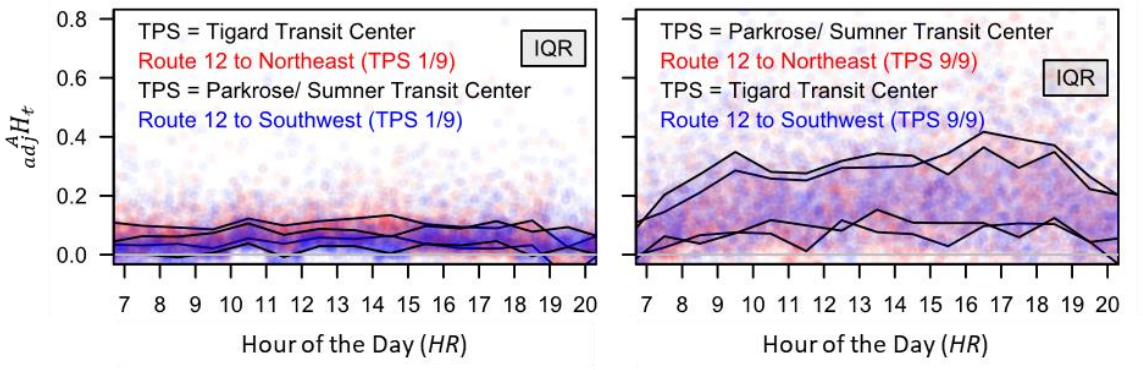

Figure 6-13 — First (left) and last (right) timepoint-segments for TriMet's high-frequency route 12. Adjusted arrival deviation index, $\left\{{}_{adj}^{A}H_{t} : t \in \mathbb{t}\right\}$. Left and right plots do not represent overlapping service.

First and last segment plots, like Figure 6-12 and Figure 6-13 can be a first step in examining if performance along a transit line is notably different in the two directions. The visual analysis can examine specific times to see commuting patterns, scheduled stability, and to potentially identify routes that require further investigation.

6.2.4. Dedicated Bus-Lanes

A potential application of the headway performance measure visuals is to compare effectiveness of transit priority. If data from before and after implementation is used, then the effect on a specific segment may be compared. More generally, segments with priority may examined together, then compared to the system as a whole. While the instances of transit priority within Portland are expanding, there were much more limited at the time the datasets were collection. Figure 6-14 is a map of the Portland Rose Lane Vision showing existing and planned transit priority. As it existed then, the downtown transit mall (Figure 6-15) was the main areas with dedicated transit lanes (PBOT, 2019).

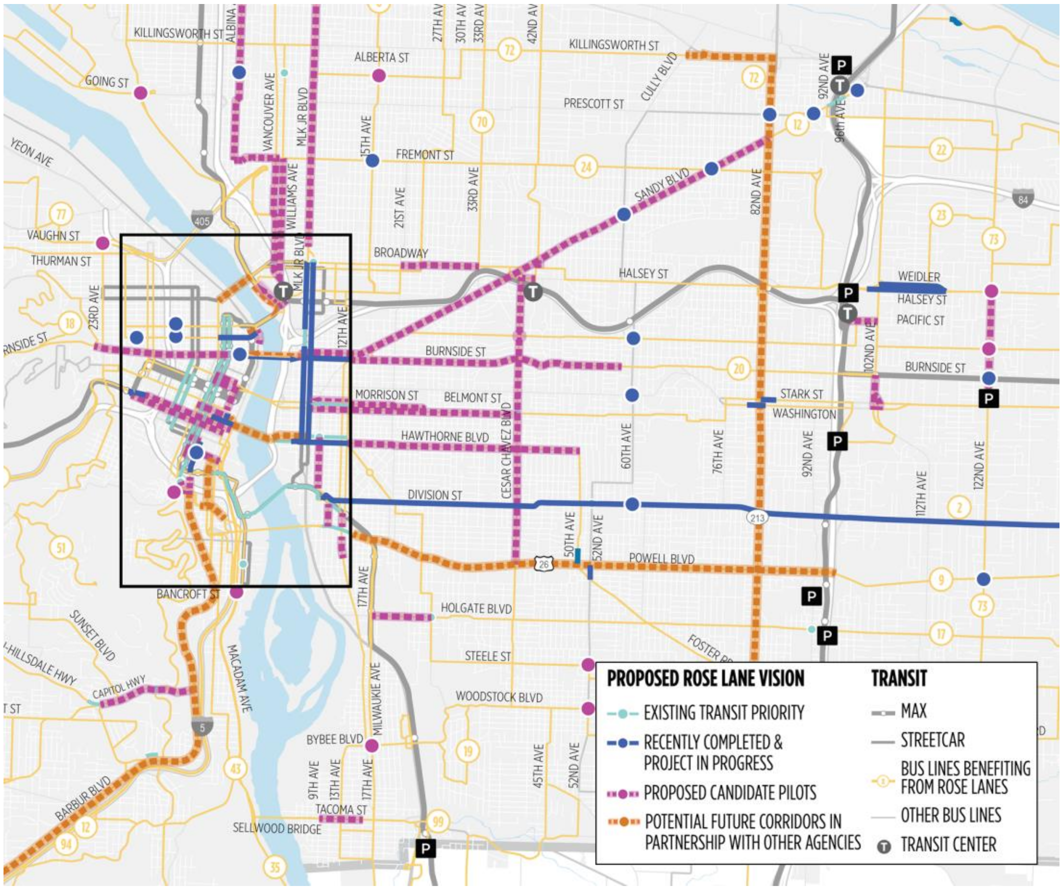

Figure 6-14 — Map of Portland Rose Lane Vision. Existing dedicated transit priority are shown in light blue.

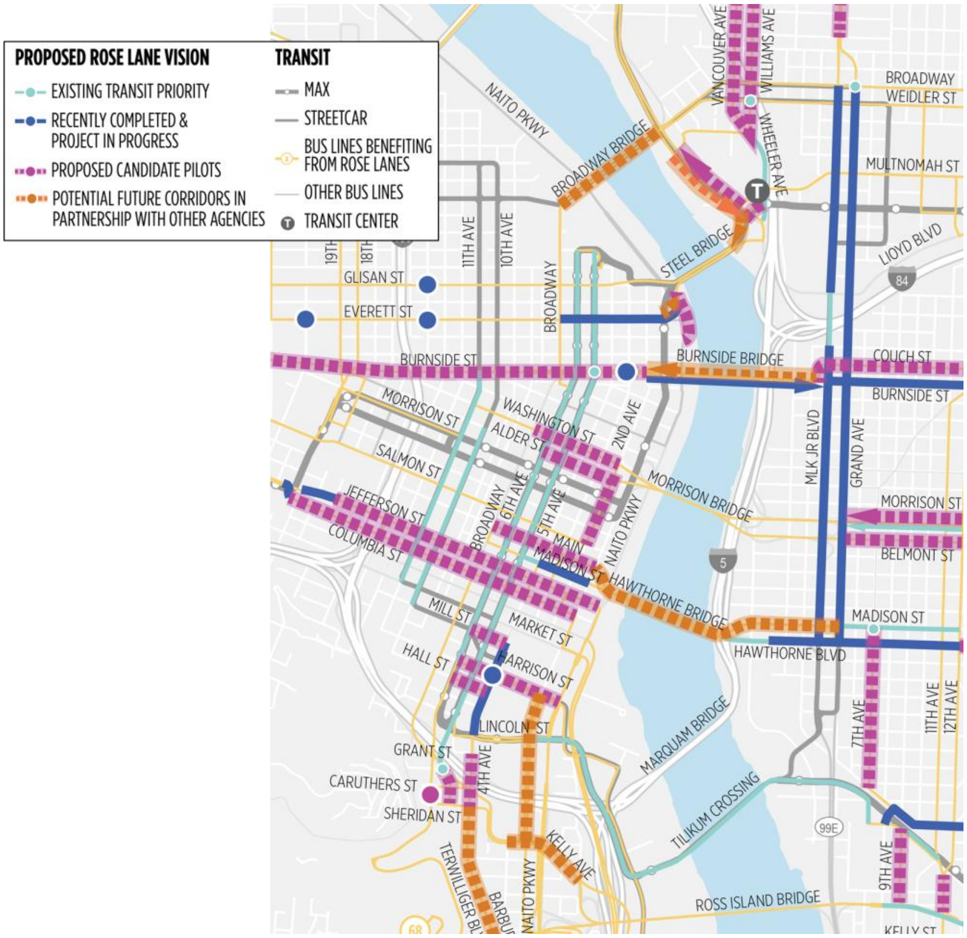

Figure 6-15 — Map of downtown Portland Rose Lane Vision. Zoom-in of black box from Figure 6-14. Existing dedicated transit priority are shown in light blue.

Areas outside of the urban core, such as the Madison Bus/Bike Lane Project were not completed until June 2019 (Graff, 2019). The Madison Project was the first of a set planned transit priority. The project has since been shown to increase bus speeds (York, 2019). Given the network and data limitations from when the data was collection, analysis into bus lane effectiveness is a potential area of further research. The aggregated data provides multiple methods for visualizing trends and identifying hotspots. Yet, sometimes a more quantitative approach is useful.

6.3. Congestion Costs

One such quantitative approach is estimating the costs associated with congestion. This next section will look at the TriMet's network as a whole, high-use routes, most expensive routes, then route 9 for an applied example.

6.3.1. Network Level Costs

TriMet provides public reports of ridership statistics and cost estimations. Table 6-1 is a simplified version of Table 4-12 and Table 4-13. For the purpose of congestion estimates, it will be assumed that the passenger, times, and other estimates are correct. This now allows the comparisons of time attributed to the congestion to be estimated.

Table 6-1 — TriMet reported system performance metrics and estimates from complete dataset.

TriMet Ridership Report (Bus Only)	***2017***	***2018***	***Weighted***	***Estimated SED***
Total Yearly Boardings	57,820,520	56,737,466	56,971,478	57,465,226
Avg. Weekday Boardings	186,800	183,800	184,449	183,915
Revenue Hours	1,529,532	1,552,044	1,547,180	1,552,648
Revenue Miles	20,923,103	21,354,739	21,261,477	21,160,004

Another useful statistics, reported by TriMet, is the cost per boarding ride. For the 2019 fiscal year, that cost was \$5.46 per boarding passenger on buses (TriMet, 2019). As a high-level overview of the methodology outlined in Sections 4.3.5 and 4.4.3, the SED and aggregated data estimated the total revenue hours as 1,552,648. Beginning with the increases in moving time and disturbance time between stops, the SED and aggregated data predict similar values. The differences are a result of using periods, $\mathcal{p}$, versus timepoint periods, t_p. For the stop level analysis, the increased in ride and recovery time (i.e.) is estimated at184,631 hours at a cost of $25,700,636 per year, given $139.20 per hour. As a function of the system total, the increases from congestion account for approximately 12% of revenue hours. Using the aggregated methodologies, the results are lower at $22.1 million (~10% of revenue hours). As a cost per boarding passenger, the SED and aggregated methodologies result in an agency cost per boarding passenger at $0.45 and $0.38, respectively. These costs per boarding rider represent about 8% of the TriMet reported operating cost per boarding ride across the system and about 40% of the average revenue per boarding ride (i.e. $1.06 in 2019 fiscal year).

At the network level, the differences in the methodologies are primarily found in the amount of data needed. The aggregated data requires fewer data points and may potentially be more easily used by typical computers found in transit agencies.

6.3.2. Route Level Costs

The methodologies outlined within this dissertation allow for a more granular look. First, Table 6-2 gives the estimated passenger boardings for the 15 highest used routes, as reported by TriMet. The TriMet estimates are based on their route ridership reports for Spring 2019. Spring 2019 doesn't overlap with the dataset, but will serve a baseline.

The passenger estimates from the Table 6-2 are used to calculate the estimate costs, resulting from congestion, per boarding ride for each route. Table 6-3 and Table 6-4 show the operational costs for each route, as reported by TriMet, and the costs per boarding ride estimated using event level data and timepoint-segment data. The first takeaway from the highest use routes is that their reported costs are lower than the system average of $5.46.

Given their high usage, these routes also have lower costs per boarding ride than the system average as calculated by the TPS data. In a few cases, the ELD calculates higher costs.

Table 6-2 — Estimated passenger boardings for 15 highest usage routes.

	Weekly Total		**Weekday Avg.**		**Sat/Sun Avg.**		**Weekly Total *ELD***	
Route	***TriMet***	***ELD***	***TriMet***	***ELD***	***TriMet***	***ELD***	**Inbound**	**Outbound**
72	4,606,821	4,010,955	70,950	70,053	17,400	17,443	1,988,957	2,021,998
20	3,700,057	3,042,447	57,150	53,288	13,810	12,840	1,526,523	1,515,924
2	2,963,800	2,913,417	46,450	45,710	10,390	10,164	1,446,783	1,466,635
75	2,861,079	2,444,527	43,050	42,279	11,820	11,687	1,224,847	1,219,681
9	2,654,071	2,359,593	41,650	41,540	9,250	9,410	1,147,484	1,212,109
12	2,585,243	2,272,258	40,000	39,800	9,580	9,592	1,148,366	1,123,893
15	2,572,207	2,207,352	41,100	39,211	8,230	7,927	1,086,129	1,121,224
57	2,430,379	2,000,488	36,150	34,483	10,460	9,859	966,629	1,033,859
4	2,145,679	2,131,494	33,250	32,866	7,900	8,011	1,053,629	1,077,865
6	2,037,743	1,629,445	31,300	28,446	7,780	7,119	844,529	784,916
17	1,890,700	1,640,586	30,950	29,431	5,310	5,156	849,517	791,069
14	1,845,336	1,564,113	28,900	27,441	6,490	6,489	912,744	651,369
8	1,801,014	1,668,102	29,900	30,178	4,640	4,602	688,877	979,225
77	1,737,921	1,522,291	27,600	26,944	5,730	5,693	736,666	785,625
19	1,649,800	1,573,843	28,150	29,129	3,490	3,591	776,229	797,167

Looking specifically at Route 12 on weekends. The off-peak period has an estimated run-moving and disturbance time of 2197 seconds for the ELD while the TPS estimates the off-peak time at 2289. The differences may be an effect of the timepoint segments having different numbers of vehicles within a given hour than the route as a whole. Route 12 is a long route and often covers multiple service hours. While the ELD analysis level groups the route by the starting hour, the TPS level allows for more granular approach. The estimates at the TPS level may also be higher, like the weekend estimates for route 4. Yet, the estimates are correlated. For all routes, the daily ELD and daily TPS estimates have a correlation of 0.56 with each other and correlations of 0.51 and 0.79, respectively, with the costs reported by TriMet.

Table 6-3 — Agency costs per boarding ride for 15 highest usage routes, as daily average, weekday average, and weekend average.

	Daily			**Weekdays**			**Weekends**		
Route	***TriMet***	***ELD***	***TPS***	***TriMet***	***ELD***	***TPS***	***TriMet***	***ELD***	***TPS***
72	$ 2.96	$ 0.15	$ 0.16	$ 2.90	$ 0.14	$ 0.16	$ 3.20	$ 0.24	$ 0.16
20	$ 3.31	$ 0.21	$ 0.22	$ 3.19	$ 0.22	$ 0.22	$ 3.81	$ 0.10	$ 0.19
2	$ 3.34	$ 0.23	$ 0.21	$ 3.28	$ 0.24	$ 0.23	$ 3.61	$ 0.15	$ 0.16
75	$ 3.78	$ 0.33	$ 0.24	$ 3.68	$ 0.31	$ 0.22	$ 4.14	$ 0.54	$ 0.37
9	$ 3.71	$ 0.26	$ 0.25	$ 3.53	$ 0.26	$ 0.25	$ 4.52	$ 0.30	$ 0.24
12	$ 3.57	$ 0.60	$ 0.34	$ 3.47	$ 0.58	$ 0.34	$ 3.99	$ 0.82	$ 0.35
15	$ 3.54	$ 0.49	$ 0.26	$ 3.36	$ 0.48	$ 0.27	$ 4.44	$ 0.57	$ 0.23
57	$ 3.27	$ 0.27	$ 0.17	$ 3.13	$ 0.28	$ 0.18	$ 3.75	$ 0.21	$ 0.13
4	$ 4.08	$ 0.19	$ 0.16	$ 3.94	$ 0.22	$ 0.16	$ 4.67	$ 0.07	$ 0.15
6	$ 3.37	$ 0.32	$ 0.27	$ 3.20	$ 0.33	$ 0.27	$ 4.05	$ 0.21	$ 0.17
17	$ 3.93	$ 0.65	$ 0.31	$ 3.82	$ 0.66	$ 0.31	$ 4.57	$ 0.54	$ 0.29
14	$ 3.04	$ 0.19	$ 0.21	$ 2.89	$ 0.20	$ 0.21	$ 3.71	$ 0.14	$ 0.15
8	$ 3.94	$ 0.31	$ 0.21	$ 3.65	$ 0.31	$ 0.21	$ 5.81	$ 0.39	$ 0.19
77	$ 4.09	$ 0.42	$ 0.32	$ 4.02	$ 0.44	$ 0.33	$ 4.43	$ 0.20	$ 0.22
19	$ 4.37	$ 0.85	$ 0.43	$ 4.34	$ 0.86	$ 0.44	$ 4.61	$ 0.60	$ 0.30

Table 6-4 — Agency costs per boarding ride for 15 highest usage routes, as daily average, and average for inbound versus outbound service.

	Daily			**Inbound**		**Outbound**	
Route	***TriMet***	***ELD***	***TPS***	***ELD***	***TPS***	***ELD***	***TPS***
72	$ 2.96	$ 0.15	$ 0.16	$ 0.14	$ 0.16	$ 0.15	$ 0.16
20	$ 3.31	$ 0.21	$ 0.22	$ 0.18	$ 0.14	$ 0.25	$ 0.29
2	$ 3.34	$ 0.23	$ 0.21	$ 0.21	$ 0.22	$ 0.24	$ 0.21
75	$ 3.78	$ 0.33	$ 0.24	$ 0.26	$ 0.22	$ 0.40	$ 0.25
9	$ 3.71	$ 0.26	$ 0.25	$ 0.22	$ 0.23	$ 0.31	$ 0.26
12	$ 3.57	$ 0.60	$ 0.34	$ 0.62	$ 0.33	$ 0.58	$ 0.35
15	$ 3.54	$ 0.49	$ 0.26	$ 0.51	$ 0.27	$ 0.47	$ 0.26
57	$ 3.27	$ 0.27	$ 0.17	$ 0.27	$ 0.15	$ 0.27	$ 0.19
4	$ 4.08	$ 0.19	$ 0.16	$ 0.19	$ 0.14	$ 0.20	$ 0.18
6	$ 3.37	$ 0.32	$ 0.27	$ 0.24	$ 0.22	$ 0.41	$ 0.31
17	$ 3.93	$ 0.65	$ 0.31	$ 0.64	$ 0.29	$ 0.66	$ 0.33
14	$ 3.04	$ 0.19	$ 0.21	$ 0.12	$ 0.18	$ 0.29	$ 0.24
8	$ 3.94	$ 0.31	$ 0.21	$ 0.52	$ 0.29	$ 0.17	$ 0.15
77	$ 4.09	$ 0.42	$ 0.32	$ 0.55	$ 0.35	$ 0.30	$ 0.30
19	$ 4.37	$ 0.85	$ 0.43	$ 0.86	$ 0.40	$ 0.84	$ 0.46

The fifteen routes reported above stand out for their usage. In contrast, the twelve routes shown in Table 6-5 stand out for their reported costs. The passenger usage estimates are shown in Table 6-6. For these routes, the estimates from the TPS model tend to be larger than the ELD estimates. Route 97 stands out as having an extremely large difference between the inbound estimates. The difference is caused by the off-peak estimate hour, which doesn't exist for Route 97. Route 97 operates from 07:00-09:00 and from 15:00-18:00. None of these times fall within the off-peak interval. Given this issue for the model formulation, the off-peak estimate was based on the 15th percentile. The sum of the 15th percentiles across timepoint segments is much smaller tchan the 15th percentile for the route as a whole.

Table 6-5 — Agency costs per boarding ride for routes costing at least $10 per boarding ride.

	Daily			**Inbound**		**Outbound**	
Route	***TriMet***	***ELD***	***TPS***	***ELD***	***TPS***	***ELD***	***TPS***
97	$ 21.97	$ 2.53	$ 6.87	$ 2.53	$10.46	$ 2.54	$ 3.06
152	$ 15.84	$ 0.98	$ 1.90	$ 0.75	$ 1.47	$ 1.26	$ 2.24
82	$ 13.08	$ 0.44	$ 1.85	$ 0.34	$ 1.73	$ 0.59	$ 2.02
84	$ 12.48	$ 0.59	$ 2.56	$ 0.98	$ 1.20	$ 0.32	$ 4.36
154	$ 11.54	$ 0.73	$ 3.14	$ 1.61	$ 1.41	$ 0.25	$ 3.91
11	$ 11.22	$ 0.35	$ 2.53	$ 0.35	$ 1.05	$ 0.34	$ 4.26
24	$ 11.09	$ 0.63	$ 2.17	$ 0.81	$ 2.55	$ 0.49	$ 1.84
32	$ 10.81	$ 0.47	$ 1.00	$ 0.74	$ 0.58	$ 0.13	$ 1.52
34	$ 10.79	$ 3.87	$ 3.46	$ 1.39	$ 2.96	$ 6.41	$ 4.05
39	$ 10.49	$ 1.02	$ 1.27	$ 0.71	$ 0.86	$ 1.36	$ 1.71
29	$ 10.41	$ 0.51	$ 2.84	$ 0.55	$ 3.37	$ 0.46	$ 2.48
30	$ 10.11	$ 0.49	$ 1.67	$ 0.67	$ 1.08	$ 0.33	$ 3.06

Table 6-6 — Estimated passenger boardings for routes costing at least $10 per boarding ride.

	Daily		**Weekdays**		**Outbound**	**Inbound**
Route	***TriMet***	***ELD***	***TriMet***	***ELD***	***ELD***	***ELD***
97	20,857	19,835	400	367	10,260	9,574
152	52,143	53,763	1,000	971	29,486	24,278
82	57,357	56,076	1,100	1,021	33,373	22,703
84	18,250	21,416	350	401	8,813	12,603
154	36,500	38,732	700	647	13,533	25,199
11	41,714	40,663	800	750	21,251	19,412
24	300,343	132,073	4,700	2,510	59, 554	72,519
32	160,079	159,270	3,000	2,917	88,811	70,460
34	140,786	138,529	2,700	2,649	69,977	68,552
39	39,107	44,241	750	844	23,423	20,818
29	41,714	68,279	800	1,035	34,433	33,846
30	166,857	171,590	2,950	2,955	79,553	92,027

At the route level, passenger costs were also estimated. However, unlike the costs to agencies, the passenger costs of riding and waiting times do not have reported values for direct comparisons. In Table 6-7, the costs per passenger are estimated for the highest-usage routes. There is a clear difference between the estimates at the two analysis levels, specifically the TPS level is much less. The same relationship holds for passenger wait times that estimates at the aggregated level are less than the estimates using ELD. The source of these differences may be the subject of further research. Ideally, an alternate method, independent from those outlined in this dissertation, could be employed that provides a third estimate and additional comparison datapoint.

Table 6-7 — Passenger riding costs per boarding ride for 15 highest usage routes, as weekday average, and average for inbound versus outbound service.

	Weekday Average		**Inbound Daily**		**Outbound Daily**	
Route	***ELD***	***TPS***	***ELD***	***TPS***	***ELD***	***TPS***
72	$ 0.35	$ 0.15	$ 0.34	$ 0.15	$ 0.40	$ 0.16
20	$ 0.74	$ 0.23	$ 0.59	$ 0.17	$ 0.80	$ 0.28
2	$ 0.54	$ 0.28	$ 0.44	$ 0.25	$ 0.54	$ 0.27
75	$ 0.75	$ 0.18	$ 0.61	$ 0.16	$ 0.94	$ 0.19
9	$ 0.52	$ 0.30	$ 0.41	$ 0.25	$ 0.62	$ 0.33
12	$ 1.19	$ 0.39	$ 1.23	$ 0.42	$ 1.16	$ 0.34
15	$ 0.88	$ 0.23	$ 0.89	$ 0.24	$ 0.86	$ 0.20
57	$ 0.56	$ 0.22	$ 0.51	$ 0.16	$ 0.56	$ 0.25
4	$ 0.37	$ 0.17	$ 0.28	$ 0.10	$ 0.35	$ 0.21
6	$ 1.33	$ 0.25	$ 1.31	$ 0.27	$ 1.38	$ 0.22
17	$ 0.53	$ 0.20	$ 0.39	$ 0.17	$ 0.64	$ 0.21
14	$ 1.31	$ 0.30	$ 1.31	$ 0.29	$ 1.25	$ 0.30
8	$ 0.27	$ 0.15	$ 0.19	$ 0.12	$ 0.36	$ 0.18
77	$ 0.43	$ 0.19	$ 0.69	$ 0.28	$ 0.24	$ 0.11
19	$ 0.90	$ 0.23	$ 1.07	$ 0.24	$ 0.65	$ 0.22

6.4. Travel Speeds

Transit travel speeds have also been a focus of ongoing research into transit performance. The same variables that were used to estimate service durations are likely related to travel speeds. Specifically, the dependent variables of the total travel time (i.e. ${}^{\Sigma T}\tilde{T}RVL_t$) and moving time (i.e. ${}^{\Sigma T}\tilde{M}OVE_t$) models may be combined with the independent variable for total distance traveled (i.e. ${}^{\Sigma}MILES_t$) to estimate average travel speed (i.e. ${}_{\mu(trvl)}^{\Sigma}MPH_t$) and the average moving speed (i.e. ${}_{\mu(move)}^{\Sigma}MPH_t$) within a timepoint segment.

Definition 6-1 — ${}_{\mu(trvl)}^{\Sigma}MPH_t$ [mph] is an *Average Speed* for all vehicles within a timepoint segment including all stops.

$$(6.4.1) \qquad \forall\, {}_{\mu(trvl)}^{\Sigma}MPH_t \in \left\{ {}_{\mu(trvl)}^{\Sigma}MPH_t : t \in \mathbb{t} \right\}, \left({}_{\mu(trvl)}^{\Sigma}MPH_t = \frac{{}^{\Sigma}MILES_t}{{}^{\Sigma T}\tilde{T}RVL_t} \right)$$

Definition 6-2 — ${}_{\mu(move)}^{\Sigma}MPH_{t}$ [mph] is an *Average Moving Speed* for all vehicles within a timepoint segment. It does not include any time stopped at or between bus stops.

$$(6.4.2) \qquad \forall\, {}_{\mu(move)}^{\Sigma}MPH_{t} \in \left\{ {}_{\mu(move)}^{\Sigma}MPH_{t} : t \in \mathbb{t} \right\}, \left({}_{\mu(move)}^{\Sigma}MPH_{t} = \frac{{}^{\Sigma}MILES_{t}}{{}^{\Sigma T}\widetilde{M}OVE_{t}} \right)$$

6.4.1. Single Variable Input

For these average speeds, the independent variables from the regression modeling of Chapter 5 serve as a staring place where single variable models are produced. Given the influence of the built environment on speed (e.g. speed limits, travel lanes, etc.) the first step in the analysis is to evaluate variables individually, but show the results of a regression model at each of ten speed-quantiles and all data. As such, eleven different models were run for each independent variables. This dissertation will not try to present the full range of results; rather it will focus on two test cases (Figure 6-16 through Figure 6-19 on pages 198 and 199) for each dependent variable that highlight the overall trends. On each plot, a blue trendline indicates a positive slope, a red trendline indicates a negative slope; solid lines were significant, and dashed lines were insignificant. Lastly, the adjusted R-squared for that model is given in purple.

For average speed, Figure 6-16 and Figure 6-17 show speed plotted against the number of vehicles (i.e. ${}^{\Sigma}VEH$) and the number of non-serviced stops (of all types) (i.e. $\sum\left[{}_{thru}^{\Sigma L}VAR\right]$) per segment. For moving speed, Figure 6-18 and Figure 6-19 the number of serviced stops (of all types) (i.e. $\sum[{}^{\Sigma L}VAR]$) and the number of boarding passengers (i.e. ${}^{\Sigma}\hat{O}NS$) per TPS are given. In all cases, and for all of the independent variables, the main takeaway is the same: inconsistency.

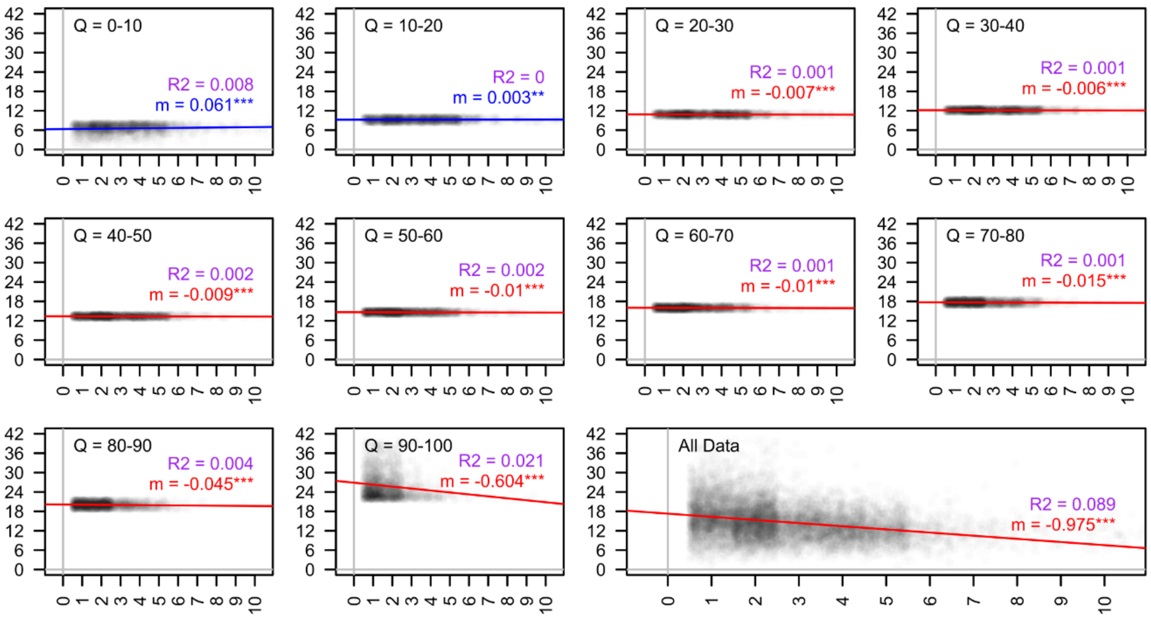

Figure 6-16 — Average travel speed ($_{\mu(trvl)}^{\Sigma}MPH$) versus the number of vehicles ($^{\Sigma}VEH$) per timepoint-segment.

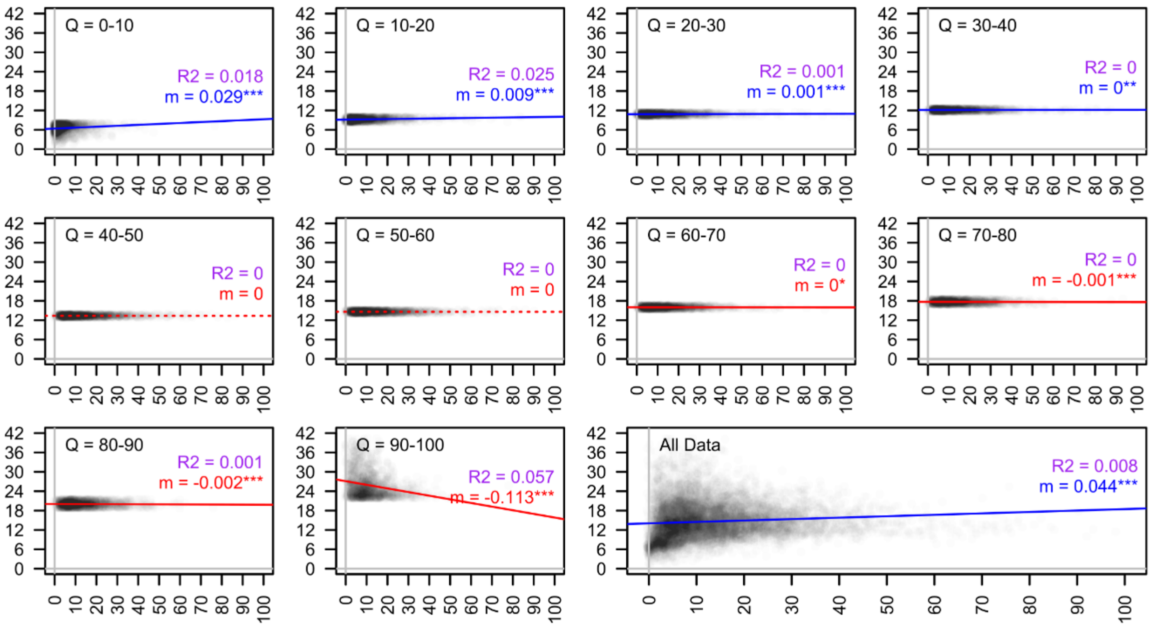

Figure 6-17 — Average travel speed ($_{\mu(trvl)}^{\Sigma}MPH$) versus the number of non-serviced stops ($\sum[_{thru}^{\Sigma L}VAR]$) per timepoint-segment.

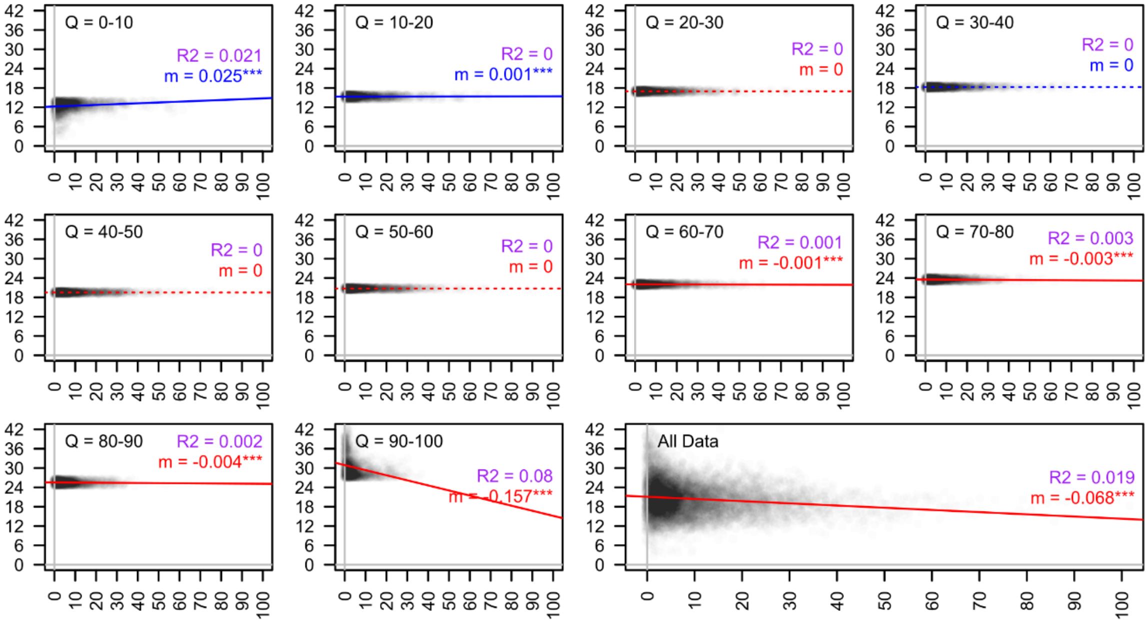

Figure 6-18 — Moving travel speed ($_{\mu(move)}^{\Sigma}MPH$) quantiles versus number of serviced bus stops ($\sum[{}^{\Sigma L}VAR]$) per timepoint-segment.

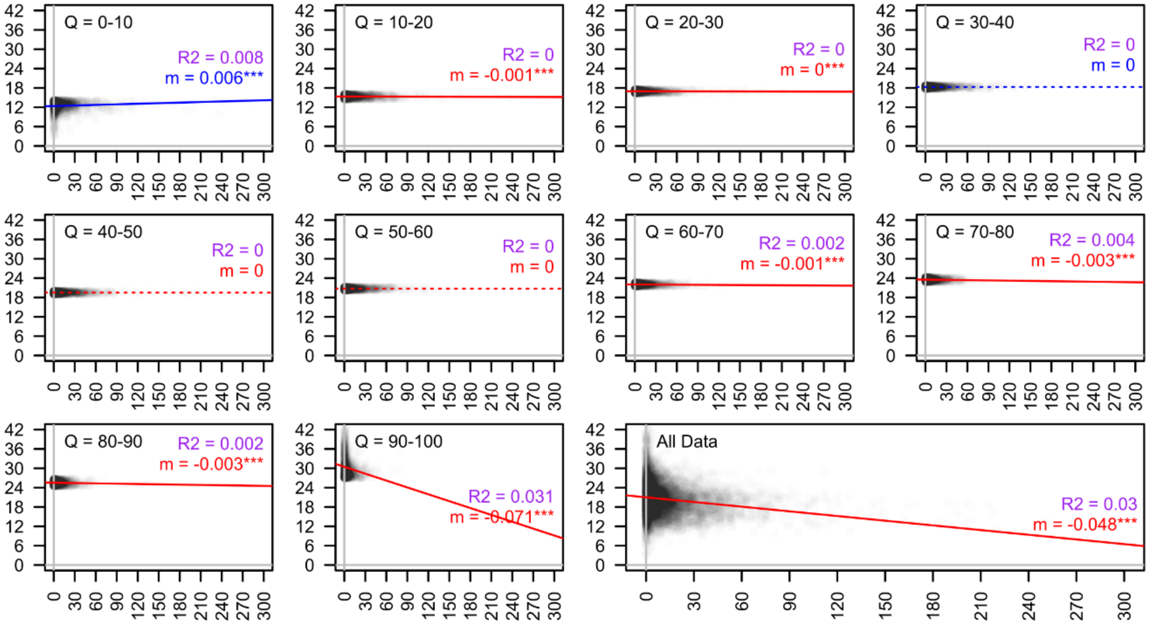

Figure 6-19 — Moving travel speed ($_{\mu(move)}^{\Sigma}MPH$) quantiles versus the boarding passengers ($^{\Sigma}\hat{O}NS$) per timepoint-segment.

The result changes between quantiles. Not one variable gave consistently signed and significant results at all quantiles for either dependent speed variable. However, some variables were consistent for the middle 80% of the data. Travel speeds are not normally distributed and the highest and lowest speed quantiles are notably different than the rest. As such, the full model relationship to speed was largely influenced by an exaggerated effect in one or both of the 10% and 90% quantiles.

6.4.2. Two-Variable Models

As a second test, two variables were tested simultaneously. Given the result of the one-variable tests four variables were selected to tested against other independent inputs. Table 6-8 shows the signs of the coefficients (using all quantiles) for the 1-variable model and 2-variable models for average speed. Sign changes have been highlighted and the results were same for moving speed.

Table 6-8 — Signs of (row) coefficients for average total travel speed models.

	1-Variable Models	**2-Variable Models**			
Variable		${}^{\Sigma}VEH$	${}^{\Sigma}\hat{O}NS$	${}^{\Sigma}\hat{O}FFS$	$\sum[{}^{\Sigma L}VAR]$
${}^{\Sigma}VEH$	−	-NA-	−	−	−
${}^{\Sigma}\hat{O}NS$	−	−	-NA-	−	−
${}^{\Sigma}\hat{O}FFS$	−	−	−	-NA-	−
${}^{\Sigma}\hat{L}IFT$	−	−	−	−	−
$\sum[{}^{\Sigma L}VAR]$	−	+	−	−	-NA-
$\sum[{}^{\Sigma L}_{thru}VAR]$	+	+	+	+	+
$\sum[{}^{\Sigma Ls}_{skd}VAR]$	−	+	+	+	+
${}^{\Sigma}VEH \times FREQ$	−	+	−	−	−
${}^{\Sigma}VEH \times W_1^{AM}$	−	+	−	−	+
${}^{\Sigma}VEH \times W_1^{PM}$	−	−	−	−	−
${}^{\Sigma}VEH \times W_0^{P}$	−	−	−	−	−
${}^{\Sigma Id}INT$	−	−	−	−	−
${}^{\Sigma Is}INT$	−	+	+	+	+

The specific explanatory power of each variable is included in the appendix (Table C-10 and Table C-11) Table 6-9 gives the ratio of explanatory powers of average moving speed to average total travel speed. With one exception (only for the single variable models), the included variables predict average total travel speed better than moving speed, as indicated by percentages less than 100. These result relates back to the models predicting service duration. While moving speed will be correlated to events a bus stops, those independent variable inputs will better account for a reduced overall speed rather than reduced moving speed.

Table 6-9 — Ratio of average moving speed adjusted R-squared to average total travel speed adjusted R-squared.

	1-Variable Models	**2-Variable Models**			
Variable		${}^{\Sigma}VEH$	${}^{\Sigma}\hat{O}NS$	${}^{\Sigma}\hat{O}FFS$	$\sum[{}^{\Sigma L}VAR]$
$\sum[{}^{\Sigma L}_{thru}VAR]$	269%	73%	54%	63%	48%
${}^{\Sigma Id}INT$	87%	56%	45%	52%	40%
${}^{\Sigma}VEH \times W_1^{AM}$	63%	43%	31%	36%	21%
${}^{\Sigma}VEH \times W_1^{PM}$	62%	46%	35%	40%	28%
${}^{\Sigma Is}INT$	59%	43%	30%	35%	21%
${}^{\Sigma}VEH$	43%	-NA-	37%	40%	38%
${}^{\Sigma}VEH \times FREQ$	40%	43%	32%	37%	26%
${}^{\Sigma}\hat{O}FFS$	36%	40%	32%	-NA-	30%
${}^{\Sigma}VEH \times W_0^{P}$	31%	43%	30%	35%	21%
${}^{\Sigma}\hat{O}NS$	30%	37%	-NA-	32%	28%
${}^{\Sigma}\hat{L}IFT$	24%	41%	29%	34%	21%
$\sum[{}^{\Sigma L}VAR]$	21%	38%	28%	30%	-NA-
$\sum[{}^{\Sigma Ls}_{skd}VAR]$	7%	58%	36%	42%	31%

Lastly, the change the adjusted R-squared was calculated after adding a second variable. The change was based on the one-variable models labeled in the columns. For moving speed (Table 6-10), adding any of the independent variables listed resulted in an improved model, but one variable stands out. On its own, $\sum[{}^{\Sigma L}_{thru}VAR]$ had an adjusted r-squared of 0.0212, which was less than each of the column variables. But as the second variable, it improved the total explanatory power by an average 249%. Generally, the number of stops not serviced has the potential to be highly misleading about a timepoint segments. It doesn't differentiate between busy segments with many vehicles and few skipped stops per vehicle and nearly empty segments with one vehicle and many skipped stops. Each of the four column-variable help put segments into a context of overall timepoint-segment usage.

Table 6-10 — Change to adjusted R-squared for moving speed one-variable (column) models adding row-variables.

	Percent Change				
Variable	${}^{\Sigma}VEH$	${}^{\Sigma}\hat{O}NS$	${}^{\Sigma}\hat{O}FFS$	$\sum[{}^{\Sigma L}VAR]$	**Average**
$\sum[{}^{\Sigma L}_{thru}VAR]$	283%	156%	161%	396%	249%
${}^{\Sigma Id}INT$	57%	67%	71%	135%	83%
${}^{\Sigma}VEH$	-NA-	42%	44%	102%	47%
$\sum[{}^{\Sigma Ls}_{skd}VAR]$	33%	19%	20%	60%	33%
${}^{\Sigma}\hat{O}NS$	11%	-NA-	32%	56%	25%
${}^{\Sigma}VEH \times W_1^{PM}$	9%	20%	21%	40%	23%
${}^{\Sigma}\hat{O}FFS$	9%	28%	-NA-	51%	22%
${}^{\Sigma}VEH \times FREQ$	<1%	12%	13%	24%	12%
${}^{\Sigma}\hat{L}IFT$	2%	3%	3%	7%	4%
${}^{\Sigma}VEH \times W_0^P$	<1%	3%	3%	3%	2%
${}^{\Sigma}VEH \times W_1^{AM}$	<1%	1%	1%	2%	1%
$\sum[{}^{\Sigma L}VAR]$	2%	<0.1%	<1%	-NA-	0%
${}^{\Sigma Is}INT$	<0.01%	<1%	<0.1%	<1%	0%

Given the influence of non-serviced stops, the contribution (as the second variable) was tested for each type of non-serviced stop. The results are given in Table 6-11 and indicate that including only non-serviced farside stops (i.e. ${}^{\Sigma L}_{thru}FAR$) results in model improvements almost as large as all non-serviced stops. Non-serviced opposite (i.e. ${}^{\Sigma L}_{thru}OPP$) and "at" (i.e. ${}^{\Sigma L}_{thru}AT$) locations also results in large gains. Unfortunately, these results come with a huge cavate. The models capture only about ten percent of total variability, which limits practical use. Similar tables for results for average total travel time (like Table 6-10 and Table 6-11) are included in the appendix as Table C-12 and Table C-13, respectively.

Table 6-11 — Change to adjusted R-squared for moving speed 1-variable (column) models adding row-variables for non-serviced stops by types.

	Total Change				Percent Change				
Variable	${}^{\Sigma}VEH$	${}^{\Sigma}\hat{O}NS$	${}^{\Sigma}\hat{O}FFS$	$\sum[{}^{\Sigma L}V]$	${}^{\Sigma}VEH$	${}^{\Sigma}\hat{O}NS$	${}^{\Sigma}\hat{O}FFS$	$\sum[{}^{\Sigma L}V]$	**Avg**
$\sum[{}_{thru}^{\Sigma L}VAR]$	0.1096	0.0471	0.0472	0.0770	283%	156%	161%	396%	249%
${}_{thru}^{\Sigma L}FAR$	0.0954	0.0563	0.0592	0.0764	247%	186%	202%	392%	257%
${}_{thru}^{\Sigma L}OPP$	0.0765	0.0616	0.0585	0.0686	198%	203%	200%	352%	238%
${}_{thru}^{\Sigma L}AT$	0.0455	0.0338	0.0336	0.0373	118%	112%	115%	192%	134%
${}_{thru}^{\Sigma L}NEAR$	0.0109	0.0019	0.0018	0.0045	28%	6%	6%	23%	16%
${}_{thru}^{\Sigma L}TC$	0.0030	0.0057	0.0042	0.0050	8%	19%	14%	25%	17%

Full models predicting transit speeds were not a primary focus of dissertation and are not included. However, preliminary investigations using timepoint segment data provide a foundation for future research. In particular, it can be shown that the aggregated independent variables, used to predict total travel time, are related to average transit speeds and can predict 10-12% of speed variability including just two variables. However, those variables are less related to the moving speed without stops and two-variable models captures an average of just 38% of the variability. Preliminary results indicate some of the same limitations as was seen when modeling disturbance stop times. The events at bus stops are indirectly correlated, but not directly applicable as a primary model input.

6.5. TriMet Route 9

TriMet's route 9 is one route that has been thoroughly studied. In addition to its importance in connecting Gresham to the urban core, it runs along an urban arterial that carries upwards of 40,000 people daily. Route 9 will serve as a case-study for how the aggregated methodology can be used to identify hotspots or other problematic areas along a route. The timepoints and timepoint-segments shown in Figure 6-20 are those used in analysis. Given the odd behavior at terminal stops on the transit mall, NW Flanders and NW Davis were not considered the center of a TPS, despite being timepoints. As such, service on the transit mall falls within a single segment.

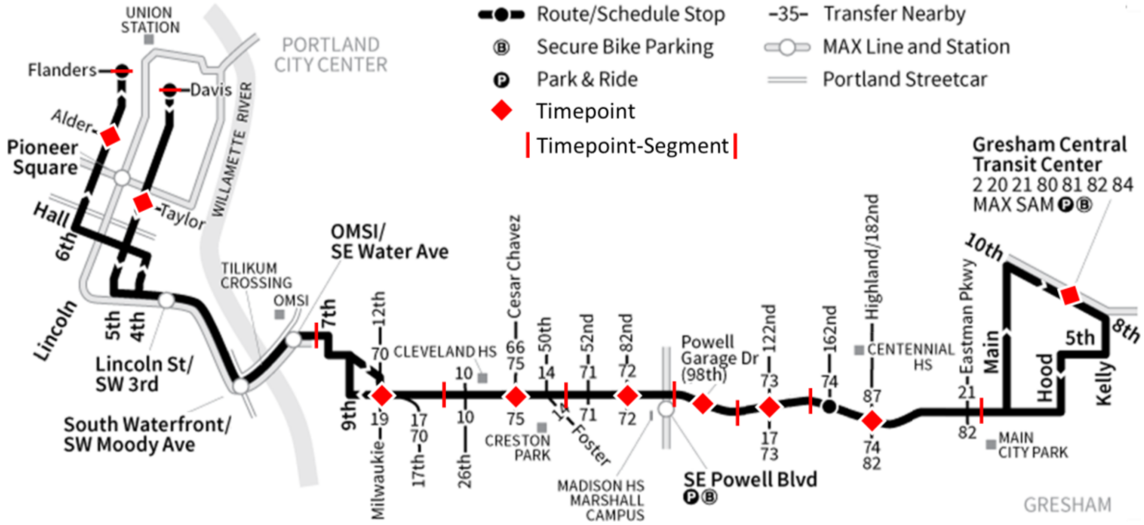

Figure 6-20 —Map of Route 9 with designated timepoint segment.

Figure 6-21, Figure 6-22, and Figure 6-23 are specific to route 9, but present the same headway performance metrics as Figure 6-1, Figure 6-2 and Figure 6-3, respectively. Both inbound and outbound trips are included, broken down by the number of vehicles in each timepoint-segment. As previously discussed, the non-normality of the scheduled arrival deviation index (Figure 6-21) is more pronounced for a single route than it was for the network as a whole. While the skew towards smaller deviations is still present for trips with five or fewer vehicle, the range of values is reduced. Looking instead at the “actual” arrival deviation index (Figure 6-22), the violin plots follow a more normal distributions.

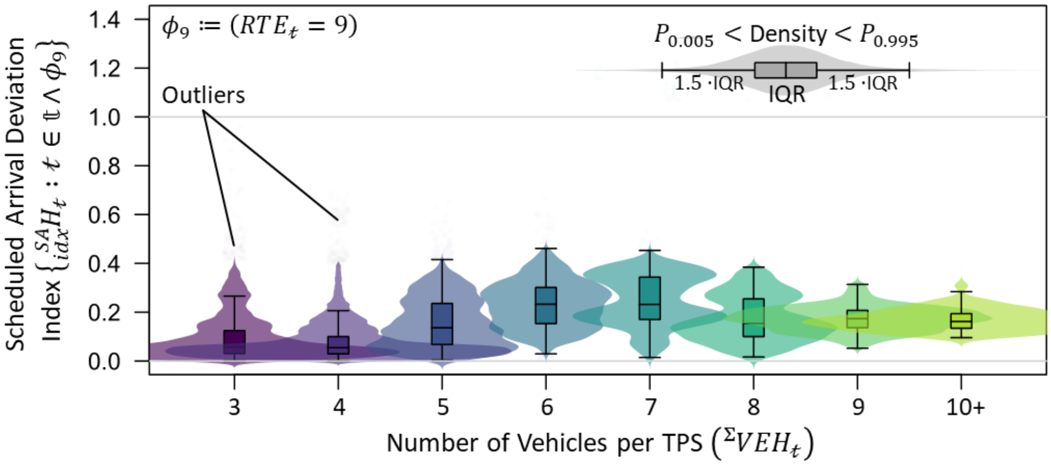

Figure 6-21 —Violin and box-plots for Route 9. Scheduled arrival deviation index $({}^{SA}_{idx}H_t)$ given number of vehicles per timepoint-segment $({}^{\Sigma}VEH_t)$.

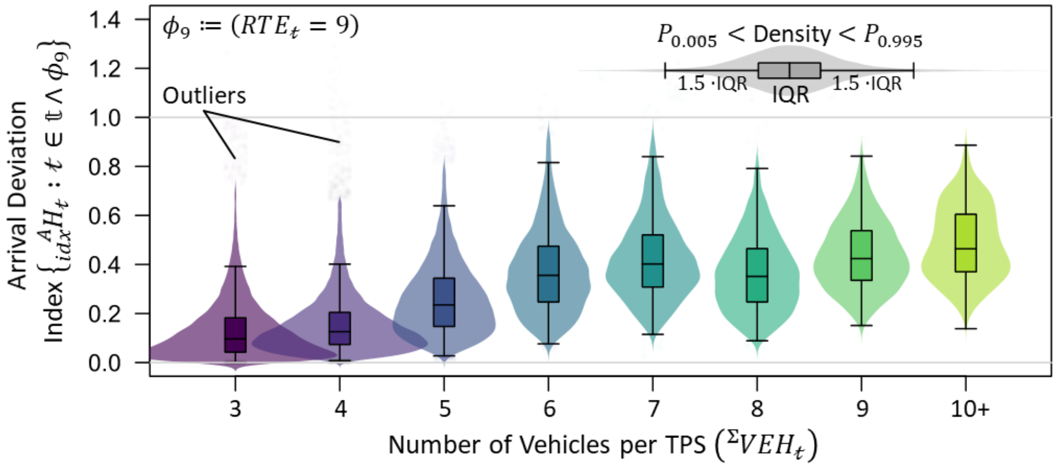

Figure 6-22 — Violin and box-plots for Route 9. Arrival deviation index $({}^{A}_{idx}H_t)$ given number of vehicles per timepoint-segment $({}^{\Sigma}VEH_t)$.

Again, an increased IQR and median is observed as the number of vehicles increases. Using the adjusted arrival deviation index (Figure 6-23), similar trends are observed. But now, segments with seven vehicles stand out, potentially indicating schedule instabilities for that number of vehicles. Given that hours with six or more vehicles are indicative of peak travel, further investigations may benefit visuals by time-of-day.

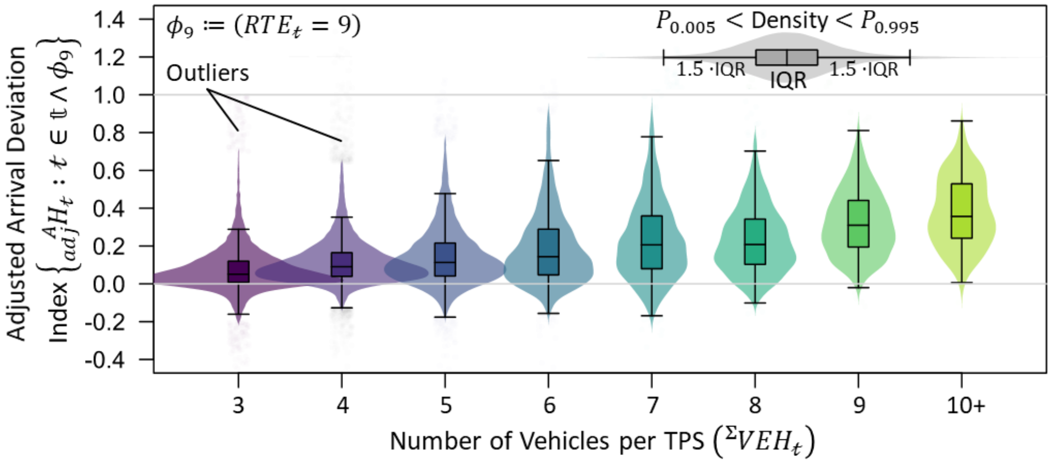

Figure 6-23 — Violin and box-plots for Route 9. Adjusted arrival deviation index $({}_{adj}^{A}H_t)$ given number of vehicles per timepoint-segment $({}^{\Sigma}VEH_t)$.

6.5.1. Inbound vs Outbound

The following graphics will focus on the adjusted deviation indexes only. First, Figure 6-24 looks at inbound service and Figure 6-25 focuses on outbound service. By time-of-day, inbound and outbound service are notably different. Given the very different demands on each direction for route 9, differences are to be expected, but still warrant explanation. First, the IQRs and overall ranges are generally higher for outbound service than for inbound; the exception is during 7:00 and 10:00 AM hours. 15:00 to 18:00 for outbound service are distinctly different than other times of day.

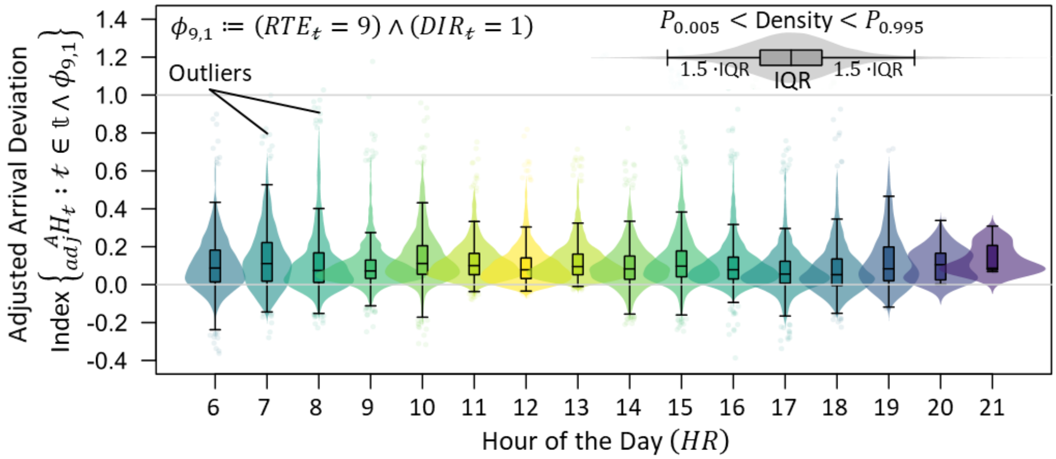

Figure 6-24 — Violin and box-plots for Route 9: *inbound* to downtown city-center. Adjusted arrival deviation index $({}_{adj}^{A}H_t)$ given hour-of-the-day (HR_t).

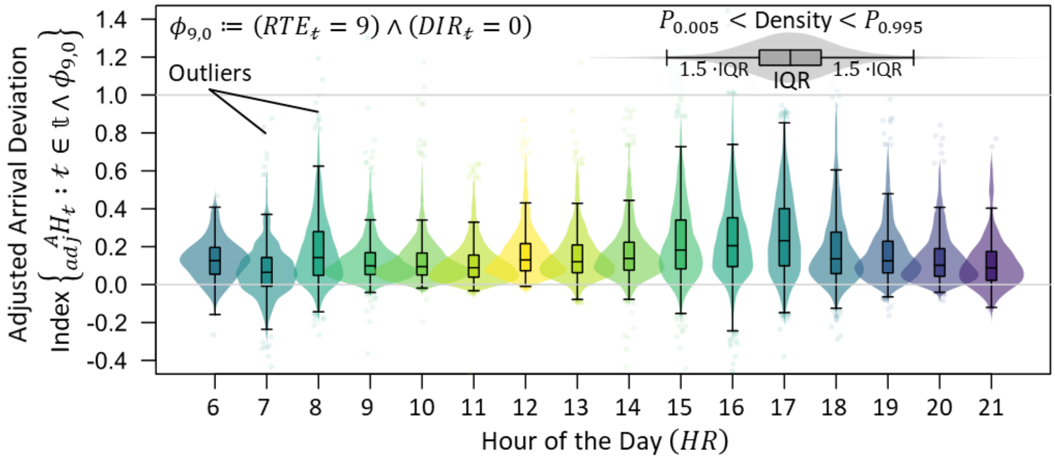

Figure 6-25 — Violin and box-plots for Route 9: *outbound* to Gresham Transit Center. Adjusted arrival deviation index $({}_{adj}^{A}H_t)$ given hour-of-the-day (HR_t).

Figure 6-24 and Figure 6-25 include all segments along the route. Figure 6-26 and Figure 6-27 add a restriction for the number of vehicles in each segment. The former limits segments to five or fewer vehicles for outbound service only. The latter limits segments to more than five vehicles, but displays both direction. In Figure 6-27, there is no directional overlap in high vehicle segments by hour.

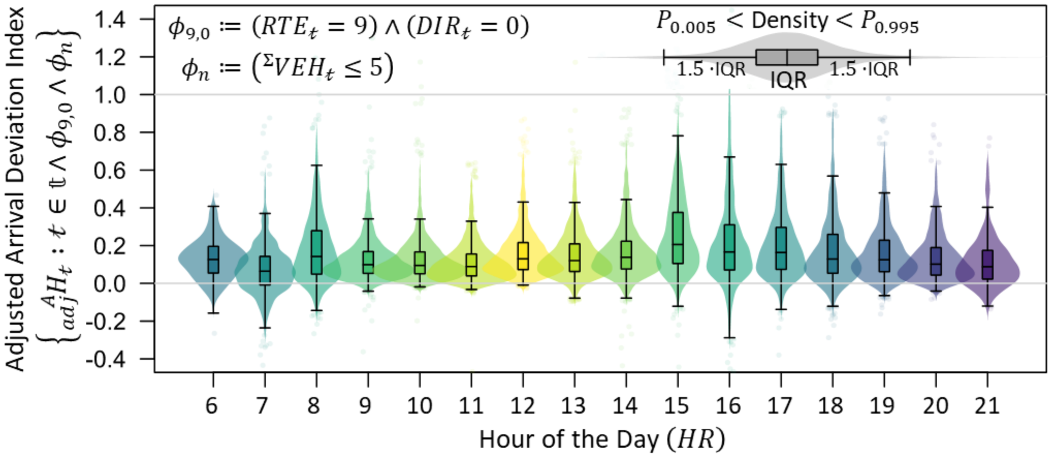

Figure 6-26 — Violin and box-plots for Route 9: *outbound* to Gresham Transit Center. Adjusted arrival deviation index (${}_{adj}^{A}H_t$), given hour-of-the-day (HR_t) and limited to five or fewer vehicles per timepoint-segment.

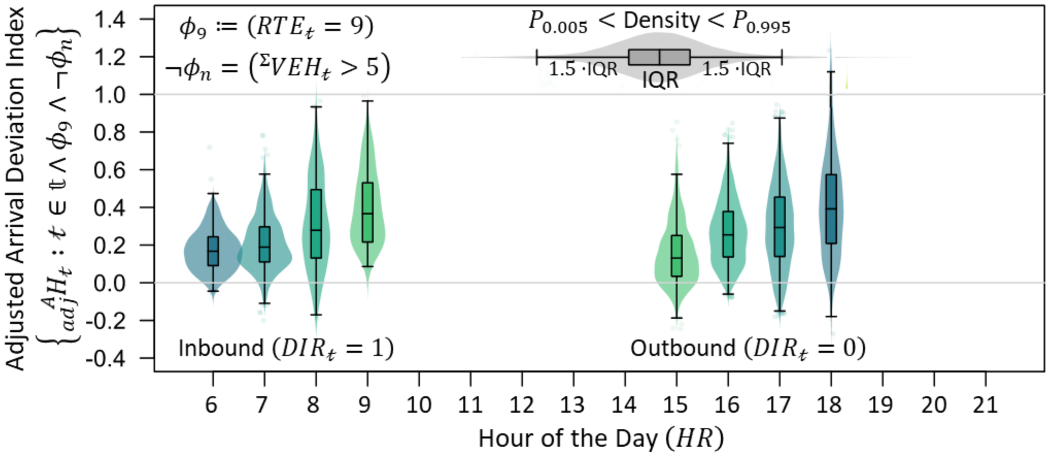

Figure 6-27 — Violin and box-plots for Route 9. Adjusted arrival deviation index (${}_{adj}^{A}H_t$), given hour-of-the-day (HR_t) and limited to more than five vehicles per timepoint-segment.

The changes to the distributions, based on number of vehicles, generally follow the trends presented in Figure 6-23. Much of the higher IQRs represent those segments with more vehicles. Examinations of the schedules and times-of-day with highest headway deviations has useful applications in planning, but doesn’t directly narrow down on the sources of those deviations along the route.

6.5.2. Timepoint-Segments

The same distributions may be produced by each timepoint segment to see how performance changes throughout a day and along a route. Headway performance may first be plotted for the first and last timepoint segments. Figure 6-28 (for inbound service) and Figure 6-29 (for outbound service) confirm what has been previously discussed: deviations generally increase along a route.

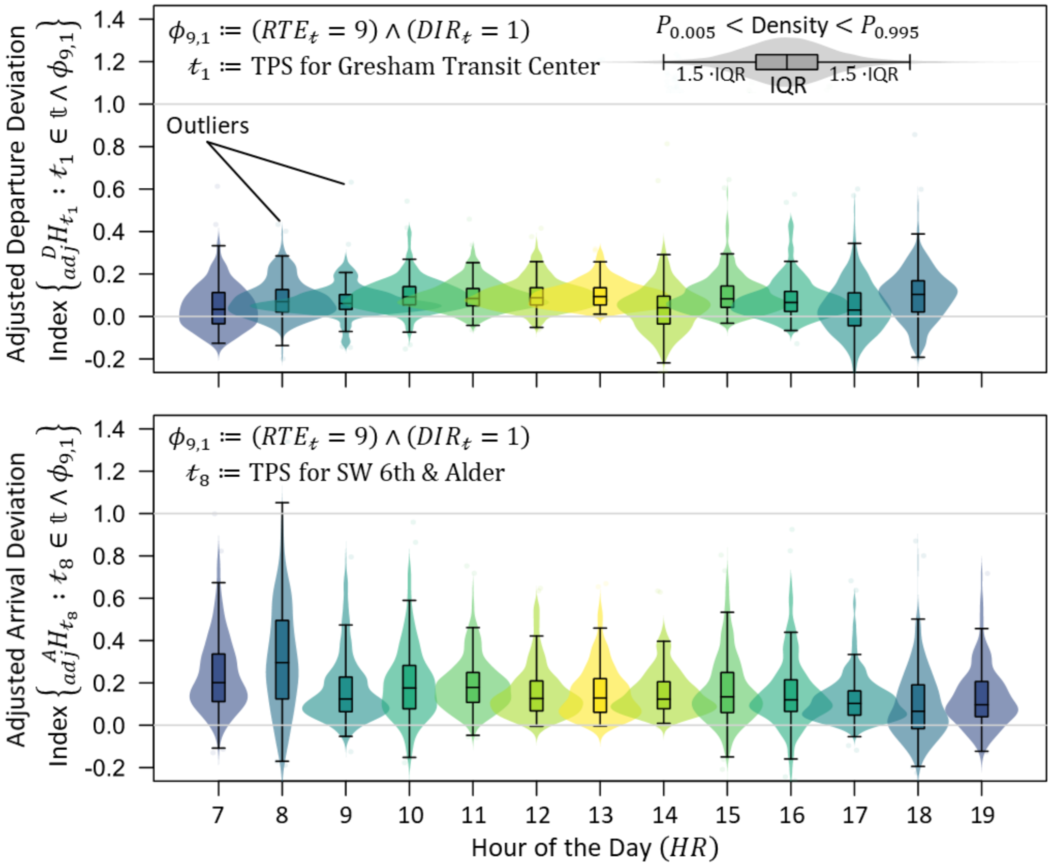

Figure 6-28 — Violin and box-plots for Route 9: *inbound* to downtown city-center. Adjusted departure deviation index $({}^{D}_{adj}H_{t})$ for first TPS (top) and adjusted arrival deviation index $({}^{A}_{adj}H_{t})$ for last TPS (bottom), given HR_{t}.

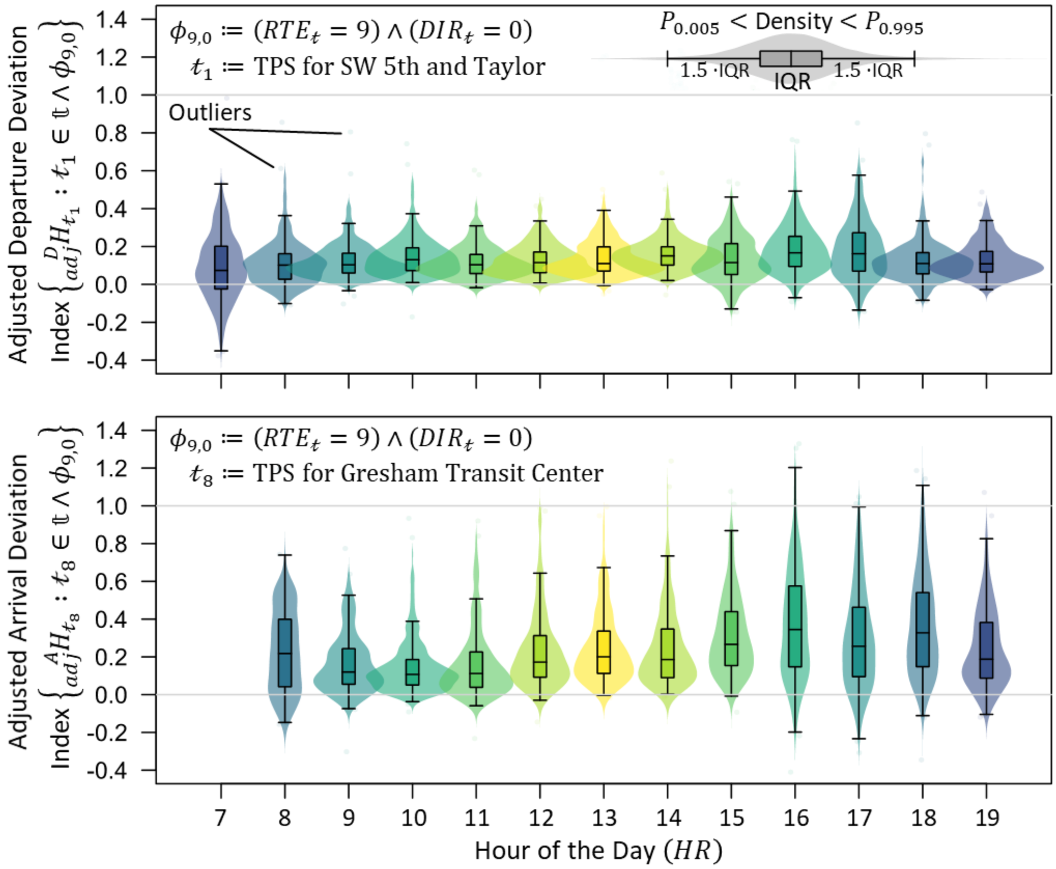

Figure 6-29 — Violin and box-plots for Route 9: *outbound* to Gresham TC. Adjusted departure deviation index (${}_{adj}^{D}H_t$) for first TPS (top) and adjusted arrival deviation index (${}_{adj}^{A}H_t$) for last TPS (bottom), given HR_t.

By plotting the adjusted deviation indexes for all segments, the segments that have the highest influence may be identified. While not included in this dissertation, all timepoint segments were plotted for route 9 in both directions. Figure 6-30 shows a single segment that was identified as having the largest change inbound. The adjusted deviation indexes for the first (arrival) and last (departure) stop of the timepoint segment are plotted. During both the AM and PM peak periods, the timepoint-segment surrounding SE Powell and Milwaukie demonstrates a large change.

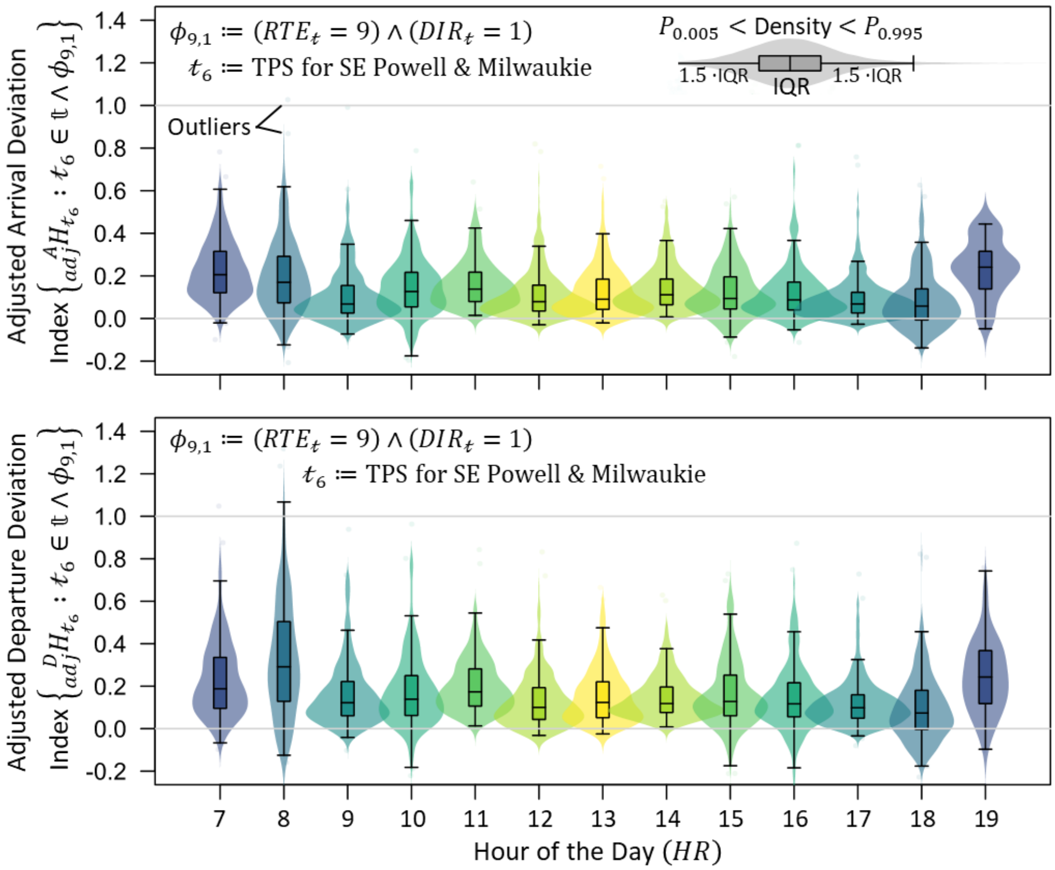

Figure 6-30 — Violin and box-plots for Route 9: *inbound* to city-center. Adjusted arrival deviation index (${}_{adj}^{A}H_t$) (top) and adjusted departure deviation index (${}_{adj}^{D}H_t$) (bottom) for sixth TPS around SE Powell & Milwaukie, given HR_t.

The identification of the timepoint segment around SE Powell and Milwaukie is an interesting confirmation of previously published investigations into Route 9 that looked at congestion hotspots (Stoll, et al., 2016) and reliability indexes (Glick & Figliozzi, 2017). Figure 6-31 and Figure 6-33 are figures taken from those publications, which were created using high-resolution transit data to examine and areas of slow travel speeds along Powell Blvd. Those methods are computationally intensive, but can provide highly detailed information about a segment. Both figures clearly show an area of slow speeds that could account for the reductions in headway performance metrics.

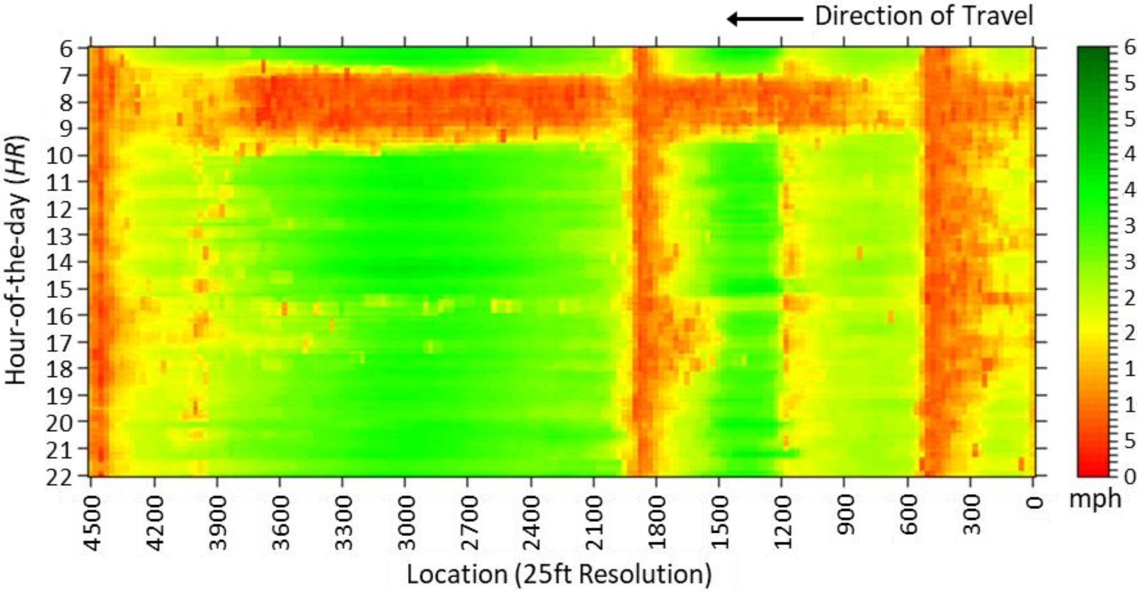

Figure 6-31 — Speed map of Route 9, by time of day, for segment of Powell Blvd leading up to SW Powell and Milwaukie Blvd (Stoll, et al., 2016).

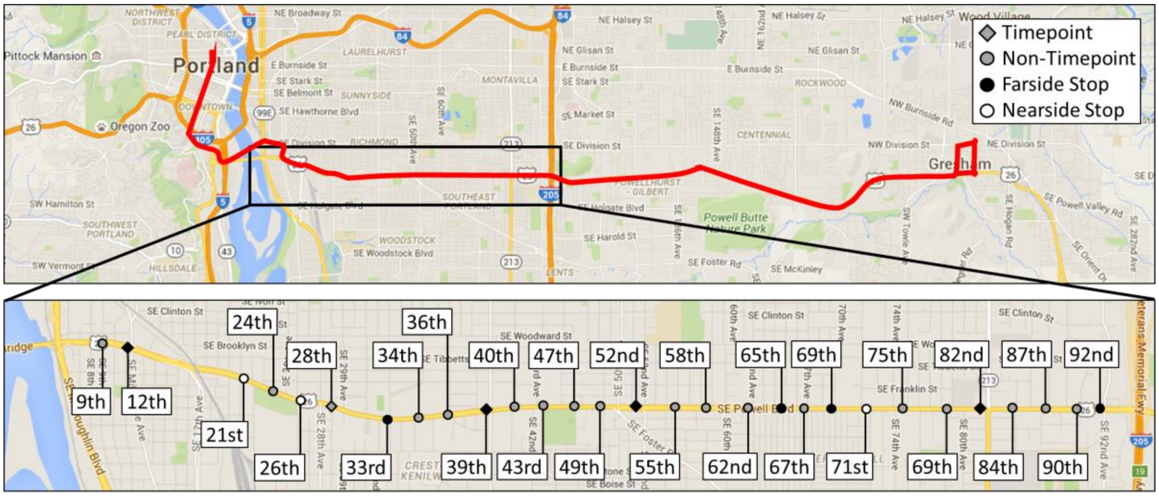

Figure 6-32 — Map of Route 9 on Powell Blvd showing locations for heat map in Figure 6-33 (Glick & Figliozzi, 2017).

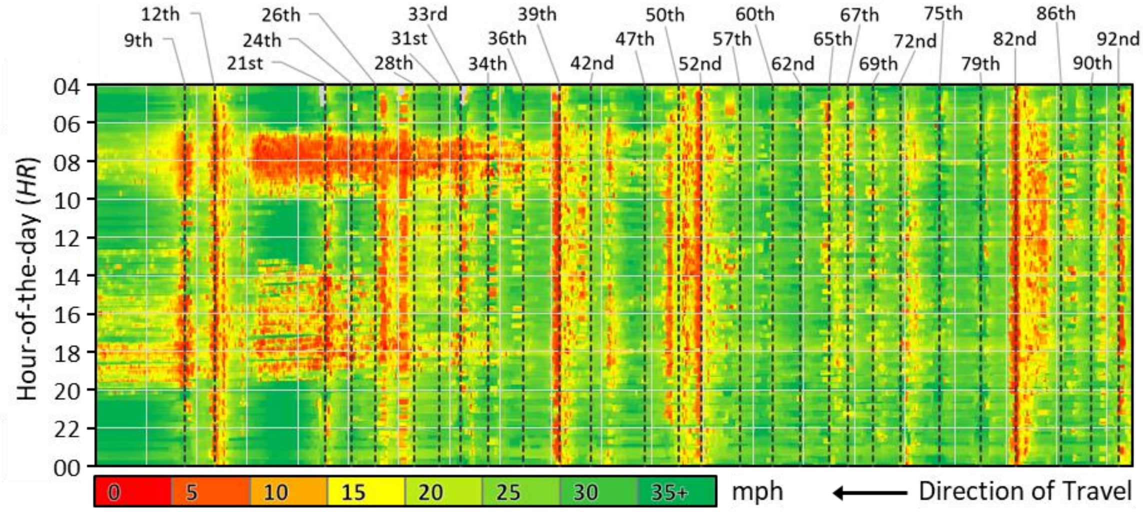

Figure 6-33 — Speed map of Route 9, by time of day, for segment of Powell Blvd identified by the map in Figure 6-32 (Glick & Figliozzi, 2017).

Using headways in an aggregated analysis is not intended to create heat maps, as it is not a high-resolution analysis, instead it serves to quickly identify where additional analysis could prove useful. The violin graphs can be produced from the aggregated data sets extremely quickly and require very little post processing. Headway performance metrics are therefore a useful visual tool, which may be augmented using other aggregated variables about passenger movements.

Congestion

Broken down by timepoint segments, the agency costs of congestion may also be examined. In Table 6-4, the average agency cost, resulting from congestion was shown to be $0.25 per boarding ride. That cost is slightly lower for inbound trips than for outbound trips at $0.23 and $0.26, respectively.

Inbound, the highest costs per boarding passenger are observed for the last segment (i.e. downtown) at $0.61. After that, both the first segment (i.e. Gresham Transit Center) and the segment around Powell & 82nd show costs at $0.31 per boarding passenger. The lowest costs are observed for SE Powell & 122nd at $0.07. Outbound trips also show high variability. The largest costs per boarding passenger is around the last stop (i.e. Gresham Transit Center) at $2.58 per boarding passenger. Yet, this value is slightly misleading. For inbound and outbound directions at that TPS, the total agency cost are 48,933 and $33,086, respectively, but these costs have vastly different boardings. Inbound there are an estimated 159,909 boardings, but just 12,813 outbound. The difference is primarily an effect of commuter patterns. Similarly, the opposite end of the transit line (i.e. the downtown segment) has an estimated 415,066 boardings outbound, but just 38,438 inbound. Given their respective total costs of 21,560 and 23,942, this equates to $0.05 and the $0.62 per boarding passenger.

6.5.3. Passenger Movements

Figure 6-34, Figure 6-35, Figure 6-36, and Figure 6-37 on the next two pages also show trends along route 9. The first two graphics show total boardings and total alightings within each timepoint segment for inbound and outbound service, respectively. The second two graphics show average boardings per vehicle and average alightings per vehicle for inbound and outbound service. From these plots, some of the reasons for increased uncertainty may be visualized.

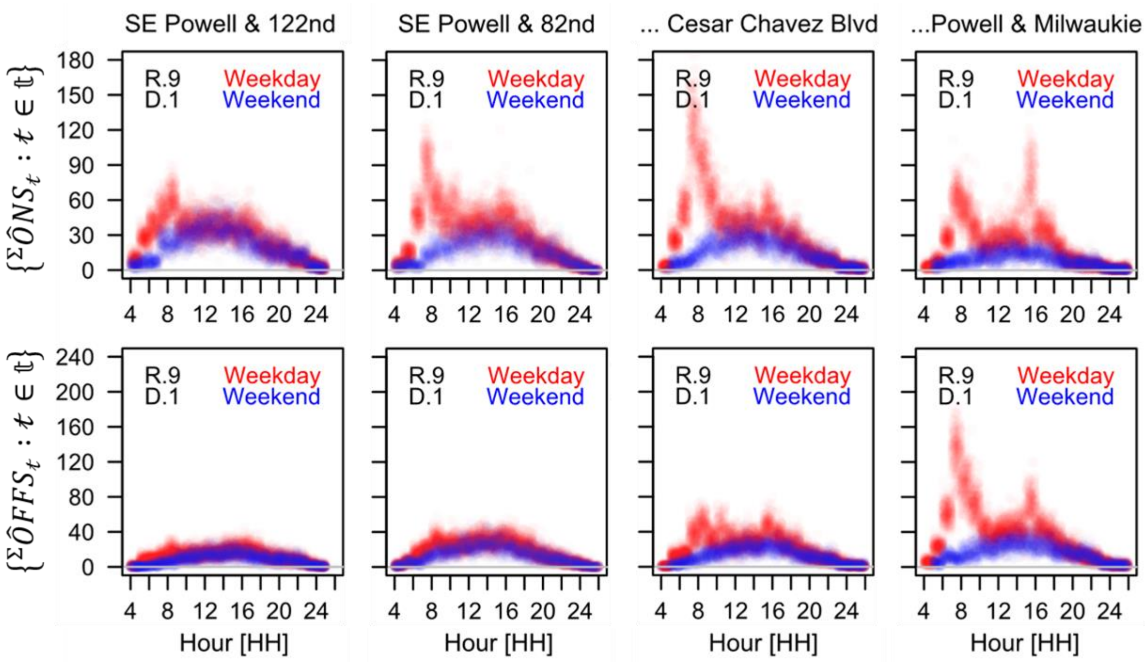

Figure 6-34 — Total boardings (top) and total alightings (bottom) for four TPS along Route 9, *inbound* to city-center. Direction of travel: left to right.

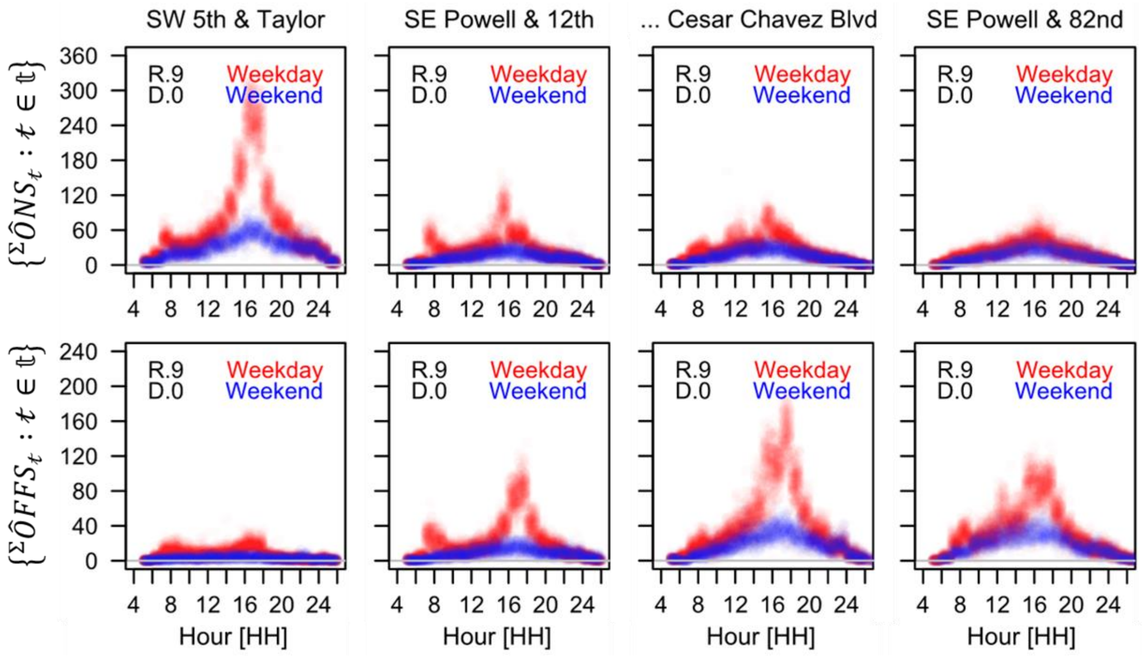

Figure 6-35 — Total boardings (top) and total alightings (bottom) for four TPS along Route 9, *outbound* to Gresham TC. Direction of travel: left to right.

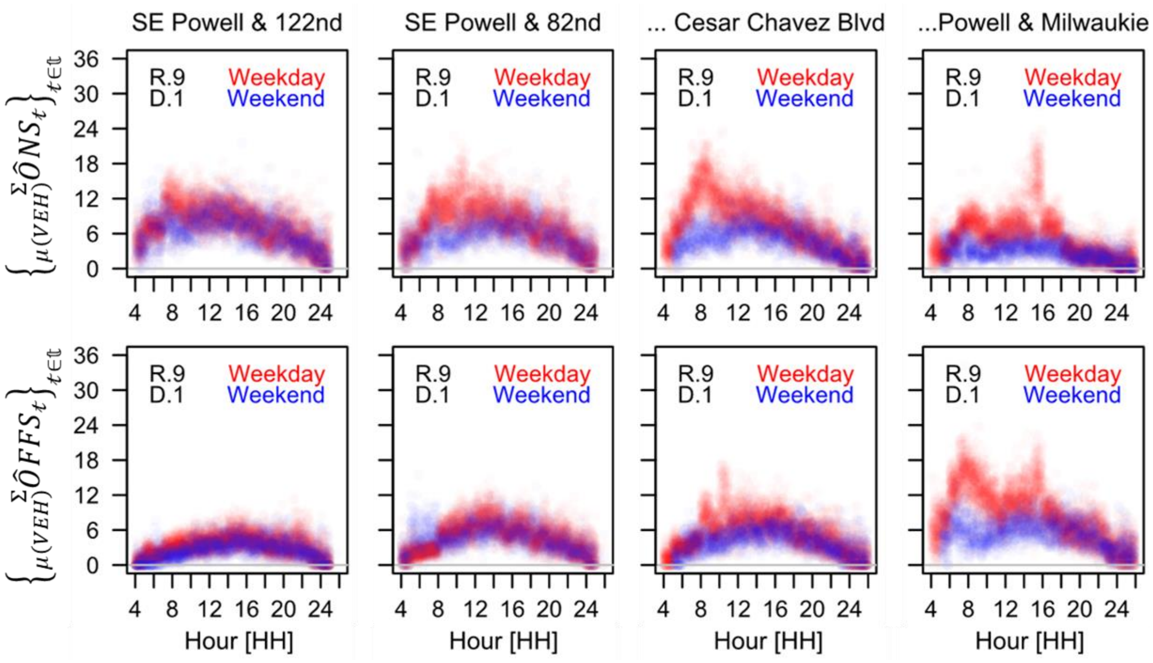

Figure 6-36 — Averages, per vehicle in TPS, for boardings (top) and alightings (bottom) for four TPS along Route 9, *inbound* to city-center. Direction of travel: left to right.

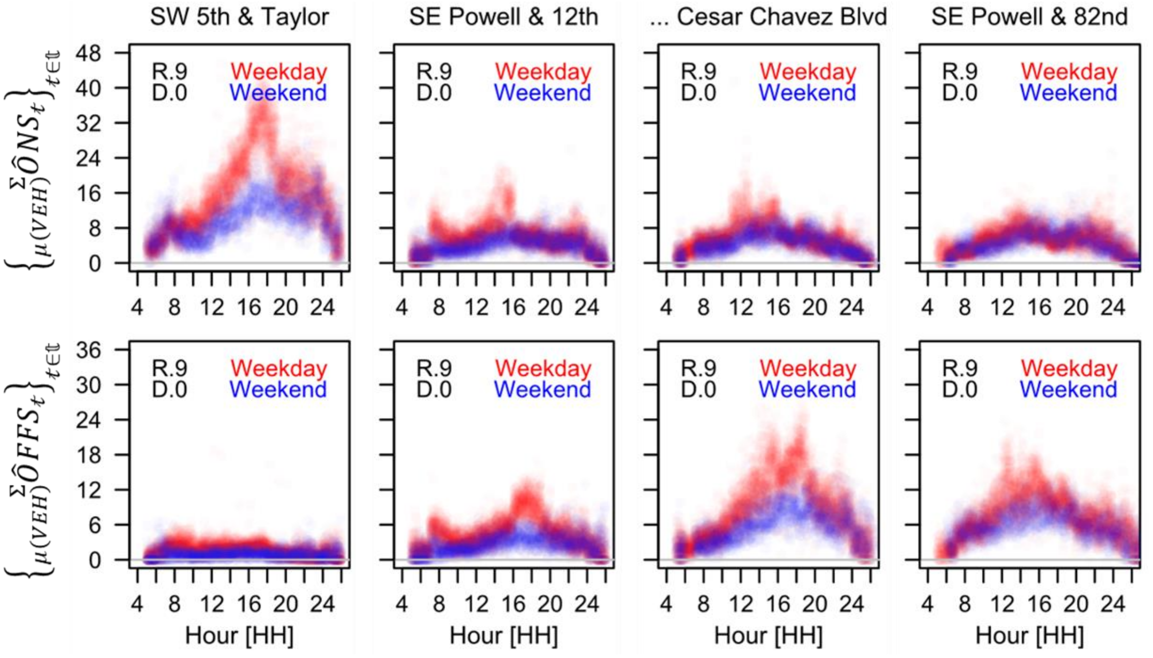

Figure 6-37 — Averages, per vehicle in TPS, for boardings (top) and alightings (bottom) for four TPS along Route 9, *outbound* to Gresham TC. Direction of travel: left to right.

The first, somewhat obvious observation, is the different demand for boardings versus alightings by time of day, which follow a commuter pattern. Second, is that the downtown transit mall and first timepoint segment on the eastside of the river account for a large percentage of boardings during the PM peak. Other segments for outbound service have few boardings but many alightings. While weekday passenger load increases along route 9 for inbound service (Figure 6-38), outbound service (Figure 6-39) starts with much more full vehicles, which are typically less full by the time they reach 82nd Ave.

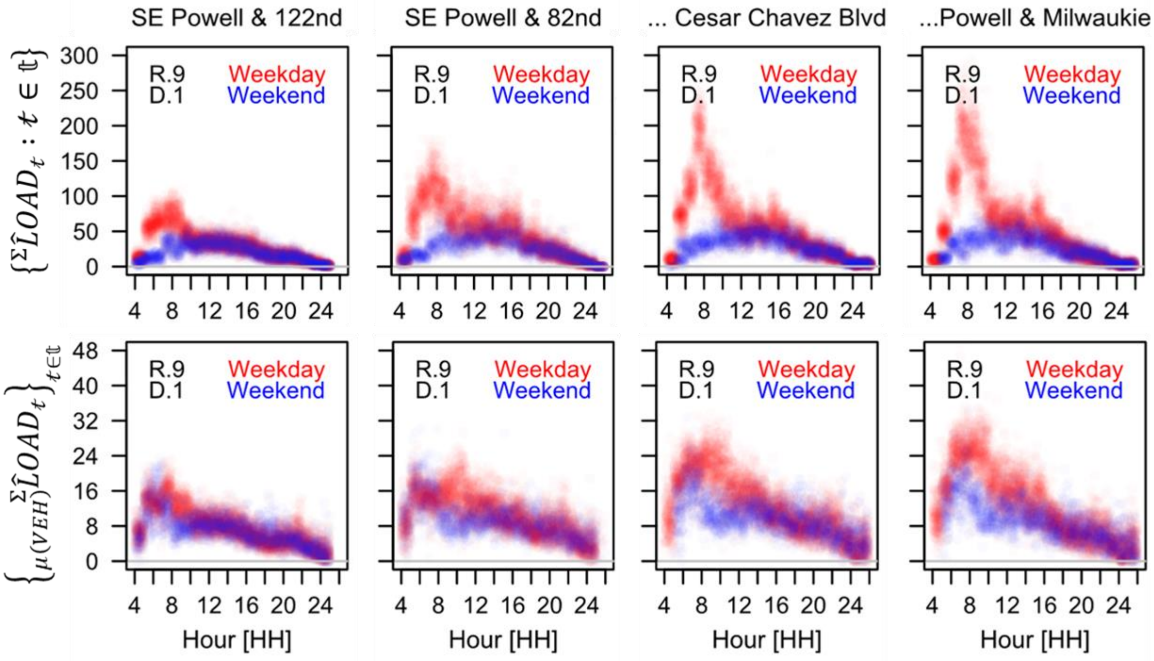

Figure 6-38 — Total estimated passenger load (top) and average passenger load (bottom), per vehicle in TPS, for four TPS along Route 9, *inbound* to city-center. Direction of travel: left-to-right.

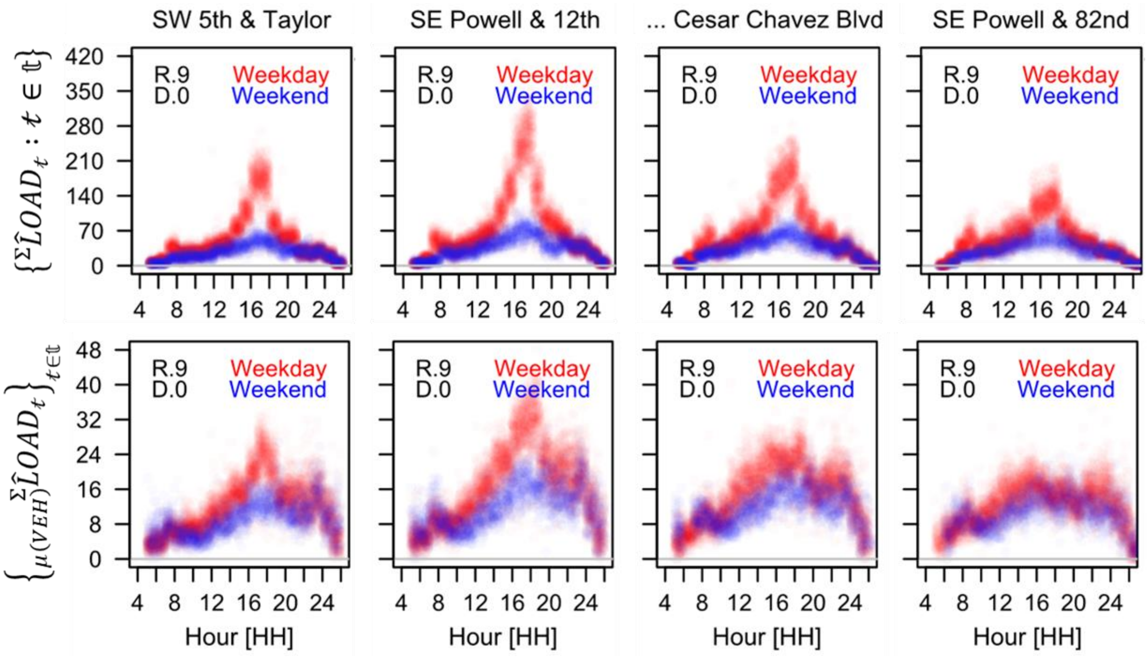

Figure 6-39 — Total estimated passenger load (top) and average passenger load (bottom), per vehicle in TPS, for four TPS along Route 9, *outbound* to Gresham Transit Center. Direction of travel: left-to-right.

6.6. Conclusion

To examine transit performance along a route, between routes, and for specific segments, the violin and IQR plots are potentially useful tool for researchers and agencies. They may be produced quickly and easily customized to examine specific locations, times, or feature sets. The methodologies are fast enough that computational burdens may be (mostly) ignored for average computers and may there be used to identify areas that require further study using higher resolution (microscopic) methodologies.

The quantitative analysis into congestion and speed both provided interesting preliminary results. Both methodologies showed the potential for useful performance metrics, but will require further research, which is discussed in Section 7.3.2.

CHAPTER 7 — CONCLUSIONS AND CONTRIBUTIONS

7.1. Overview

New technologies and the broadened availability of data and data collection systems have continued to influence how agencies, the public, and researchers understand and evaluate transit. Modern methodologies improve the decision-making process; yet, the increased amount of data results in a trade-off between the scope of analysis and the level of detail. This research focuses on balancing that tradeoff. It is therefore useful to return to the opening paragraph of Chapter 1, which frames the research problem:

> "Public transit routes comprise a network that serves multiple, and often conflicting, objectives: maximize ridership, provide fast and reliable travel times, increase accessibility for disadvantaged individuals and communities, and reduce costs. The realization of these objectives requires both a baseline understanding of the factors affecting each objective and, perhaps more importantly, tools that can help policy makers evaluate the tradeoffs between the objectives." (Page 1)

Chapter 1 further outlines the general pressures facing transit systems. In summary, transit systems have the potential to improve congestion, air quality, energy consumption, and safety; but, they must operate within a complex intersection of demographic trends, policy decisions, and economic forces. Transit agencies are themselves highly complex organizations that are typically slow to change practices, but are expected to provide an ever-improving level-of-service, while meeting new regulations, balancing revenues with operational costs, and maintaining transparency to governments and the public.

One of the primary constraints on transit operations is currently costs. As a broad overview of transit in the United States, the National Transit Database publishes timeseries data (National Transit Database, 2019) and a report of summaries and trends (National Transit Database, 2018). Figure 7-1 and Figure 7-2, show that operating expenses are trending upward at a slightly higher rate than vehicle revenue-hours and unlinked passenger trip (boarding ride).

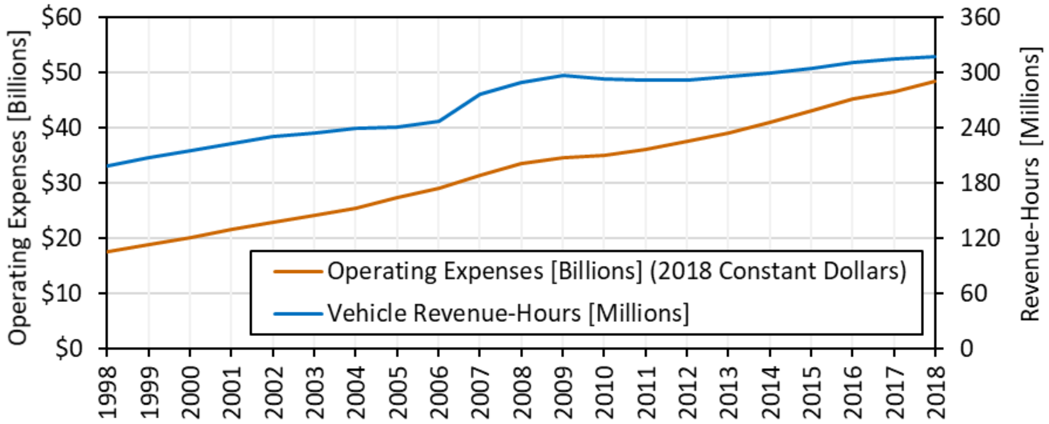

Figure 7-1 — Operating Expenses and Vehicle Revenue Hours: Time Series. Recreated Exhibit 1 from NTD 2018 Report (National Transit Database, 2018).

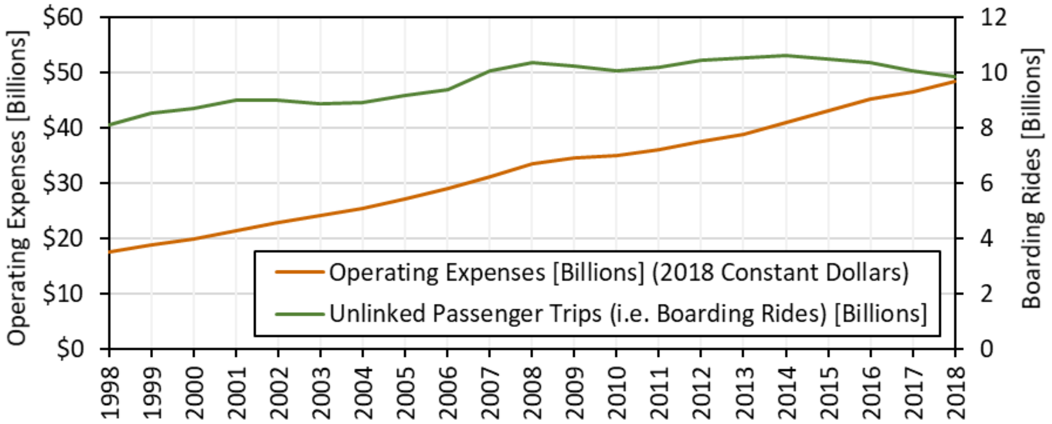

Figure 7-2 — Operating Expenses and Unlinked Passenger Trips (i.e. Boarding Rides): Time Series. Recreated Exhibit 2 from NTD 2018 Report.

Figure 7-3 shows the total operating expenses as a ratio to revenue-miles and boarding passengers. For the United States, the average operating cost per revenue-mile has increased by an average of $3.09 per year. Similarly, the operating cost per boarding ride has increased by an average of $0.13 per year. If the data set is restricted to just ten recent years (2009-2018), then the change in the number of riders has a negative, but statistically insignificant, trend. Yet, since 2012, the change in the number of riders per year has been negative and is statistically significant, averaging 10.8 million fewer riders per year. These national trends are also observed for TriMet (Figure 7-4).

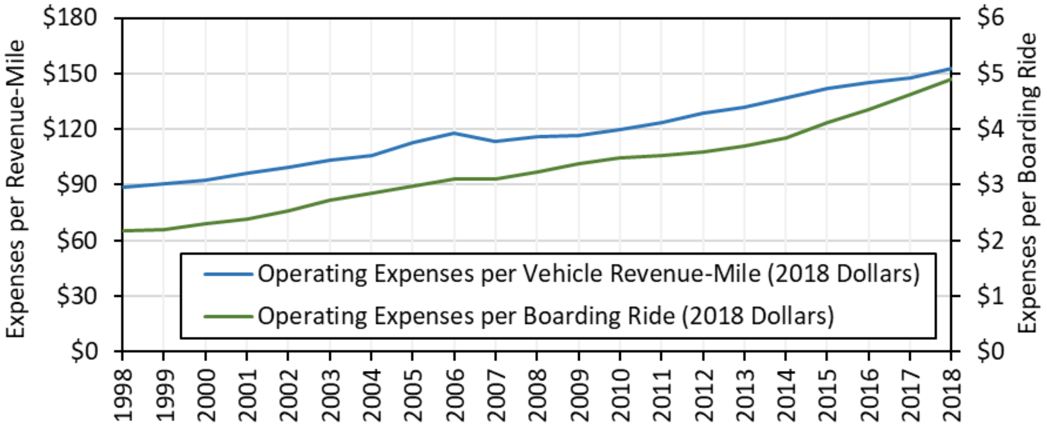

Figure 7-3 — Operating expenses per vehicle revenue-mile and per boarding ride: Time Series. Values in 2018 constant dollars.

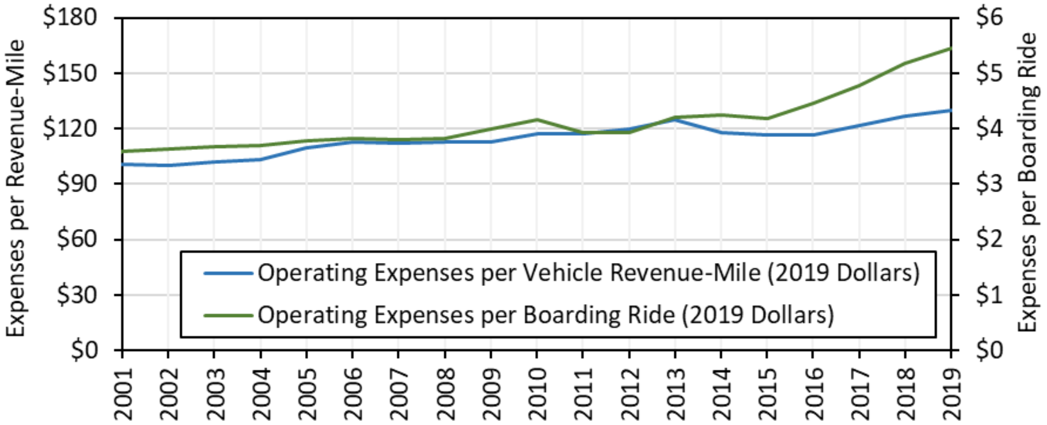

Figure 7-4 — Operating expenses per vehicle revenue-mile and per boarding ride: Time Series from TriMet. Values in 2019 constant dollars (TriMet, 2019).

The purpose of this research was not to answer the question of why ridership has been generally declining, but these trends and the other pressures facing transit systems are important to understanding why new performance metrics are needed. Beyond that, increasing operational costs provide context for why the new systems need to be based around existing data collection systems.

7.1.1. Motivation

As previously discussed, transit systems are expected to improve their level-of-service, despite other difficulties. As one example, lengthening transit lines and adding stops to match the new demands of urban sprawl can potentially reduce service attractiveness for existing passengers, if travel times or travel time uncertainty increases. As such, potential ridership gains by expanding service may be lost in other downstream locations. Unfortunately, the demographic and urbanization trends are likely to continue; therefore, service is likely to require expansions. Given the conflict between providing access and maintaining service quality, tools and methods are required to analyze, identify, and improve areas of existing service.

Current analysis methodologies typically examine performance at either a microscopic or macroscopic scale. The former focuses on specific locations, segments, transfer points, and transit trips. The latter examines performance over larger time periods, complete transit routes, or the network as a whole. If applied to larger systems, microscopic methodologies can often suffer from computational limitations, but macroscopic methods are often too coarse to identify hotspots or issues of specific areas. While both methods have useful applications, this research focuses on an alternative, middle approach.

7.2. Contributions

This mesoscopic methodology uses information identified from microscopic analyses to guide variable selection and to examine trends in aggregate. Higher resolution data is aggregated to reduce the computation burden, but maintains a sub-route level of detail for each hour of operation. The aggregated analysis reduces variability caused by singular atypical events, but still preserves enough detail for a detailed statistical analysis of routes, days, and times. Mesoscopic performance measures allow for segments to be studied either in the context of other segments or individually. Overall, the approach improves realism over previous macroscopic methods; thus, allowing for an evaluation of the key factors influencing transit operations and service variability.

7.2.1. Timepoint-Segments

A key contribution of this approach is the use of timepoint-segments, which are potentially a broadly applicable division of transit routes and transit systems. This potential for application is primarily a result of how fix-route transit is defined within the United States; specifically, with timepoint stops. Timepoint-segments are an application of an existing system for many agencies, therefore reducing the "cost" of entry for the methodologies outlined by this research.

7.2.2. Data Cleaning Methodology

The data sources for this research are additionally widespread and generally available to transit agencies, with the exception of high-resolution GPS data. While HRD is less available, it is not rare and it continues to grow in usage across agencies. In Chapter 3, a methodology for merging and cleaning the data sources is proposed that reduces reliance on many assumptions of previous studies. First, broken passenger counters are identified; these vehicles represent approximately one out of every eight vehicles on a given day. 80% of these vehicles showed zero passenger movements and would be traditionally excluded from analysis; however, 20% of these broken passenger counters represent vehicles that are recording data incorrectly (i.e. many more passenger boardings than alightings or the reverse). Previous methodologies would often allow for these counts because they are not necessarily outliers; yet, they represent observations that are not representative of actual operations.

Second, for many previous studies using stop event data, high-usage locations and many first or last stops are often excluded. While such locations do not typically behave like the rest of the transit network, this research provides a method to include these locations. First, probability distributions, representative of specific locations, times, and routes, are estimated then used to fill in missing or broken data stochastically. The replaced (i.e. “fixed”) values are probabilistically representative of a location, but not of that stop specifically. Using door open duration (i.e. ${}^{T}DWL$) as an example, both the mean and variance of the observed door open times increase when “fixed” data is included, as compared to the dataset excluding broken passenger counters and outliers. With the increased variance, created by including stochastically generated values, the proposed models are less likely to show statistically significant results than if problematic data was excluded completely.

Third, the cleaning methodology utilizes sufficient statistics, which allows for the entire dataset to be examined in parts while still representing the whole. As such, the probability distributions are not limited by the number of the datapoints that can be loaded into active RAM. Rather, the parameter estimation of the probability distributions can represent all available data, (mostly) without limitations of computer used to calculate the sufficient statistics. It should be noted that 12GB of RAM is at the lower end of system requirements. Luckily, typical office computers are increasingly configured with 16GB – 32GB of RAM as part of their default and cheaper systems.

7.2.3. Regressions

The stop level analysis required about 50 million data points to examine a year of non-zero archived datapoints. A computer with less than 12GB of RAM cannot load this data into an analysis program without excluding portions. The problem is further exaggerated when high-resolution data is included. Yet, the requirements for an aggregated data set are much lower, requiring 10x less RAM after processing. It should also be noted that the computer used for this research was capable of evaluating larger data-sets than those typically available by agencies. However, the purpose of that part of the analysis, specifically from Section 5.3, is not the focus of this research. Instead it serves as a baseline to show that the proposed methodologies can produce similar results with much reduced computational requirements.

In Chapter 5, multiple types of regression analysis were included and compared. The choices of which variables, and model types was largely based on previously published literature, but not entirely. In addition to a new analyses level, several updated classes of variables were included simultaneously in order to provide comparisons for stop types, traffic signals, vehicle interactions, and time-of-day. The coefficients of independent variables at the aggregated level, which are summations of stop-event variables, were expected to be similar. Mostly, this was true, but the differences highlight a benefit of the aggregated models: they can be more easily applied to total travel time than other data sources by examining moving duration and unplanned stops.

For every model, multiple divisions of the transit network were tested. In general, splitting the network into models specific to individual locations and times did not result in notable gains for the overall predictive power. One key takeaway is that additional complexity will not results in useful gains, most of the time. This is not to say that some models didn't perform extremely well. In particular, models focused on the downtown transit mall generally captured more variability than other areas. If an analysis is focused only on that area, or at a specific time, then specialized models may be useful. However, if only a small subset of the network is needed, then stop event data may be more appropriate for that application because the computation limitations are reduced.

7.2.4. Headways and Congestion

The visuals (i.e. violin and IQR plots) provide several useful way to examine transit performance along a route, between routes, and for specific segments. In particular, the violin plots may be produced quickly and customized to many different aspects of the transit system for segment, route, date, time, and system level analysis. The use of the adjusted deviation index allows for multiple routes to be compared simultaneously that have different scheduled headways and to see where routes are generally running with less variability than the schedule (i.e. values less than 0) or more (i.e. values greater than 0). As just one example, Figure 6-25 shows that headway variability is typically higher than the schedule during the middle of the day, but have the potential for less in early morning and later evening. Additionally, during the PM-peak, headway performance is shown to be highly variable with the majority of segments having much higher variability than the service schedule. The headways visuals, proposed in this research, are a tool to better identify where further evaluation is warranted.

7.3. Final Conclusions

The methodologies for data cleaning and results provide a foundation for how timepoint-segments and subsequent analyses may prove useful to researchers and agencies. The coefficients of regression results make sense in the context of previous methods using stop event data, while allowing for entire networks to be examined using ten times fewer data points. The visuals using headway performance metrics are potentially a tool for identifying areas that require a closer examination and for evaluating performance along routes. The methods are fast enough that computational considerations are reduced. More areas may be examined quickly before investing time in higher resolution methods.

7.3.1. Limitations

While timepoints are broadly used by agencies, the data analysis at the timepoint-segment level has a notable limitation. Specifically, where to define the divisions between consecutive timepoint-segments. This research defined the divisions only partially formulaically; some routes require manual separations. Identifying which routes would require manual definitions was not fully addressed; and therefore, an algorithmic way of defining all segments was not provided.

Another importantly limitation is the amount of estimation that was required for transit centers that were the first or last stop served. In many cases, the vehicles reported no passenger movements and no service durations. In particular, the Beaverton Transit Center reported the lowest usage of any transit center, despite being the most heavily used. The passenger estimates and the "fixed" data is based on the data points that remained, and resulted in reasonable estimates, when compared to officially reported values. Yet, an alternative approach for first and last stops, especially for those that are transit centers, may be warranted. Also relating to first and last stops, timepoint-segments at the beginning and end of transit lines often exhibited odd behaviors. These odd behaviors were one reason why a pseudo-timepoint was used for downtown Portland on Route 9, rather than the end of the line.

Areas of Improvement

With regards to timepoint-segments additional attention could have been given to the divisions of each route, beginning with first and last segments of routes. Odd behaviors lingered through the analysis that could not be fully explained without manually setting the divide for each route. Yet, an alternative approach could be employed, which is discussed in Section 7.3.2.

For regression modeling. Additional models, focused on high-usage and low-usage segments may have allowed for the issues of economies of scale to be addressed. In particular, the separation of high and low usage stops could allow for a more complete comparison of the square of the sum of passenger boardings and alightings versus the sum of the square.

The quantitative estimates of costs attributed to congestion would benefit from additional time and focus. The agency costs make sense and are reflected in the costs reported by TriMet. However, the passenger in-vehicle and waiting time estimates between the aggregated and stop-event level were inconsistent. The source of those inconstancies was examined and improved for most high-frequency routes, but not for low-frequency routes. There are two likely, and nonexclusive, sources of issue: first, some low-frequency routes did not include off-peak travel hours, which is required to estimate run times: and second, the stop-level methodology was originally tested on higher usage routes and may not have been properly calibrated to this dataset and to low usage locations. It would likely prove beneficial to utilize a third independent method of estimating passenger waiting times to better identify and correct for the discrepancies at the different analysis levels.

7.3.2. Future Research

In the future, the divisions between timepoint segments would be defined more precisely if only the trip patterns were considered, rather than the routes and directions. Trip patterns are unique to a single (unspecific) vehicle and may include multiple sequential routes of service, partial routes, and deadheads. With that change alone, some of the corrections for timepoint segments and routes, that needed to be manually entered, may be correct without adjustments. One example is for trips that loop without an official break between trips (e.g. entering then immediately exiting the downtown core). Officially, the route and direction of the trip change, but the vehicle does not behave as if two separate trips are occurring. Instead the single vehicle treats the outbound/inbound portion of the trip as a continuation of the inbound/outbound. Frequent or well-informed riders will sometimes board a vehicle, when it is still officially a different route and/or direction, knowing that particular vehicle will continue a pattern that includes the passenger's preferred route.

Trip patterns, as a column within the archived data, was not included directly for about half of the dataset. Using trip patterns would have required significant data processing to relate the identification numbers across trips. When breadcrumb data sets were included for later months, it became possible to directly link trip pattern numbers across all data sets. Additionally, updates the GTFS data improve cross compatibility. Future research would like benefit from using trip patterns for a few different reasons. First, trip pattern numbers are unique and not repeated. If a trip pattern changes, the previous number is discarded and a new number is used. Second, some routes use several trip patterns throughout a regular day, like Route 9. Route 9 operates two main patterns for each direction. Either vehicles begin/end at the Gresham Transit Center or at the Powell Garage at 99th. But, Route 9 is also part of larger patterns. Over a year on weekdays, Route 9 inbound was divided between two to ten patterns with 80% of days having five to seven unique patterns. Third, route numbers changes. A prominent example in this dataset was Route 4, which split into Routes 4 and 2 partway through the data set. For this change, a "fake" Route 3 was added to distinguish between route 4 from before versus after the change. With trip patterns, the differences would be captured automatically.

For regression modeling, models for total time, specific to each timepoint-segment along one transit line, could potentially provide insight into how operations change along a route. While the overall performance of the system was shown to be relative constant when multiple models were used, models along a route could prove a useful analysis tool to compare coefficient changes and identify hotspots. Changes and trends could be compared to the visual methods of the headway analysis for an additional quantitative approach. Finally, additional independent variables are known to influence transit operations. One example is transfer points, which were calculated within the dataset, but not utilized. The number of stops transferring to/from one stop from/to nearby stops, or between routes at the same stop are part of the GTFS datasets. These values are the same regardless of time of day. Important information may be gained if efforts are made to estimate the number of passengers transferring. In later GTFS files, the transfers are additionally indexed by trip pattern, in addition to route and direction.

Endnote

The original archived AVL/APC and GTFS datasets will be uploaded as a compressed archive for anyone to access in the future.

REFERENCES

80th Oregon Legislative Assembly, 2019. *House Bill 2001*, Salem, OR: s.n.

Albright, E. & Figliozzi, M., 2013. Factors Influencing Effectiveness of Transit Signal Priority and Late Bus Recovery at Signalized-Intersection Level. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 2311.

Bartholdi III, J. & Eisenstein, D., 2012. A Self_Coordinating Bus Route to Resist Bus Bunching. *Transportation Research Part B: Methodological,* 46(4), pp. 481-491.

Berkow, M., Wolfe, M., Monsere, C. & Bertini, R., 2008. Using Signal System Data and Buses as Probe Vehicles to Define the Congested Regime on Arterials. *Transportation Research Record: Journal of the Transportation Research Board,* pp. 35-45.

Bertini, R. & El-Geneidy, A. M., 2003. Generating Transit Performance Measures with Archived Data. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 1841, pp. 109-119.

Bertini, R. L. & El-Geneidy, A. M., 2004. Modeling Transit Trip Time Using Archived Bus Dispatch System Data. *Journal of Transportation Engineering,* 103(1), pp. 56-67.

Bertini, R. & Tantiyanugulchai, S., 2004. Transit Buses as Traffic Probes: Use of Geolocation Data for Empirical Evaluation. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 1870, pp. 35-45.

Bivand, R., Keitt, T. & Rowlingson, B., 2019. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. *R Package.*

Bureau of Planning and Sustainability, 2020. *Better Housing by Design - Adopted Staff Report*, City of Portland, OR: Denver Igarta.

Casella, G. & Berger, R. L., 2002. *Statistical Inference.* 2nd ed. Pacific Grover, CA: Thomson Learning Inc.

Cathey, F. & Dailey, D., 2002. Transit Vehicles as Traffic Probe Sensors. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 1804, pp. 23-30.

Chakroborty, P. & Kikuchi, S., 2004. Using Bus Travel Time Data to Estimate Travel Times on Urban Corridors. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 1870, pp. 18-25.

Cotugno, A. et al., 2008. *Metro Travel Forcasting 2008 Trip-Based Demand Model Methology Report,* s.l.: Planning Department - Transportation Research and Modeling Services.

Daganzo, C. F., 2009. A Headway-Based Approach to eliminate bus Bunching: Systematic Analysis and Comparsisons.. *Transportation Research Part B: methodological,* 43(10), pp. 319-921.

Delgado, F., Munoz, J. C. & Giesen, R., 2012. How Much Can Holding and/or Limiting Boarding Improve transit Performance?. *Transportation Research Part B: Methodological,* 46(9), pp. 1202-1217.

Diab, E., Bertini, R. L. & El-Geneidy, A., 2016. *Bus transit service reliability: Understanding the impacts of overlapping bus service on headway delays and determinants of bus bunching.* Washington D.C., Transportation Research Board Annual Meeting.

Dowle, M., Srinivasan, A. & al., e., 2019. *Package 'data.table',* s.l.: CRAN.

Dueker, K. J., Kimpel, T. J., Strathman, J. G. & Callas, S., 2004. Determinants of Bus Dwell Time. *Journal of Public Transportation,* 7(1), pp. 21-40.

El-Geneidy, A. M. & Vijayakumar, N., 2011. The Effects of Articulated Buses on Dwell and Running Times. *Journal of Public Transportation,* 14(3), pp. 63-86.

Feng, W., 2014. *Analyses of Bus Travel Time Reliability and Transit Signal Priority at the Stop-to-Stop Segment Level.* s.l.:Dissertation and Thesis.

Feng, W., Figliozzi, M. & Bertini, R., 2015. Quantifying the joint impacts of stop locations, signalized intersections, and traffic conditions on bus travel time. *Public Transport,* 7(3), pp. 391-408.

Fox, J. & Weisberg, S., 2018. *An {R} Companion to Applied Regression.* Third ed. Thousand Oaks, CA: Sage.

Fricker, J. D., 2011. *Bus Dwell Time Analysis Using Onboard Video..* Washington D.C., s.n.

FTA, Office of Planning & Environment, 2016. *Performance-Based Planning and Programming,* Washington, DC: s.n.

Furth, P. G., 2005. Sampling and Estimation Techniques for Estimated Bus System Passenger-Miles. *Journal of Transportation and Statistics,* 8(2), pp. 87-100.

Furth, P. G. & Halawani, A. T. M., 2018. How Well Does the Traffic System Protect Transit from Congestion? Measuring Route-Level Costs That Congestion Imposes on transit Operators and Users. *Journal of Advanced Transportation,* Volume 2018.

Furth, P. G. & Muller, T. H. J., 2006. Service Reliability and Hidden Waiting Time: Insights from Automatic Vehicle Location Data. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 1955, pp. 79-87.

Glick, T. B., Feng, W., Bertini, R. L. & Figliozzi, M. A., 2015. Exploring Applications of Second-Generation Archived Transit Data for Estimating Performance Measures and Arterial Travel Speeds. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 2538, pp. 44-53.

Glick, T. B. & Figliozzi, M. A., 2017. Measuring the Determinants of Bus Dwell Time. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 2647, pp. 109-117.

Glick, T. B. & Figliozzi, M. A., 2018. Evaluation of Route Changes Utilizing High-Resolution GPS Bus Transit Data. *Transportation Research Record: Journal of the Transportation Research Board,* 2672(8), pp. 199-209.

Glick, T. B. & Figliozzi, M. A., 2019. Analysis and Application of Log-Linear and Quantile Regression Models to Predict Bus Dwell Time. *Transportation Research Record: Journal of the Transportation Research Board,* 2673(10), pp. 118-128.

Glick, T. B. & Figliozzi, M. A., 2019. *Analyzing the Impact of Bus Stop Queuing and Bus Interactions on Bus Dwell Times.* Washington D.C., Transportation Research Board.

Glick, T. B., Figliozzi, M. A. & Crumley, M., 2020. Examining the Impact of Overlapping Bus Service on Dwell Times and Bunching. *Manuscript.*

Glick, T. & Figliozzi, M., 2017. Traffic and Transit Travel Time Reliability Indexes and Confidence Intervals: Novel Methodologies for the Corridor and Segment Level. *Transportation Reseserach Record: Journal of the Transportation Research Board,* 2649(1), pp. 28-41.

González, E. M., Romana, M. G. & Álvaro, O., 2012. Bus Dwell-Time Model of Main Urban Route Stops: Case Study in Madrid, Spain. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 2274, pp. 126-134.

Google Developers, 2019. *Google Transit APIs > Static Transit > Reference.* [Online] Available at: https://developers.google.com/transit/gtfs/reference/ [Accessed December 2019].

Graff, G., 2019. *SW Madison Bus/Bike Lane Project,* Portland: s.n.

Grolemund, G. & Wickham, H., 2011. Dates and Times Made Easy with {lubridate}. *Journal of Statistical Software,* 40(3), pp. 1-25.

Grömping, U., 2006. Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software*, 17(1), pp. 1-27.

Güneralp, B. et al., 2017. Global scenarios of urban density and its impacts on building energy use through 2050. *Proceedings of the National Academy of Sciences,* 114(34), pp. 8945-8950.

Hall, R. & Vyas, N., 2000. Buses as a Traffic Probe: Demonstration Project. *Transportation Research Record: Journal of the Transportation Research Board,* 1731(1), pp. 96-103.

Kate Brown, 2020. *State of Oregon, Updated Mitigation Measures on Coronavirus Response,* Salem: Office of the Governor.

Keeling, K. L., Glick, T. B., Crumley, M. & figliozzi, M. A., 2019. Evaluation of Bus-Bicycle and Bus/Right-Turn Traffic Delays and Conflicts. *Transportation Research Record: Journal of the Transportation Research Board,* 2673(7), pp. 443-453.

Kittelson & Associates, et. al. , 2013. *Transit Capacity and Quality of Service Manual, Third Edition.* Washington D.C.: The National Academies Press.

Klik, M., Collet, Y. & Facebook, 2019. *Package 'fst',* s.l.: CRAN.

Lee, Y.-J. & Vuchic, V. R., 2005. Transit Network Design with Variable Demand. *Journal of Transportation Engineering,* 131(1).

Levinson, H., 1983. Analyzing Transit Travel Time Performance. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 915, pp. 1-6.

Li, F., Duan, Z. & Yang, D., 2012. Dwell Time Estimation Models for Bus Rapid Transit Stations. *Journal of Modern Transportation,* 20(3), pp. 168-177.

Mandl, C. E., 1980. Evaluation and Optimization of Urban Public Trnsportation Networks. *European Journal of Operational Research,* 5(6), pp. 396-404.

Ma, X. & Wang, Y., 2014. Development of a Data-Driven Platform for Transit Performance Measures Using Smart Card and GPS Data. *Journal of Transportation Engineering*, 140(12).

Metro, 2020. *Metro and Covid-19: Setps we've taken.* [Online]
Available at: https://www.wmata.com/service/covid19/COVID-19.cfm
[Accessed July 2020].

Milkovits, M. N., 2008. Modeling the Factors Affecting Bus Stop Dwell Time: Use of Automatic Passenger Counting, Automatic Fare Counting, and Automatic Location Data. *Transportation Research Record: Journal of the Transportation Research Board*, Volume 2072, pp. 125-130.

MobilityData, 2019. *General Transit Feed Specification Reference.* [Online]
Available at: https://gtfs.org/reference/static
[Accessed 14 December 2019].

Moore, A., Figliozzi, M. & Bigazzi, A., 2014. Modeling the Impact of Traffic Conditions on the Variability in Midblock PM2.5 Urban Arterial Concentrations. *Transporation Research Record: Journal of the Transportation Research Board*, Volume 2428, pp. 35-43.

Moore, A., Figliozzi, M. & Monsere, C., 2012. Bus Stop Air Quality: An Empirical Analysis of Exposure to Particulate Matter at Bus Stop Shelters. *Transportation Research Record: Journal of the Transportation Research Board* , Volume 2270, pp. 78-86.

Nason, N. R. & Williams, K. J., 2019. *Status of the Nation's Highways, Bridges, and Transit: Conditions and Performance 23rd Edition: Report to Congress,* Washington, DC: Federal Highway Administration.

National Transit Database, 2018. *National Transit Summaries and Trends,* s.l.: Federal Transit Administration.

National Transit Database, 2019. *TS2.1 - Service Data and Operating Expenses Time-Series by Mode.* [Online]
Available at: https://www.transit.dot.gov/ntd/data-product/ts21-service-data-and-operating-expenses-time-series-mode-2
[Accessed October 2020].

Noch, M., 2019. *Resources for Practitioners - ITS Transit Fact Sheets.* [Online]
Available at: https://www.pcb.its.dot.gov/factsheets/avl/avl_overview.aspx
[Accessed 2019].

OpenMobilityData, 2019. *TriMet GTFS.* [Online]
Available at: https://transitfeeds.com/p/trimet/43
[Accessed November 2019].

Palma, A. d. & Lindsey, R., 2001. Optimal Timetables for Public Transportation. *Transportation Research Part B: Methodological,* 35(8), pp. 789-813.

PB Farradyne Inc.; Battelle, 2001. *Tri-Met 5-Year Intelligent Transportation System Plan,* Portland, OR: TriMet.

PBOT, 2019. *Draft Rose Lane Vision Open House,* Portland: City of Portland.

Pebesma, E. J. & Bivand, R. S., 2005. Classes and methods for spatial data in R. *R News*, 5(2), pp. 9-13.

ProofWiki contributors, 2020. *Definition: Indexing Set.* [Online]
Available at:
https://proofwiki.org/w/index.php?title=Definition:Indexing_Set&oldid=468930
[Accessed September 2020].

ProofWiki contributors, 2020. *Definition: Indexing Set/Function.* [Online]
Available at:
https://proofwiki.org/w/index.php?title=Definition:Indexing_Set/Function&oldid=414915
[Accessed September 2020].

ProofWiki contributors, 2020. *Definition: Set.* [Online]
Available at:
https://proofwiki.org/w/index.php?title=Definition:Set&oldid=490546
[Accessed September 2020].

ProofWiki contributors, 2020. *Definition: Set/Definition by Predicate.* [Online]
Available at:
https://proofwiki.org/w/index.php?title=Definition:Set/Definition_by_Predicate&oldid=491369
[Accessed September 2020].

Rose, J., 2009. *The educated commuter: How TriMet tallies MAX boardings.* [Online]
Available at:
https://www.oregonlive.com/commuting/2009/09/the_educated_commuter_how_trim.html
[Accessed February 2020].

RStudio Team, 2019. *RStudio: Integrated Development for R.*. Boston, MA: s.n.

Shyr, O. F., Andersson, D. E., Cheng, Y.-H. & Hsiao, Y.-H., 2017. What explains rapid transit use? Evidence from 97 urbanized areas. *Transportation Research Part A: Policy and Practice,* Volume 100, pp. 162-169.

Siddiqui, F., 2018. *Falling transit ridership poses an 'emergency' for cities, experts fear.* [Online]
Available at: https://www.washingtonpost.com/local/trafficandcommuting/falling-transit-ridership-poses-an-emergency-for-cities-experts-fear/2018/03/20/ffb67c28-2865-11e8-874b-d517e912f125_story.html
[Accessed 2020].

Simon, J. & Furth, P. G., 1985. Generating a Bus Route O-D Matrix from On-Off Data. *Journal of Transportation Engineering,* 111(6), pp. 583-593.

Skabardonis, A. & Geroliminis, N., 2005. *Real-time Estimation of Travel Times on Signalized Arterials.* University of Maryland, s.n.

Slavin, C., Feng, W. & Figliozzi, M., 2012. *An Evaluation of the Impacts of an Adaptive Coordinated Traffic Signal System on Transit Performance: a Case Study on Powell Boulevard.* Santiago, Chile, s.n.

Slavin, C. & Figliozzi, M., 2011. *Air Quality and Multimodal Evaluation of an Adaptive Traffic Signal System: A Case Study on Powell Boulevard.* s.l., s.n.

Stoll, N., 2016. *Potential of Using High Resolution Bus GPS Data to Assess Traffic Speeds,* s.l.: PDXScholar.

Stoll, N. B., Glick, T. B. & & Figliozzi, M. A., 2016. Using High-Resolution Bus GPS Data to Visualize and Identify Congestion Hot Spots in Urban Arterials. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 2539, pp. 20-29.

Strathman, J., Kimpel, T. J. & Callas, S., 2005. *Rail APC Validation and Sampling for NTD and Internal Reporting at TriMet*, Portland, OR: TriMet.

Sun, L. et al., 2014. Models of Bus Boarding and Alighting Dynamics. *Transportation Research Part A: Policy and Practice*, Volume 69, pp. 447-460.

Tirachini, A., 2013. Estimation of Travel Time and the Benefits of Upgrading the Fare Payment Technology in Urban Bus Services. *Transportation Research Part C: Emerging Technologies*, Volume 30, pp. 239-256.

Transit Cooperative Research Program, 2013. *Report 165: Transit Capacity and Quality of Service Manual, Third Edition,* Washington, D.C.: s.n.

Transit, 2020. *How coronavirus is disrupting public transit.* [Online]
Available at: https://transitapp.com/coronavirus
[Accessed July 2020].

TriMet Developer Resources, 2019. *Unofficial/Proposed GTFS Data Elements.* [Online]
Available at: https://developer.trimet.org/gtfs_ext.shtml
[Accessed December 2019].

TriMet, 2019. *Approved Budget,* Portland, OR: Budget and Grants Administration Department.

TriMet, 2019. *Route Ridership Report: Weekdays,* Portland, OR: s.n.

TriMet, 2019. *Title VI Program Update,* Portland, OR: TriMet.

TriMet, 2019. *TriMet Service and Ridership Information,* Portland, OR: s.n.

TriMet, 2020. *Ridership and Performance Statistics - COVID-19 Ridership Impact.* [Online]
Available at: https://trimet.org/about/performance.htm#weekly
[Accessed July 2020].

TriMet, 2020. *Studies and Surveys.* [Online]
Available at: https://trimet.org/research/#onboard
[Accessed February 2020].

TriMet, 2020. *TriMet City Center Map.* [Online]
Available at: https://trimet.org/maps/img/citycenter.png
[Accessed November 2020].

TriMet, 2020. *TriMet Interactive map.* [Online]
Available at: https://ride.trimet.org/
[Accessed November 2020].

TriMet, 2020. *TriMet System Map.* [Online]
Available at: https://trimet.org/maps/img/trimetsystem.png
[Accessed November 2020].

TriMet, 2020. *TriMet System Maps.* [Online]
Available at: https://trimet.org/maps/img/trimetsystem.png
[Accessed 2020].

United Nations, 2019. *World Urbanization Prospects 2018: Highlights,* New York, NY: Population Division.

United States Census Bureau, 2020. *Means of Transportation to Work.* [Online]
Available at: https://data.census.gov/
[Accessed 2020].

Wikipedia contributors, 2020. *Algebra of sets,* s.l.: Wikipedia, The Free Encyclopedia.

Wikipedia contributors, 2020. *Algebra of sets,* s.l.: Wikipedia, The Free Encyclopedia.

Wikipedia contributors, 2020. *Pacific Time Zone,* s.l.: Wikipedia, The Free Encyclopedia.

Wikipedia contributors, 2020. *Partition of a set,* s.l.: Wikipedia, The Free Encyclopedia.

Wikipedia contributors, 2020. *Set (mathematics),* s.l.: Wikipedia, The Free Encyclopedia.

Wikipedia contributors, 2020. *Set-builder notation,* s.l.: Wikipedia, The Free Encyclopedia.

Worldometers.info, 2020. *United States Population.* [Online]
Available at: https://www.worldometers.info/world-population/us-population/
[Accessed 2020].

Yan, Y., Liu, Z., Meng, Q. & Jiang, Y., 2013. Robust Optimization Model of Bus transit Network Design with Stochastic Travel Time. *Journal of Transportation Engineering,* 139(6).

York, T., 2019. *Bus lanes speed up trips on 10 TriMet lines.* [Online]
Available at: https://news.trimet.org/2019/11/bus-lanes-speed-up-trips-on-10-trimet-lines/
[Accessed October 2020].

Zeileis, A. & Grothendieck, G., 2005. zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software,* 14(6), pp. 1-27.

APPENDIX A — NOTATION AND DEFINITIONS

A.1. Introduction

The notation used throughout this dissertation is a non-typical variant of set-builder notation (Wikipedia contributors, 2020) (ProofWiki contributors, 2020), which relies heavily on indexed families (Wikipedia contributors, 2020), indexing sets (ProofWiki contributors, 2020), indexing functions (ProofWiki contributors, 2020), and predicated logic (ProofWiki contributors, 2020). The choice of this notation was a compromise between necessary complexity and readability.

A.2. Set-Builder Notation

Set-builder notation is used extensively to differentiate between individual values and collections of values. Without any additional notation (e.g. superscripts, subscripts, accents, etc.), VAR, the placeholder variable defined in Definition 3-1 represents any single value that is contained in ${}^{VAR}\Omega$, the set of all possible values for a single observation (i.e. the sample space) of VAR.

Definition A-1 — ${}^{VAR}\Omega$ is the *Sample Space* for a single value of VAR. ${}^{VAR}\Omega$ contains all possible values for a single element VAR.

VAR has the same properties as observational (i.e. recorded) data of the same name; but, does not refer to a specific record from, nor is VAR contained in the dataset. The complete set of observed values is denoted ${}^{VAR}\mathcal{S}$ and individual values (i.e. $\{VAR_1, \dots, VAR_n\}$) are specified using indexes. Unless otherwise noted, the index i is used to index all observations in ${}^{VAR}\mathcal{S}$, where i is contained in the index set, $I = \{1, \dots, n\}$. VAR_i

is therefore any single value contained in ${}^{VAR}\mathcal{S}$. In equation (A.1), which shows several equivalent notations for ${}^{VAR}\mathcal{S}$, $i \in I$ is a domain that defines which VAR_i are included in the set.

Definition A-2 — ${}^{VAR}\mathcal{S}$ is the *Complete Set* observed values recorded in the dataset.

(A.1) $${}^{VAR}\mathcal{S} = \{VAR_1, \dots, VAR_n\} = \{VAR_i : i \in I\} = \{VAR_i\}_{i \in I} \ ,$$

For: $I = \{i : (i \in \mathbb{N}_1) \wedge (i \leq n)\}$.

Additionally, equation (A.1) also uses the logical "and" operator (i.e. "$\wedge$") and the colon symbol ":" (see Definition A-3), and the number set $\mathbb{N}_1$ (see Definition A-4 and equation (A.2). For the combination of negation and a logical "and" or "or" conjunctions, $\neg(A \wedge B) = (\neg A \vee \neg B)$ and $\neg(A \vee B) = (\neg A \wedge \neg B)$.

Definition A-3 — "$\neg$", "$\wedge$", "$\vee$", "$\Rightarrow$", and ":" are *Logical Operators* meaning *not*, *and*, *or*, *if-then*, and *such-that*. For ":", $\{A : B\}$ is defined as the set of all *A such-that B* is true. The order of operations is to evaluate parenthesis, "$\neg$", "$\wedge$" and "$\vee$", quantifiers, then conditionals. A truth table for each operation are given in Table A-1.

Table A-1 — Truth table for logical operators.

		Negation		***And***	***Or***	***If-Then***		***Such-That***
A	***B***	$\neg A$	$\neg B$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$	$A \Leftarrow B$	$A : B$
T	**T**	F	F	T	T	T	T	T
T	**F**	F	T	F	T	F	T	F
F	**T**	T	F	F	T	T	F	F
F	**F**	T	T	F	F	T	T	T

Definition A-4 — $\mathbb{B}$, $\mathbb{N}_1$, $\mathbb{N}_0$, $\mathbb{Z}$, and $\mathbb{R}$, are *Number Sets*. $\mathbb{B}$ is the binary set containing $\{0, 1\}$. $\mathbb{N}_0$ and $\mathbb{N}_1$ contain non-negative and positive integers, respectively. $\mathbb{Z}$ contains all integers. $\mathbb{R}$ is the set of all real numbers (i.e. any non-infinite quantity that can be represented as an infinite decimal expansion).

(A.2) $$\left.\begin{matrix}\{0,1\} = \mathbb{B} \\ \{1,2,3\ldots\} = \mathbb{N}_1\end{matrix}\right\} \subset \overbrace{\mathbb{N}_0 = \{0\} \cup \mathbb{N}_1}^{\{0,1,2,3,\ldots\}} \subset \overbrace{\mathbb{Z} = -\mathbb{N}_1 \cup \mathbb{N}_0}^{\{\ldots,-2,-1,0,1,2,\ldots\}} \subset \mathbb{R}\ ,$$

Given that: $\forall x \in \mathbb{R}, \big((-\infty < x < \infty) \wedge (x \neq \emptyset)\big)$; $kA = \{ka : a \in A\}$ if k is a constant and A is a Set*; and*, $\emptyset$ denotes a "Null" (i.e. missing) or "Na" (i.e. not applicable) value.

A.2.1. Partitions

Often, it will prove useful to partition the index set I depending on each values of VAR_i (Wikipedia contributors, 2020). Each i will belong to one of three partitions: I_0 (i.e. the zero partition), J (i.e. the non-zero partition), or K (i.e. the non-real partition), which contains values that are infinite, *Null*, or *Na*. There exists no i belonging to the intersection of any two partitions; any such intersection would result in { }, an empty set. The union of I_0, J, and K will be the complete index set I. Equation (A.3) defines the partition conditions using ϕ, a placeholder variable for a *True* logical statement, which is defined below equation. In equation (A.3), it is also possible for one or two partitions to empty sets, but not all three, assuming $I \neq \{\,\}$. Each partition is defined as a non-strict subsets of I. For such subsets, their superset (i.e. I) does not need to be included in notation, as it is implied. As an example, $\{VAR_i : i \in J\}_{i\in I}$ includes redundant information and may be simplified to $\{VAR_i\}_{i\in J}$ or $\{VAR_i : i \in J\}$.

(A.3) $$\forall i \in I, \left(i \in \begin{Bmatrix} I_0, & \text{if}\, VAR_i = 0 \\ J, & \text{if}\, VAR_i \in \mathbb{R}\backslash\{0\} \\ K, & \text{if}\, VAR_i \notin \mathbb{R} \end{Bmatrix} : \phi\right)\ ,$$

Given that: $\phi \coloneqq (I_0 \cup J \cup K = I) \wedge (I_0 \cap J = I_0 \cap K = J \cap K = \{\,\})$; *and,* $\mathbb{R}\backslash\{0\}$ is the "Set Difference" between $\mathbb{R}$ *and* $\{0\}$, *such that:* if A and B are sets, then $A\backslash B = \{x \in A : x \notin B\}$.

Partition index sets (i.e. I_0, J, and K) are defined by the variable or function in the brackets. For example, if $x \in X = \{2, 1, 0, \emptyset\}$, $y \in Y = \{1, 0, 0, 3\}$, and $z \in Z = X + Y$, then equation (A.4) shows the partitions for sets, X, Y, and Z. For Z, it is important that the number of elements in X and Y is the same (Wikipedia contributors, 2020).

$$\begin{array}{ccc} \{x_i\}_{i \in I_0} = \{x_3\} & \{y\}_{i \in I_0} = \{y_2, y_3\} & \{z_i\}_{i \in I_0} = \{z_3\} \\ \{x_i\}_{i \in J} = \{x_1, x_2\} & \{y_i\}_{i \in J} = \{y_1, y_4\} & \{z_i\}_{i \in J} = \{z_1, z_2\} , \\ \{x_i\}_{i \in K} = \{x_4\} & \{y_i\}_{i \in K} = \{\,\} & \{z_i\}_{i \in K} = \{z_4\} \end{array} \tag{A.4}$$

given that: $\forall z \in Z, \big(z_i = x_i + y_i : (x \in X) \wedge (y \in Y)\big) \therefore Z = \{3, 1, 0, \emptyset\}$.

The non-real partition, K, is not explicitly part of most function for this dissertation. However, codes often required explicit instructions on how to respond to missing, null, infinite, or otherwise non-real data values.

A.2.2. Subsets

Throughout this research, specified indices can also denote a subset for another index set; primarily, script lower-case letters will be used. For example, if $\mathscr{s}$ is defined to be a subset of J, then $\{VAR_i : i \in \mathscr{s}\}$ is a subset of $\{VAR_i : i \in J\}$. Index membership in $\mathscr{s}$ may be defined explicitly or conditionally using $\Phi_{\mathscr{s} \subseteq J}(\,\cdots)$ (i.e. a conditional predicate function with defined parameters) (ProofWiki contributors, 2020). In equation (A.5), VAR_i is included in the set if, and only if, the predicate is true; and (if true), the index i is defined as a member of $\mathscr{s}$.

$$\begin{aligned} \{VAR_i\}_{i \in \mathscr{s}} &= \{VAR_i : i \in \mathscr{s}\} \\ &= \big\{VAR_i : \Phi_{\mathscr{s} \subseteq J}(\,\cdots)\big\} \\ &= \{VAR_i : \Phi_{\mathscr{s}}(\,\cdots) \wedge (i \in J)\} \\ &= \{VAR_i : \Phi_{\mathscr{s}}(\,\cdots)\}_{i \in J} \end{aligned} \tag{A.5}$$

Where possible, additional parameters will be added to predicate functions reduce the number of indices need and keep notation simple. However, several subsets will sometimes be needed. For these cases, all necessary indices will be listed. If $\mathscr{a}$ and $\mathscr{b}$ are independently defined subsets of J, then $\{VAR_i : i \in (\mathscr{a} \cap \mathscr{b})\}$ contains all VAR_i for which $i \in J$ and both $\Phi_{\mathscr{a}}(\cdots)$ and $\Phi_{\mathscr{b}}(\cdots)$ are true for their parameters.

$$\begin{aligned} \{VAR_i\}_{i \in (\mathscr{a} \cap \mathscr{b})} &= \{VAR_i : i \in (\mathscr{a} \cap \mathscr{b})\} \\ &= \left\{VAR_i : \Phi_{\mathscr{a} \subseteq J}(\cdots) \wedge \Phi_{\mathscr{b} \subseteq J}(\cdots)\right\} \\ &= \{VAR_i : \Phi_{\mathscr{a}}(\cdots) \wedge \Phi_{\mathscr{b}}(\cdots) \wedge (i \in J)\} \\ &= \{VAR_i : \Phi_{\mathscr{a}}(\cdots) \wedge \Phi_{\mathscr{b}}(\cdots)\}_{i \in J} \end{aligned} \tag{A.6}$$

When multiple subsets are needed, then a double-struck lowercase letter is used to denote a family of subsets. For example, $\mathbb{s}$ may be defined as the family containing $\{\mathscr{s}_1, \dots, \mathscr{s}_n\}$, where each element in $\mathbb{s}$ is defined as subset of J. In this case, $\bigcup \mathbb{s}$ is the union (i.e. $\bigcup \mathbb{s} = \mathscr{s}_1 \cup \cdots \cup \mathscr{s}_n$) and $\bigcap \mathbb{s}$ is the intersection (i.e. $\bigcap \mathbb{s} = \mathscr{s}_1 \cap \cdots \cap \mathscr{s}_n$) of all $\mathscr{s} \in \mathbb{s}$. Each term in equation (A.7) therefore identifies the same set of VAR_i.

$$\{VAR_i\}_{i \in \bigcap \mathbb{s}} = \{VAR_i : i \in \bigcap \mathbb{s}\} = \{VAR_i : i \in (\mathscr{s}_1 \cap \cdots \cap \mathscr{s}_n)\}\,, \tag{A.7}$$

Given that: $\mathbb{s} = \{\mathscr{s}_1 \subseteq J, \dots, \mathscr{s}_n \subseteq J\}$.

A.2.3. Summations

Typically, brackets will be used to denote variable sets while unbracketed items will be single values. While $\{VAR_i : i \in \mathscr{s}\}$ refers to a collection of variables, $VAR_{\mathscr{s}}$ will denote the total (i.e. the sum) of all $VAR_i \in \{VAR_i : i \in \mathscr{s}\}$. This summation notation will be used extensively for variables aggregated at the timepoint segment level.

Definition A-5 — $VAR_{\mathscr{s}}$ [u] is the *Sum* of all observed values of $VAR_i \in \{VAR_i\}_{i \in \mathscr{s}}$ with [u] units, which are the same as VAR_i.

(A.8) $$VAR_{\mathscr{s}} = \sum_{i \in \mathscr{s}}[VAR_i] = \sum[\{VAR_i : i \in \mathscr{s}\}]$$

A.3. Variable Modifiers

In addition to indexes, the notation of many variables will use superscripts and subscripts to the left of base variables (Definition A-6). These modifiers are used to indicate categories of variables and to distinguish variables that rely on the same base name.

Definition A-6 — Left-superscripts and left-subscripts:

- X, as a left-superscript, $({}^{X}VAR)$, denotes a variable of category X.
  - Xa, Xb, and Xc, as left-superscripts, $({}^{Xa}VAR, {}^{Xb}VAR, {}^{Xc}VAR)$ denote variables of categories Xa, Xb, and Xc, which are sub-categories of X.
- Y, as a left-subscript, $({}^{X}_{Y}VAR)$, may indicate:
  - A normalizing factor, Y, for variable of category X.
  - A variable of category X limited to scope Y.
  - A new variable relating to, but not necessarily derived from ${}^{X}VAR$.

Lastly, a right-superscript (e.g. VAR^{a}) will indicate an exponential relationship of VAR raised to the a power.

A.4. Example Variable

The generic variable, VAR, is used in to show the mains structures of variables, yet VAR is not analyzed itself not is it actually part of any dataset. As an example, using a variable from the data set, Table A-2 shows several ways that individual values or sets of values for door open duration (i.e. ${}^{T}DWL$) may be denoted.

Table A-2 — Variable notation example using Door Open Duration.

Notation	**Definition**
${}^{T}DWL$	Any single value of *Door Open Duration*. Does not refer to any specific value in the dataset, but has the same properties.
${}^{T}DWL_i$	Any one value of *Door Open Duration* found the data set. Entry can be a zero, a positive number, or missing.
${}^{T}\widehat{D}WL_i$	Any one corrected value of *Door Open Duration* found the data set. Entry can be a zero or a positive number
$\{{}^{T}\widehat{D}WL_i : i \in I\} = \{{}^{T}\widehat{D}WL_i\}_{i \in I}$	The complete set of all corrected values of *Door Open Duration* found the data set. Includes all values.
$\{{}^{T}\widehat{D}WL_i : i \in J\} = \{{}^{T}\widehat{D}WL_i\}_{i \in J}$	The complete set of all corrected, numeric, and non-zero values of Door Open Duration found the data set.
$\{{}^{T}\widehat{D}WL_i : {}^{T}\widehat{D}WL_i < 180\}_{i \in J}$	A subset of correction, numeric, and non-zero values of *Door Open Duration* found the data set, such that all values are less than 180.
${}^{\Sigma T}\widehat{D}WL$	Any single value of an *Aggregated Door Open Duration* for a timepoint segment. Does not refer to any specific value in the dataset, but has the same properties. The hat (i.e. (^) implies that only values within defined limits are possible.
${}^{\Sigma T}\widehat{D}WL_t$	Any one specific value of *Aggregated Door Open Duration*, calculated from corrected data, for a timepoint segment found the data set.
$\{{}^{\Sigma T}\widehat{D}WL_t\}_{t \in \mathbb{t}}$	The complete set of all values of *Aggregated Door Open Duration*, calculated from corrected data, for all timepoint segments found the data set.

APPENDIX B — DEFINITION TABLES

B.1. Introduction

This appendix uses some of the notation introduced in Appendix A. For most tables, subscripts are not included for each variable. As a reminder, variables that lack a right-subscript have the same properties as observational (i.e. recorded) data of the same name; but, do not refer to a specific record from, nor are they contained in the dataset.

B.1.1. Summary Tables

Units, Abbreviations, and Initialisms

Table B-1 — Unit definitions.

Category	**Units**	**Definition**
Number Sets	$\mathbb{B}$	Binary. Contains $\{0,1\}$
	$\mathbb{N}_1$	Positive integer set containing $\{1, 2, 3, \dots\}$
	$\mathbb{N}_0$	Non-negative integer set containing $\{0, 1, 2, 3, \dots\}$
	$\mathbb{Z}$	Integer set containing $\{\dots, -2, -1, 0, 1, 2, \dots\}$
	$\mathbb{R}$	Real number set containing all values, x, that can be represented by an infinite decimal expansion.
Distance	miles	Distance in Miles
	feet	Distance in Feet
	meters	Distance in Meters
Time	sec	Seconds
	$\mathcal{M}$sec	Seconds after midnight
Passengers	pax	Passengers
	pax^2	Passengers-squared

Table B-2 — Abbreviations and initialisms for non-variables.

Name	Definition
APC	Automatic Passenger Counts
AVL	Automatic Vehicle Location
BCD	BreadCrumb Data
Bus-bay	Bus catchment area
UTC	Coordinated Universal Time
ELD	Event Level Data
file.fst	File using “fst” data structure
file.csv	File using Comma Separated Values
GB	Gigabytes
GTFS	General Transit Feed Specification
GPS	Global Positioning System
HRD	High Resolution Data
ITS	Intelligent Transportation Systems
PDT	Pacific Daylight Time (UTC-07:00)
PLT	Pacific Local Time $:= \begin{cases} 02{:}00 \text{ PST} \rightarrow 03{:}00 \text{ PDT on } 2^{\text{nd}} \text{ Sun. in Mar.} \\ 02{:}00 \text{ PDT} \rightarrow 01{:}00 \text{ PST on } 1^{\text{st}} \text{ Sun. in Nov.} \end{cases}$
PST	Pacific Standard Time (UTC-08:00)
RAM	Random Access Memory
SDD	Stop Disturbance Data
SED	Stop Event Data
SCATS	Sydney Coordinated Adaptive Traffic System
TPS	Timepoint-Segment
TSP	Transit Signal Priority
TriMet	Tri-County Metropolitan District of Oregon
VIF	Variance Inflation factor

Indexing Variables and Index Sets

Some variables are needed to create indexes and index sets. Unique values of these variables, unique combinations, and other mathematical definitions are used to create the script indexes used for set-builder notation.

Table B-3 — Variables used to define index sets.

Variable	Definition	References & Page #	
$DATE$	Actual *Calendar Date* as defined by Pacific Local Time (PLT).	Definition 3-2	33
${}_{SVC}DATE$	*Service Date* defined by the TriMet service schedule for specific routs and lines.	Definition 3-3	33
DAY	*Day-of-the-Week* for which an observation was recorded.	Definition 3-23	44
DIR	*Direction of Travel* for TriMet routes. 1 is typically inbound to the Portland city center.	Definition 3-21	43
HR	*Service Hour* defined as the rounded down hour of PLT and is recorded as an integer value between 0 and 23	Definition 3-17	41
LOC	*Location Identification Number* for TriMet's bus stops.	Definition 3-22	44
RTE	*Route Identification Number* for TriMet's network. It is unique to each transit route, but not to the direction of travel.	Definition 3-20	43
$TRIP$	*Trip Identification Number* that is unique to one vehicle, for one day, for one complete route and direction.	Definition 3-24	51
VEH	*Vehicle Identification Number* that is unique to each bus or train in TriMet's Network	Definition 3-12	38

Table B-4 — Index set and subset definitions.

Subset	Definition	References & Page #	
$\mathcal{a}$	$\forall \mathcal{a}, \big((\mathcal{a} \subset I) \land (\mathcal{a} \in \mathbb{a})\big)$. Includes all VAR_i recorded on each unique $TRIP_i$. Family of $\mathcal{a}$ is contained in $\mathbb{a} = \{\mathcal{a}_1, \dots, \mathcal{a}_n\}$.	Definition 3-24	51
a	Represents all ordered events from a single trip, such that $a \in \mathcal{a} = \{a_1, a_2, \dots, a_n\}$ and $(i \mapsfrom a)$ is a function mapping index i from index a.		
$\dot{\mathcal{a}}$	$\forall \dot{\mathcal{a}}, \big((\dot{\mathcal{a}} \subseteq \mathcal{a}) \land (\dot{\mathcal{a}} \in \dot{\mathbb{a}})\big)$. Subset of $\mathcal{a}$ that includes all VAR_i that recorded at a scheduled bus-stop locations. The family of $\dot{\mathcal{a}}$ is contained in $\dot{\mathbb{a}}$.	Definition 3-24	51
$\dot{a}$	Represents all ordered event at scheduled locations, such that $\dot{a} \in \dot{\mathcal{a}} = \{\dot{a}_1, \dot{a}_2, \dots, \dot{a}_n\} \subseteq \mathcal{a}$ and $(i \mapsfrom \dot{a})$ and $(a \mapsfrom \dot{a})$ are functions mapping indexes i and a, respectively from index $\dot{a}$.		
$\mathcal{b}$	$\forall \mathcal{b}, \big((\mathcal{b} = \bigcap(\ell, \mathcal{d})), (\mathcal{b} \in \mathbb{b})\big)$. Unique real intersection of location ℓ and service day $\mathcal{d}$. Family of $\mathcal{b}$ is contained in $\mathbb{b} = \{\mathcal{b}_1, \dots, \mathcal{b}_n\}$.	(4.3.3)	73
b	Represents all ordered events in $\mathcal{b}$, such that $b \in \mathcal{b} = \{b_1, b_2, \dots, b_n\}$ and $(i \mapsfrom b)$ is a function that maps index i from index b.		72
$\dot{\mathcal{b}}$	$\forall \dot{\mathcal{b}}, \Big(\big(\dot{\mathcal{b}} = \bigcap(\ell, \mathcal{d}, \mathcal{r}_d)\big), \big(\dot{\mathcal{b}} \in \dot{\mathbb{b}}\big)\Big)$. Unique real intersection of location ℓ, service day $\mathcal{d}$, and route-direction $\mathcal{r}_d$. Family of $\dot{\mathcal{b}}$ is contained in $\dot{\mathbb{b}} = \{\dot{\mathcal{b}}_1, \dots, \dot{\mathcal{b}}_n\}$.	(4.3.6)	76
$\dot{b}$	Represents all ordered events in $\dot{\mathcal{b}}$, such that $\dot{b} \in \dot{\mathcal{b}} = \{\dot{b}_1, \dot{b}_2, \dots, \dot{b}_n\}$ and $\big(i \mapsfrom \dot{b}\big)$ is a function that maps index i from index $\dot{b}$.		76
$\mathcal{d}_0$	$\forall \mathcal{d}_0, \big((\mathcal{d}_0 \subset I) \land (\mathcal{d}_0 \in \mathbb{d}_0)\big)$. Includes all VAR_i recorded on a unique $DATE_i$. Family of $\mathcal{d}_0$ is contained in $\mathbb{d}_0 = \{\mathcal{d}_{0_1}, \dots, \mathcal{d}_{0_n}\}$.	Definition 3-2	33
$\mathcal{d}$	$\forall \mathcal{d}, \big((\mathcal{d} \subset I) \land (\mathcal{d} \in \mathbb{d})\big)$. Includes all VAR_i recorded on a unique ${}_{SVC}DATE_i$. Family of $\mathcal{d}$ is contained in $\mathbb{d} = \{\mathcal{d}_1, \dots, \mathcal{d}_n\}$.	Definition 3-3 (3.2.1)	33

Table B-4 (Continued) — Index set and subset definitions.

Subset	Definition	References & Page #	
h	$\forall h, \big((h \subset I) \wedge (h \in \mathbb{h})\big)$. Includes all VAR_i recorded during a unique HR_i. Family of h is contained $\mathbb{h} = \{h_0, \dots, d_{23}\}$.	Definition 3-17	41
$\dot{h}$	Modified index for hours that combines off-peak hours into a single index.	(4.3.9)	78
ℓ	$\forall \ell, \big((\ell \subset I) \wedge (\ell \in \mathbb{l})\big)$. Includes all VAR_i recorded at a unique LOC_i. Family of ℓ is contained in $\mathbb{l} = \{\ell_1, \dots, \ell_n\}$.	Definition 3-22	44
p	$\forall p, \Big(\big(p = \cap(r_w, \dot{h})\big), (p \in \mathbb{p})\Big)$. Unique real intersection of route-direction-day r_w and modified hours $\dot{h}$. Family of p is contained $\mathbb{p} = \{p_1, \dots, p_n\}$	(4.3.10)	79
r	$\forall r, \big((r \subset I) \wedge (r \in \mathbb{r})\big)$. Includes all VAR_i recorded for a unique RTE_i. Family of r is contained in $\mathbb{r} = \{r_1, \dots, r_n\}$.	Definition 3-20	43
r_d	$\forall r_d, \big((r_d \subseteq r) \wedge (r_d \in \mathbb{r}_d)\big)$. Partition of r. Includes all VAR_i recorded for a unique RTE_i and DIR_i. Family of r_d is contained in $\mathbb{r}_d = \{\, r_{d_1}, \dots, r_{d_n}\}$.	Definition 3-21	43
r_w	$\forall r_w, \big((r_w = \cap(r_d, w)), (r_w \in \mathbb{r}_w)\big)$. Unique real intersection of a route-direction r_d and weekdays or weekends w. Family of r_w is contained in $\mathbb{r}_w = \{r_{w_1}, \dots, r_{w_n}\}$.	(4.3.8)	78
$\Psi_m(s)$	A random sample of size m taken from a $s \subseteq I$, where m is user defined and strictly less than $\lVert s \rVert$, the number of elements in s.	Definition 5-1	105
$\psi(m)$	A function to define a sample size for $\Psi_m(s)$. $\psi_1(m)$, $\psi_2(m)$, … are functions defined within this dissertation.	Definition 5-2 (5.2.1) & (5.2.2)	105 111
t	$\forall t, (t \subset I)$. Includes all VAR_i recorded for one timepoint-segment. A Timepoint segment has one timepoint, one route-direction, and spans one hour. Family of t is contained in $\mathbb{t}$.		84
$\dot{t}$	$\forall \dot{t}, (\dot{t} \subseteq t)$. Subset of t. Includes all VAR_i recorded at each unique stop.		84
$\dot{t}$	Represents the set of ordered events for one $\dot{t}$, such that $\dot{t} \in \dot{t} = \{\dot{t}_1, \dots, \dot{t}_n\}$. Headways calculations focus the first and last stops only (i.e. $\dot{t}_1$ and $\dot{t}_n$).	(4.4.1)	84
t_p	$\forall t_p, \Big(\big(t_p = \cap(t, \dot{h})\big), (t_p \in \mathbb{t}_p)\Big)$. Unique real intersection of a timepoint segment t and modified hour index $\dot{h}$. Family of t_p is contained in $\mathbb{t}_p = \{t_{p_1}, \dots, t_{p_n}\}$.	(4.4.9)	96
v	$\forall v, \big((v \subseteq I) \wedge (v \in \mathbb{v})\big)$. Includes all VAR_i recorded on each unique VEH_i. Family of v is contained in $\mathbb{v} = \{v_1, \dots, v_n\}$.	Definition 3-12	38
v_d	$\forall v_d, \big((v_d = \cap(v, d)), (v_d \in \mathbb{v}_d)\big)$. Unique real intersection of vehicle index v and date index d. Family of v_d is contained in $\mathbb{v}_d = \{v_{d_1}, \dots, v_{d_n}\}$.	(3.3.3)	39
w	$\forall w, \big((w \subseteq I) \wedge (w \in \mathbb{w})\big)$. Includes all VAR_i recorded on weekdays (i.e. Mon. – Fri.) or on weekends (i.e. Sat. – Sun.). Family of w contained in $\mathbb{w} = \{w_0, w_1\}$.	Definition 3-23	44
z	$\forall z, \big((z = \cap(\ell, r_d, h, w)), (z \in \mathbb{z})\big)$. Unique real intersection of location ℓ, route-direction r_d, hour h, and weekday/weekend w. Family of z is contained in $\mathbb{z} = \{z_1, \dots, z_n\}$.	(3.3.8)	44

Table B-5 — Flagged index sets and subsets.

Subset	Definition	Reference & Page #	
$\breve{\mathscr{f}}$	Union of all $\breve{\mathbb{V}}_d$.		40
$\breve{\mathscr{f}}_{\{1,2,3\}}$	Union of $\breve{\mathscr{f}}$ and $\breve{\mathscr{g}}_{\{1,2,3\}}$. $\{1\}$ indicates ONS_i. $\{2\}$ indicates $OFFS_i$, and $\{3\}$ indicates ${}^{T}DWL_i$.		49
$\breve{\mathscr{f}}_4$	Equal to $\breve{\mathscr{g}}_4$. Flagged set of global and local outliers for ${}^{T}BAY_i$		
$\breve{\mathscr{g}}_{\{1,2,3,4\}}$	Flagged sets of global and local outliers defined as the union of their respective $\breve{\mathscr{g}}'_{\{1,2,3,4\}}$ and $\breve{\mathscr{g}}''_{\{1,2,3,4\}}$		49
$\breve{\mathscr{g}}'_1$	Flagged global outlier for ONS_i	(3.3.7)	43
$\breve{\mathscr{g}}'_2$	Flagged global outlier for $OFFS_i$		
$\breve{\mathscr{g}}'_3$	Flagged global outlier for ${}^{T}DWL_i$		
$\breve{\mathscr{g}}'_4$	Flagged global outlier for ${}^{T}BAY_i$		
$\breve{\mathscr{g}}''_1$	Flagged local outlier for ONS_i	(3.3.17)	49
$\breve{\mathscr{g}}''_2$	Flagged local outlier for $OFFS_i$		
$\breve{\mathscr{g}}''_3$	Flagged local outlier for ${}^{T}DWL_i$		
$\breve{\mathscr{g}}''_4$	Flagged local outlier for ${}^{T}BAY_i$		
$\breve{\mathbb{V}}_d$	Flagged vehicle-day combination. Used to identify broken passenger counters.	(3.3.4)	40

Events, Times, and Durations

Table B-6 — Service and non-service event variables

Event	Units	Definition	Reference & Page #	
${}^{E}DSTB$	$\mathbb{B}$	Disturbance events. (Stopping) event where a vehicle stops outside of a bus-bay that is part of secluded service.	Definition 3-6	35
${}^{E}SVC$		Service events. (Stopping) event where a vehicle stops within a bus-bay that is part of secluded service.	Definition 3-5	35
${}^{E}THRU$		Thru events. (Non-stopping) event where a vehicle does not stop within a bus-bay that is part of secluded service.	Definition 3-7	35
${}^{t}SKD$	$\mathcal{M}$sec	Officially scheduled departure time at a bus stop in TriMet's network.	Definition 3-4	34
${}^{t}ARR$		Vehicle arrival time defined as observed time that: Service $({}^{E}SVC)$ ~ a vehicle enters a bus-bay. Disturbance $({}^{E}DTSB)$ ~ a vehicle stops moving for more than five-seconds outside a bus-bay. Thru $({}^{E}THRU)$ ~ a vehicle passes a bus stop.	Definition 3-8	35
${}^{t}DEP$		Vehicle departure time defined as observed time that: Service $({}^{E}SVC)$ ~ a vehicle exits a bus-bay. Disturbance $({}^{E}DTSB)$ ~ a vehicle stops moving for more than five-seconds outside a bus-bay. Thru $({}^{E}THRU)$ ~ a vehicle passes a bus stop.	Definition 3-9	36

Table B-7 — Service duration variables.

Duration	Unit	Definition	Reference & Page #	
${}^{T}BAY$	[sec]	*Bus-Bay (Stop) Duration* and is recorded in integer seconds defined by the difference between arrival time and departure time.	Definition 3-11 (3.2.2)	36
${}^{T}\tilde{B}AY$		${}^{T}BAY$ with added jitter to make continuous for values greater than or equal to 1.	(3.3.14)	47
${}^{T}\hat{B}AY$		Corrected ${}^{T}\tilde{B}AY$ based on flags and probability distributions.	(3.3.18)	50
${}^{\Sigma T}\hat{B}AY$		Summation of all ${}^{T}\hat{B}AY$ for all vehicles within one timepoint segment.		85
${}^{T}DSTB$	[sec]	*Disturbance Duration* of unscheduled stops between consecutive bus stop locations.	(4.2.2)	57
${}^{T}\tilde{D}STB$		${}^{T}DSTB$with added jitter to make continuous for values greater than or equal to 1.		58
${}^{\Sigma T}\tilde{D}STB$		Summation of all ${}^{T}\tilde{D}STB$ for all vehicles within one timepoint segment.		148
${}^{T}DWL$	[sec]	*Door Open Duration* at bus stops and is recorded in integer seconds defined by the total time vehicle doors are open at a bus stop.	Definition 3-10	36
${}^{T}\tilde{D}WL$		${}^{T}DWL$ with added jitter to make continuous for values greater than or equal to 1.	(3.3.15)	47
${}^{T}\hat{D}WL$		Corrected ${}^{T}\tilde{D}WL$ based on flags and probability distributions.	(3.3.19)	50
${}^{\Sigma T}\hat{D}WL$		Summation of all ${}^{T}\hat{D}WL$ for all vehicles within one timepoint segment.		85
${}^{T}MOVE$	[sec]	*Moving Duration* between consecutive bus-stop locations, excluding the disturbance duration.	(4.2.3)	58
${}^{T}\tilde{M}OVE$		${}^{T}MOVE$ with added jitter to make continuous for values greater than or equal to 1.		58
${}^{\Sigma T}\tilde{M}OVE$		Summation of all ${}^{T}\tilde{M}OVE$ for all vehicles within one timepoint segment.		148
${}^{T}\tilde{T}RVL$	[sec]	*Total Travel Duration* from when a vehicle begins servicing passengers as its first stop and stops serving passengers at its last stop.	Definition 4-3 (4.2.4)	58
${}^{\Sigma T}\tilde{T}RVL$		Summation of all ${}^{T}\tilde{T}RVL$ for all vehicles within one timepoint segment.		159
${}^{\Sigma T}_{\mu(skd)}VAR$	[sec]	Average duration of service (per scheduled stop location) for a timepoint segment.		88
${}^{\Sigma T}_{\mu(svc)}VAR$		Average duration of service (per serviced stop location) for a timepoint segment.		
${}^{\Sigma T}_{\mu(veh)}VAR$		Average duration of service (per vehicle) for a timepoint segment.		

Passenger Movements

Table B-8 — Passenger movement variables.

Variable	Units	Definition	References & Page #	
$\hat{L}IFT$	$\mathbb{B}$	Cleaned and binary version of *Wheelchair Ramp Deployment.*	Definition 4-14	71
${}^{\Sigma}\hat{L}IFT$	$\mathbb{N}_0$	Summation of all $\hat{L}IFT$ for all vehicles within one timepoint segment.		85
$LOAD$	pax	*Estimated Passenger Load* onboard a vehicle at a given location.	Definition 3-15	38
$\hat{L}OAD$		Corrected $LOAD$ based on flags and probability distributions.	(3.3.21)	52
$OFFS$	pax	*Number of passengers Alighting* (i.e. exiting) a vehicle at a bus stop.	Definition 3-14	38
$\hat{O}FFS$		Corrected $OFFS$ based on flags and probability distributions.	(3.3.22)	52
${}^{\Sigma}\hat{O}FFS$		Summation of all $\hat{O}FFS$ for all vehicles within one timepoint segment.		85
$\hat{O}FFS^2$	pax²	Square term of $\hat{O}FFS$.	Definition 4-13	67
${}^{\Sigma}(\hat{O}FFS^2)$		Sum of $\hat{O}FFS^2$ at the TPS level	Figure C-5	266
$({}^{\Sigma}\hat{O}FFS)^2$		Square term of $\hat{O}FFS^2$.		85
ONS	Pax	*Number of passengers Boarding* (i.e. entering) a vehicle at a bus stop.	Definition 3-13	38
$\hat{O}NS$		Corrected ONS based on flags and probability distributions.	(3.3.20)	51
${}^{\Sigma}\hat{O}NS$		Summation of all $\hat{O}NS$ for all vehicles within one timepoint segment.		85
$\hat{O}NS^2$	pax²	Square term of $\hat{O}NS$.	Definition 4-13	67
${}^{\Sigma}(\hat{O}NS^2)$		Sum of $\hat{O}NS^2$ at the TPS level	Figure 5-21	138
$({}^{\Sigma}\hat{O}NS)^2$		Square term of $\hat{O}NS^2$.		85
${}^{\Sigma}_{\mu(skd)}VAR$	pax	Average number of passenger movements (per scheduled stop location) for a timepoint segment.		88
${}^{\Sigma}_{\mu(svc)}VAR$		Average number of passenger movements (per serviced stop location) for a timepoint segment.		
${}^{\Sigma}_{\mu(veh)}VAR$		Average number of passenger movements (per vehicle) for a timepoint segment.		

Locations

Table B-9 — Bus stop location variables.

Location	Unit	Definition	References & Page #	
^{L}VAR	$\mathbb{B}$	*Location-type* variable with binary units.	Definition 4-4	*59*
^{L}AT	$\mathbb{B}$	*At* bus stop location, as shown in Figure 4-5.	Definition 4-6	59
^{L}FAR		*Farside* bus stop location, as shown in Figure 4-3.	Definition 4-6	59
$^{L}MALL$		Stops located on the *downtown transit mall*.	Definition 4-12	65
$^{L}NEAR$		*Nearside* bus stop location, as shown in Figure 4-2.	Definition 4-6	59
^{L}OPP		*Opposite* bus stop location, as shown in Figure 4-4.	Definition 4-6	59
$^{L}P\&R$		Stop located within a quarter-mile of an official *park-and-ride* facility	Definition 4-11	64
^{L}SIG		Stops located near *signalized intersections*	Definition 4-7	61
^{L}TC		Stops located at a *Transit Center*	Definition 4-10	64
^{L}TP		*Timepoint* stops. Used to define and maintain serviced schedules.	Definition 4-5	59
^{Ls}VAR	$\mathbb{B}$	*Signalized Location-type* variable with binary units. Can be applied to all ^{L}VAR, listed above.	Definition 4-8 (4.3.1)	*62*
^{Lu}VAR	$\mathbb{B}$	*Unsignalized Location-type* variable with binary units. Can be applied to all ^{L}VAR, listed above.	Definition 4-9 (4.3.2)	62
$^{\Sigma L}VAR$	$\mathbb{N}_0$	*Serviced location-type* variable. Summation of all ^{L}VAR for all vehicles within one timepoint segment.		91
$^{\Sigma L}_{skd}VAR$		*Scheduled location-type* variable. Total number of stops on the service schedule for all vehicles within one timepoint segment.		
$^{\Sigma L}_{thru}VAR$		*Thru location-type* variable. Total number of stops on the service schedule that were not served by any vehicles in the timepoint-segment.		
$^{\Sigma Ls}VAR$	$\mathbb{N}_0$	*Signalized Serviced location-type* variable. Summation of all ^{Ls}VAR for all vehicles within one timepoint segment.		91
$^{\Sigma Ls}_{skd}VAR$		*Signalized Scheduled location-type* variable. See definition for$^{\Sigma L}_{skd}VAR$.		
$^{\Sigma Ls}_{thru}VAR$		*Signalized Thru (Non-serviced) location-type* variable. See definition for$^{\Sigma L}_{thru}VAR$.		
$^{\Sigma Lu}VAR$	$\mathbb{N}_0$	*Unsignalized Serviced location-type* variable. Summation of all ^{Lu}VAR for all vehicles within one timepoint segment.		91
$^{\Sigma Lu}_{skd}VAR$		*Unsignalized Scheduled location-type* variable. See definition for$^{\Sigma L}_{skd}VAR$.		
$^{\Sigma Lu}_{thru}VAR$		*Unsignalized Thru (Non-serviced) location-type* variable. See definition for$^{\Sigma L}_{thru}VAR$.		

Vehicle Interactions

Table B-10 — Bus interaction variables.

Variable	Unit	Definition	References & Page #	
^{I}VAR	$\mathbb{N}_0$	Interaction-type variable.		72
$^{I}JUMP$	$\mathbb{N}_0$	*Jumping Interaction* for $VEH_{i\leftrightarrow b}$ and gives the number of Bus B interactions from Scenario (3) in Figure 4-13 and Table 4-6.	Definition 4-18 (4.3.4)	74 74
$^{I}LEAD$		*Leading Interaction* for $VEH_{i\leftrightarrow b}$ and gives the number of Bus A interactions from Scenario (2). Figure 4-13 and Table 4-6.	Definition 4-15 (4.3.4)	73 74
$^{I}TAIL$		*Tailing Interaction* for $VEH_{i\leftrightarrow b}$ and gives the number of Bus B interactions from Scenario (2). Figure 4-13 and Table 4-6.	Definition 4-16 (4.3.4)	73 74
$^{I}WAIT$		*Waiting Interaction* for $VEH_{i\leftrightarrow b}$ and gives the number of Bus A interactions from Scenario (3). Figure 4-13 and Table 4-6.	Definition 4-17 (4.3.4)	74 74
^{I}INT	$\mathbb{N}_0$	Sum of all interactions for $VEH_{i\leftrightarrow b}$ for one bus-bay stopping event.	Definition 4-19	75
^{Is}VAR	$\mathbb{N}_0$	Interaction-type variables from vehicles of the *Same* route.		76
^{Id}VAR		Interaction-type variables from vehicles of the *Different* routes.		
^{Is}INT	$\mathbb{N}_0$	Sum of all interactions for $VEH_{i\leftrightarrow b}$ for one bus-bay stopping event between vehicles of the *same* route.	Definition 4-20	76
^{Id}INT		sum of all Interactions for $VEH_{i\leftrightarrow b}$ for one bus-bay stopping event between vehicles of *different* routes.	Definition 4-21	76

Headways

Table B-11 — Headway and headway performance metrics variables.

Variable	Unit	Definition	References & Page #	
${}^{A}H$	sec	*Arrivals, Departures*, and *Scheduled Service* between two consecutive vehicles of the same route servicing a given stop.	Definition 4-22 (4.3.7)	77
${}^{D}H$				
${}^{S}H$				
${}^{A}_{avg}H$	sec	*Mean* (avg) *Headway* between vehicles *arriving* (A) at the first stop or *departing* (D) the last stop of a timepoint-segment.	Definition 4-25 (4.4.2)	93
${}^{D}_{avg}H$				
${}^{A}_{mad}H$	sec	*Mean Absolute Deviation* (mad) for *arrivals* at the first stop and *departures* at the last stop of a TPS. ${}^{\{A,D\}}_{mad}H$ are defined as the absolute difference between headways and mean headway.	Definition 4-26 (4.4.3)	94
${}^{D}_{mad}H$				
${}^{A}_{idx}H$	ratio	*Headway Deviation Indexes* (idx) for *arrivals* at the first stop and *departures* at the last stop of a TPS.	Definition 4-27 (4.4.4)	94
${}^{D}_{idx}H$				
${}^{SA}_{avg}H$	sec	*Mean Headway* for *scheduled arrivals* at the first stop and *scheduled departures* at the last stop of a TPS.	Definition 4-28 (4.4.5)	95
${}^{SD}_{avg}H$				
${}^{SA}_{mad}H$	sec	*Mean Absolute Deviation* (mad) for *scheduled arrivals* at the first stop and scheduled departures at the last stop.	Definition 4-29 (4.4.6)	95
${}^{SD}_{mad}H$				
${}^{SA}_{idx}H$	ratio	*Headway Deviation Indexes* (idx) for *scheduled arrivals* at the first stop and *scheduled departures* at the last stop of a TPS.	Definition 4-30 (4.4.7)	95
${}^{SD}_{idx}H$				
${}^{A}_{adj}H$	ratio	*Adjusted Deviation Indexes* for *Arrivals* at the first stop and *Departures* at the last stop, respectively.	Definition 4-31 (4.4.8)	96
${}^{D}_{adj}H$				

Congestion

Table B-12 — Congestion duration and cost variables.

Variable	**Unit**	**Definition**	**References & Page #**	
${}^{C}VAR_{\mathscr{p}}$		*Congestion-type* variable. Increase in period $\mathscr{p}$ over baseline in period $\mathscr{p} = \mathscr{p}_0$.		
${}^{CT}VAR_{\mathscr{p}}$	sec	*Congestion Duration-type* variable. Increase elapsed time in period $\mathscr{p}$ over baseline in period $\mathscr{p} = \mathscr{p}_0$.		*79*
${}^{C\$}VAR_{\mathscr{p}}$	\$, as USD	*Congestion Cost-type* variable. Increase in accrued costs in period $\mathscr{p}$ over baseline in period $\mathscr{p} = \mathscr{p}_0$.		
${}^{CT}_{avg}\widetilde{E}XW_{\mathscr{p}}$	sec	*Excess wait Times.* The average increase in *Passenger* time, per passenger, excess wait times in period $\mathscr{p}$ over baseline in $\mathscr{p} = \mathscr{p}_0$.	(4.3.20)	*83*
${}^{C\$}_{avg}EXW_{\mathscr{p}}$	USD	*Excess Wait Costs.* The average increase in *Passenger* costs, per passenger, attributed to ${}^{CT}_{avg}\widetilde{E}XW_{\mathscr{p}}$.	(4.3.21)	*83*
${}^{\Sigma CT}_{avg}\widetilde{E}XW_{\mathscr{t}_p}$	sec	*Aggregated Excess Wait Times.* The average increase in *Passenger* time, per passenger, from excess wait times in modified timepoint segment $\mathscr{t}_p$ over baseline in $\mathscr{t}_p = \mathscr{t}_{p_0}$.	(4.4.14)	*99*
${}^{\Sigma C\$}_{avg}\widetilde{E}XW_{\mathscr{t}_p}$	USD	*Aggregated Excess Wait Costs.* The average increase in *Passenger* costs, per passenger, attributed to ${}^{\Sigma CT}_{avg}\widetilde{E}XW_{\mathscr{t}_p}$.	(4.4.15)	*99*
${}^{CT}_{avg}\tilde{R}\&B_{\mathscr{p}}$	sec	*Ride and Buffer Time.* The average increase in *Passenger* time, per passenger, from in-vehicle time and buffer-time in period $\mathscr{p}$ over baseline in $\mathscr{p} = \mathscr{p}_0$.	(4.3.18)	*82*
${}^{C\$}R\&B_{\mathscr{p}}$	USD	*Ride and Buffer Costs*. The average increase in *Passenger* costs, per passenger, attributed to ${}^{CT}_{avg}\tilde{R}\&B_{\mathscr{p}}$.	(4.3.19)	*82*
${}^{\Sigma CT}_{avg}R\&B_{\mathscr{t}_p}$	sec	*Aggregated Ride and Buffer Time.* The average increase in *Passenger* time, per passenger, from in-vehicle time and buffer-time in modified timepoint segment $\mathscr{t}_p$ over baseline in $\mathscr{t}_p = \mathscr{t}_{p_0}$.	(4.4.12)	*98*
${}^{\Sigma C\$}_{avg}R\&B_{\mathscr{t}_p}$	USD	*Aggregated Ride and Buffer Costs*. The average increase in *Passenger* costs, per passenger, attributed to ${}^{\Sigma CT}_{avg}R\&B_{\mathscr{t}_p}$.	(4.4.13)	*98*
${}^{CT}_{avg}\tilde{R}\&R_{\mathscr{p}}$	sec	*Ride and Recovery Time.* The average increase in *Agency* travel time and associated recovery time, per trip, in period $\mathscr{p}$ over baseline in $\mathscr{p} = \mathscr{p}_0$	(4.3.14)	*81*
${}^{C\$}_{avg}R\&R_{\mathscr{p}}$	USD	*Rider and Recovery Costs.* The increase in *Agency* costs, per trip, attributed to ${}^{CT}_{avg}\tilde{R}\&R_{\mathscr{p}}$.	(4.3.15)	*81*

Table B-12 (Continued) — Congestion duration and cost variables.

Variable	Unit	Definition	References & Page #	
${}^{\Sigma CT}_{avg}\tilde{R}\&R_{\mathscr{t}_p}$	sec	*Aggregated Ride and Recovery Time.* The average increase in *Agency* travel time and associated recovery time, per trip, in modified timepoint segment $\mathscr{t}_p$ over baseline in $\mathscr{t}_p = \mathscr{t}_{p_0}$.	Definition 4-32 (4.4.10)	*97*
${}^{\Sigma C\$}_{avg}R\&R_{\mathscr{t}_p}$	USD	*Aggregated Rider and Recovery Costs.* The increase in *Agency* costs, per trip, attributed to ${}^{\Sigma C\$}_{avg}R\&R_{\mathscr{t}_p}$.	(4.4.11)	*97*
${}^{CT}_{avg}\tilde{R}CV_{\mathscr{p}}$	sec	*Recovery Time.* The average increase in *Agency* recovery time, per trip, in period $\mathscr{p}$ over baseline in $\mathscr{p} = \mathscr{p}_0$.	(4.3.16)	*82*
${}^{C\$}_{avg}RCV_{\mathscr{p}}$	USD	*Recovery Costs.* The increase in *Agency* costs, per trip, attributed to ${}^{CT}_{avg}\tilde{R}CV_{\mathscr{p}}$.	(4.3.17)	*82*
${}^{CT}_{avg}\tilde{R}UN_{\mathscr{p}}$	sec	*Ride Time.* The average increase in *Agency* travel time and associated recovery time, per trip, in period $\mathscr{p}$ over baseline in $\mathscr{p} = \mathscr{p}_0$.	Definition 4-23 (4.3.12)	*80* *80*
${}^{C\$}_{avg}RUN_{\mathscr{p}}$	USD	*Rider Costs.* The increase in *Agency* costs, per trip, attributed to ${}^{CT}_{avg}\tilde{R}UN_{\mathscr{p}}$	(4.3.13)	81

B.2. Regression Model Variables

The following two tables include the variables found in the reported regression models. Other variables were tested, but are not listed below.

Table B-13 — Variable definitions for stop event level linear regression models.

Category	Variable	Unit	Definition
Passenger Movements	$\hat{O}NS$	pax	(Corrected) Number of passenger boardings (entering) a vehicle at a bus stop
	$\hat{O}FFS$		(Corrected) Number of passenger alightings (exiting) a vehicle at a bus stop
	$\hat{O}NS^2$	pax²	Square of $\hat{O}NS$
	$\hat{O}FFS^2$		Square of $\hat{O}FFS$
	$\hat{L}IFT$	$\mathbb{B}$	(Corrected) Wheelchair Ramp Deployment. 1 if ramp deploys, 0 otherwise.
Bus Stop Locations (L)	${}^{L}TP$	$\mathbb{B}$	1 if timepoint stop location, 0 otherwise.
	${}^{L}TC$		1 if stop is located at a *Transit Center, 0 otherwise.*
	${}^{L}MALL$		1 if stop is located on the downtown transit mall (i.e. 5th or 6th Ave) , 0 otherwise.
	${}^{L}NEAR$		1 if stop is a “nearside” stop location, 0 otherwise. (Figure 4-2)
	${}^{L}FAR$		1 if stop is a “farside” stop location, 0 otherwise. (Figure 4-3)
	${}^{L}OPP$		1 if stop is an “opposite” stop location, 0 otherwise. (Figure 4-4)
	${}^{L}AT$		1 if stop is an “at” stop location, 0 otherwise. (Figure 4-5)
	${}^{L}P\&R$		1 if stop is located within ¼ mile from a designated park-and-ride facility, 0 otherwise.
Traffic Signal	${}^{L}SIG$	$\mathbb{B}$	1 if stop located near a signalized intersection, 0 otherwise
Weekday	$WDAY$	$\mathbb{B}$	1 if weekday, 0 otherwise.
High-Freq RTE	$FREQ$	$\mathbb{B}$	1 if high-frequency route, 0 otherwise.
Weekdays	W_1^{AM}	$\mathbb{B}$	1 if TPS corresponds to 07:00-08:59 on weekdays, 0 otherwise.
	W_1^{PM}		1 if TPS corresponds to 15:00-17:59 on weekdays, 0 otherwise.
Weekends	W_0^{P}	$\mathbb{B}$	1 if TPS corresponds to 12:00-18:59 on weekends, 0 otherwise.
Vehicle Interactions at Bus Stops	${}^{I}INT$	$\mathbb{N}_0$	Number of interactions with other vehicles a stop.
	${}^{I}LEAD$		Number of interactions with other vehicles at a bus stop as the *“leading”* vehicle (See Figure 4-13 and Table 4-6)
	${}^{I}TAIL$		Number of interactions with other vehicles at a bus stop as the “*tailing* “vehicle (See Figure 4-13 and Table 4-6)
	${}^{I}WAIT$		Number of interactions with other vehicles at a bus stop as the “*waiting”* vehicle (See Figure 4-13 and Table 4-6)
	${}^{I}JUMP$		Number of interactions with other vehicles at a bus stop as the “*jumping”* vehicle (See Figure 4-13 and Table 4-6)
Same Route (Is)	${}^{Is}INT$	$\mathbb{N}_0$	Number of interactions with other vehicles at bus stops for vehicles within the same route.

Table B-14 — Variable definitions for aggregated TPS linear regression models.

Category	Variable	Unit	Definition
Number of Vehicles	${}^{\Sigma}VEH$	$\mathbb{N}_0$	Number of vehicles within a TPS
Total Distance	${}^{\Sigma}MILES$	Miles	Total distance traveled by all vehicles within a TPS
Total Passenger Movements (Σ)	${}^{\Sigma}\hat{O}NS$	pax	Total number of boarding (entering) passengers for all vehicles within a TPS
	${}^{\Sigma}\hat{O}FFS$		Total number of alighting (exiting) passengers for all vehicles within a TPS
	${}^{\Sigma}\hat{O}NS^2$	pax^2	Sum of ONS^2 for all vehicles within a TPS
	${}^{\Sigma}\hat{O}FFS^2$		Sum of ONS^2 for all vehicles within a TPS
	${}^{\Sigma}\hat{L}IFT$	$\mathbb{N}_0$	Total number of wheelchair lift activations by all vehicles within a TPS
Total Serviced Bus Stops Locations (ΣL)	${}^{\Sigma L}TC$	$\mathbb{N}_0$	Number of serviced *transit center* bus stops locations, by all vehicles, within a TPS
	${}^{\Sigma L}MALL$		Number of serviced bus stop locations on the downtown transit mall, by all vehicles, within a TPS
	${}^{\Sigma L}NEAR$		Number of serviced *nearside* bus stops, by all vehicles, within a TPS (Figure 4-2).
	${}^{\Sigma L}FAR$		Number of serviced *farside* bus stops, by all vehicles, within a TPS (Figure 4-3).
	${}^{\Sigma L}OPP$		Number of serviced *opposite* bus stops, by all vehicles, within a TPS (Figure 4-4).
	${}^{\Sigma L}AT$		Number of serviced *at* bus stops, by all vehicles, within a TPS (See Figure 4-5).
Total Non-Serviced Bus Stop Locations $\binom{\Sigma L}{thru}$	${}^{\Sigma L}_{thru}TC$	$\mathbb{N}_0$	Number of non-serviced *transit center* bus stops location, by all vehicles, within a TPS.
	${}^{\Sigma L}_{thru}NEAR$		Number of non-serviced *nearside* bus stops, by all vehicles, within a TPS (Figure 4-2).
	${}^{\Sigma L}_{thru}FAR$		Number of non-serviced *farside* bus stops, by all vehicles, within a TPS (Figure 4-3)
	${}^{\Sigma L}_{thru}OPP$		Number of non-serviced *opposite* bus stops, by all vehicles, within a TPS (Figure 4-4)
	${}^{\Sigma L}_{thru}AT$		Number of non-serviced *at* bus stops, by all vehicles, within a TPS (Figure 4-5)

Table B-14 (Continued) — Variable definitions for aggregated TPS linear regression models.

Category	Variable	Unit	Definition
Total Scheduled Bus Stop Locations near Traffic Signals $\binom{\Sigma Ls}{skd}$	${}^{\Sigma Ls}_{skd}TC$	$\mathbb{N}_0$	Number of *transit center* bus stops located near signalized intersections on the service schedule for all vehicles within a timepoint segment.
	${}^{\Sigma Ls}_{skd}NEAR$		Number of *nearside* bus stops (Figure 4-2) located near signalized intersections on the service schedule for all vehicles within a timepoint segment.
	${}^{\Sigma Ls}_{skd}FAR$		Number of *farside* bus stops (Figure 4-3) located near signalized intersections on the service schedule for all vehicles within a timepoint segment.
	${}^{\Sigma Ls}_{skd}OPP$		Number of *opposite* bus stops (Figure 4-4) located near signalized intersections on the service schedule for all vehicles within a timepoint segment.
	${}^{\Sigma Ls}_{skd}AT$		Number of *at* bus stops (See Figure 4-5) located near signalized intersections on the service schedule for all vehicles within a timepoint segment.
Weekdays	$WDAY$	$\mathbb{B}$	1 if stop event occurred on a weekday, 0 otherwise
High-Frequency RTE	$FREQ$	$\mathbb{B}$	1 if RTE is a frequent service route, 0 otherwise
Weekdays	W_1^{AM}		1 if TPS corresponds to 07:00-08:59 on weekdays, 0 otherwise.
	W_1^{PM}		1 if TPS corresponds to 15:00-17:59 on weekdays, 0 otherwise.
Weekends	W_0^P		1 if TPS corresponds to 12:00-18:59 on weekends, 0 otherwise.
High-Frequency RTE	${}^{\Sigma}VEH \times FREQ$	$\mathbb{B} \times \mathbb{N}_0$	Number of vehicles in TPS times $FREQ$ (as binary).
Weekdays	${}^{\Sigma}VEH \times W_1^{AM}$		Number of vehicles in TPS times W_1^{AM} (as binary).
	${}^{\Sigma}VEH \times W_1^{PM}$		Number of vehicles in TPS times W_1^{PM} (as binary).
Weekends	${}^{\Sigma}VEH \times W_0^P$		Number of vehicles in TPS times W_0^P (as binary).
High-Frequency RTE	${}^{\Sigma}MILES \times FREQ$	$\mathbb{B} \times$ miles	Number of miles travel by all vehicles in TPS times $FREQ$ (as binary).
Weekdays	${}^{\Sigma}MILES \times W_1^{AM}$		Number of miles travel by all vehicles in TPS times W_1^{AM} (as binary).
	${}^{\Sigma}MILES \times W_1^{PM}$		Number of miles travel by all vehicles in TPS times W_1^{PM} (as binary).
Weekends	${}^{\Sigma}MILES \times W_0^P$		Number of miles travel by all vehicles in TPS times W_0^P (as binary).

Table B-14 (Continued) — Variable definitions for aggregated TPS linear regression models.

Category	Variable	Unit	Definition
Total Different $ROUTE$ Vehicle Interactions (ΣId)	$^{\Sigma Id}INT$	$\mathbb{N}_0$	Number of interactions at bus stops between vehicles of different RTE within a TPS
	$^{\Sigma Id}LEAD$		Number of different RTE vehicle interactions at bus stops as the *leading* vehicle within TPS
	$^{\Sigma Id}TAIL$		Number of different RTE vehicle interactions at bus stops as the *tailing* vehicle within TPS
	$^{\Sigma Id}WAIT$		Number of different RE vehicle interactions at bus stops as the *waiting* vehicle within TPS
	$^{\Sigma Id}JUMP$		Number of different RTE vehicle interactions at bus stops as the *jumping* vehicle within TPS
Total Same $ROUTE$ Interactions (ΣIs)	$^{\Sigma Is}INT$	$\mathbb{N}_0$	Number of interactions at bus stops for vehicles of within the same RTE within a TPS

APPENDIX C — SUPPLEMENTARY FIGURES AND TABLES

C.1. Passenger Movement Graphics

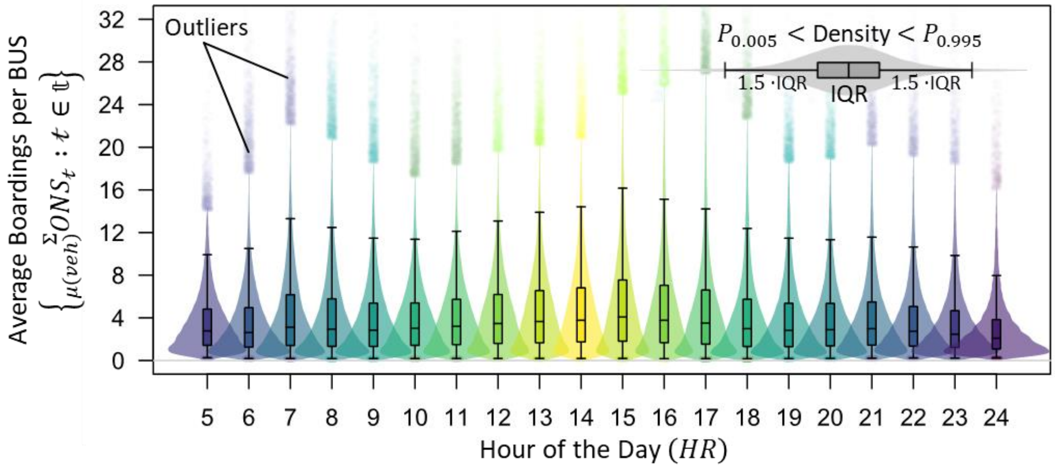

Figure C-1 — Violin and box-plots for all timepoint segments. Average boardings per vehicle, $\left\{{}_{\mu(veh)}^{\Sigma}\hat{O}NS_{t} : t \in \mathbb{t}\right\}$, given hour-of-the-day (HR_{t}).

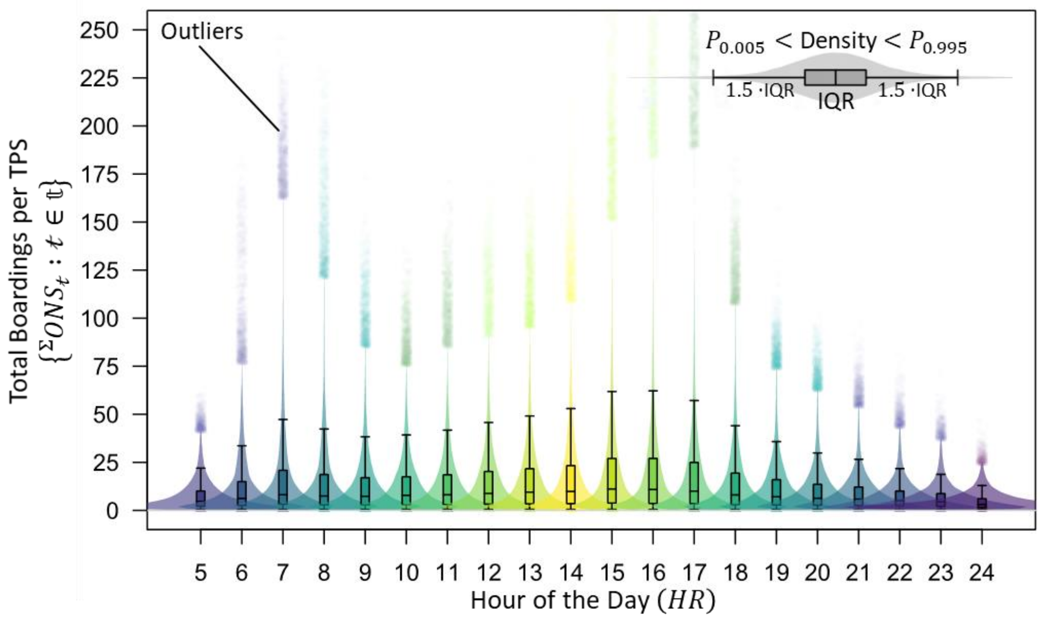

Figure C-2 — Violin and box-plots for all timepoint segments. Total boardings per TPS, $\left\{{}^{\Sigma}\hat{O}NS_{t} : t \in \mathbb{t}\right\}$, given hour-of-the-day (HR_{t}).

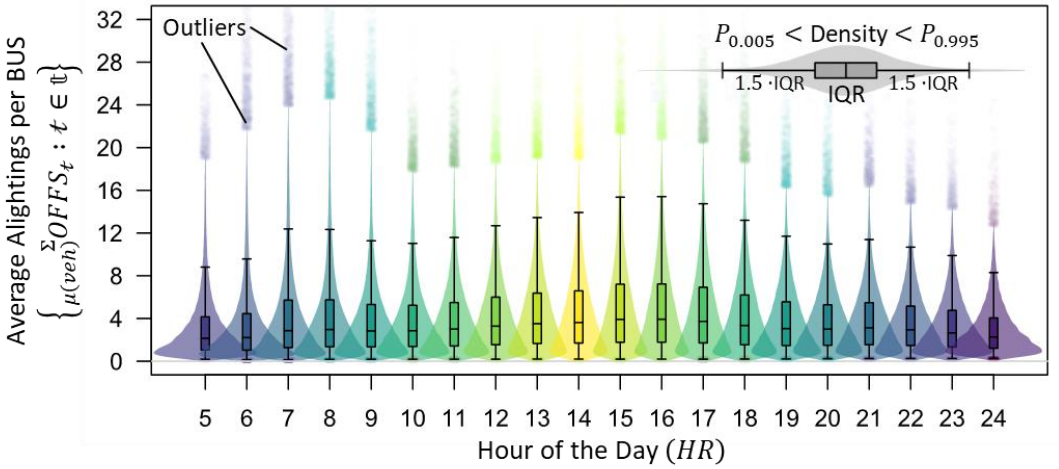

Figure C-3 — Violin and box-plots for all timepoint segments. Average alightings per vehicle, $\left\{{}_{\mu(veh)}^{\Sigma}\hat{O}FFS_t : t \in \mathbb{t}\right\}$, given hour-of-the-day (HR_t).

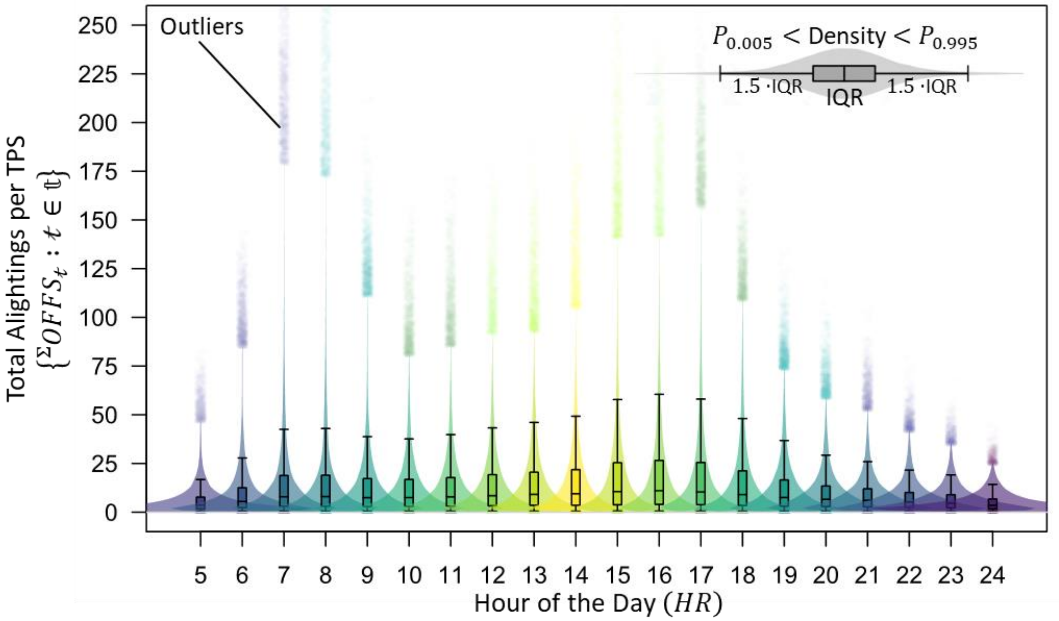

Figure C-4 — Violin and box-plots for all timepoint segments. Total alightings per TPS, $\left\{{}^{\Sigma}\hat{O}FFS_t : t \in \mathbb{t}\right\}$, given hour-of-the-day (HR_t).

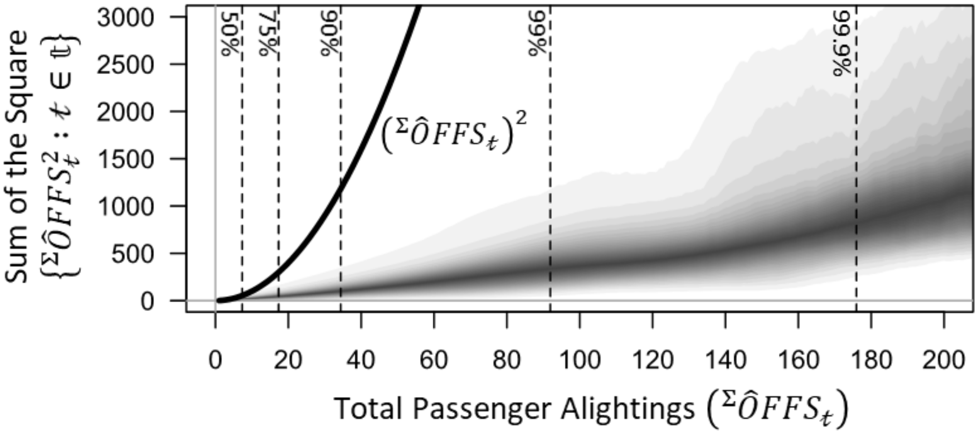

Figure C-5 — Sum of the square and square of the sum for passenger alightings.

C.2. Models with Composite Variables

Table C-1 — $\{{}^{\Sigma T}\widehat{D}WL_{t} : t \in \mathbb{t}\}$ aggregated linear regression model using composite frequency and time variables based on ${}^{\Sigma}VEH_{t}$. In-text summary and comparison given by Table 5-14 on page 143.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	-3.81	0.0996		
Number of Vehicles	${}^{\Sigma}VEH$	5.07	0.0462	6.78%	9.23%
Total Passenger Movements	${}^{\Sigma}\hat{O}NS$	3.79	0.0067	13.28%	18.10%
	${}^{\Sigma}\hat{O}FFS$	1.49	0.0067	7.64%	10.41%
	$({}^{\Sigma}\hat{O}NS)^{2}$	-0.003	0.0000	5.46%	7.44%
	$({}^{\Sigma}\hat{O}FFS)^{2}$	-0.001	0.0000	3.22%	4.39%
	${}^{\Sigma}\hat{L}IFT$	39.52	0.1017	3.46%	4.72%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	15.07	0.0554	0.88%	1.20%
	${}^{\Sigma L}MALL$	9.99	0.0310	1.14%	1.55%
	${}^{\Sigma L}NEAR$	5.17	0.0187	5.94%	8.10%
	${}^{\Sigma L}FAR$	4.89	0.0271	4.01%	5.46%
	${}^{\Sigma L}OPP$	5.89	0.0382	1.51%	2.06%
	${}^{\Sigma L}AT$	7.43	0.0451	0.93%	1.27%
Total Serviced Bus Stop Locations Near Traffic Signals (ΣLs)	${}^{\Sigma Ls}TC$	4.24	0.0908	0.49%	0.67%
	${}^{\Sigma Ls}NEAR$	1.23	0.0234	4.97%	6.77%
	${}^{\Sigma Ls}FAR$	1.58	0.0329	3.50%	4.77%
	${}^{\Sigma Ls}OPP$	1.30	0.0716	0.60%	0.82%
	${}^{\Sigma Ls}AT$	5.37	0.0738	0.74%	1.01%
High-Frequency RTE	${}^{\Sigma}MILES \times FREQ$	1.64	0.0182	5.35%	7.29%
Weekdays	${}^{\Sigma}MILES \times W_{1}^{AM}$	-3.44	0.0233	0.37%	0.51%
	${}^{\Sigma}MILES \times W_{1}^{PM}$	-0.96	0.0204	1.05%	1.43%
Weekends	${}^{\Sigma}MILES \times W_{0}^{P}$	1.91	0.0301	0.31%	0.43%
Total Vehicle Interactions at Bus Stops Between Different RTE (ΣId)	${}^{\Sigma Id}LEAD$	2.98	0.1255	0.81%	1.10%
	${}^{\Sigma Id}WAIT$	25.22	0.3374	0.36%	0.49%
	${}^{\Sigma Id}JUMP$	-1.26	0.3266	0.16%	0.22%
Total Vehicle Interactions at Bus Stops Within the Same RTE (ΣIs)	${}^{\Sigma Is}INT$	-2.51	0.1377	0.42%	0.57%

$n = 4{,}525{,}801$ Adjusted R-Squared $= 73.38\%$

p-value $\ll 0.001$ or all variables

Table C-2 — $\{{}^{T}\widehat{B}AY_i : t \in \mathbb{t}\}$ aggregated linear regression model using composite frequency and time variables based on ${}^{\Sigma}VEH_{t}$. In-text summary and comparison given by Table 5-17 on page 148.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	-20.32	0.1915		
Number of Vehicles	${}^{\Sigma}VEH$	19.92	0.1009	8.56%	10.92%
Total Passenger Movements	${}^{\Sigma}\hat{O}NS$	5.54	0.0120	12.29%	15.69%
	${}^{\Sigma}\hat{O}FFS$	2.04	0.0079	9.09%	11.61%
	$({}^{\Sigma}\hat{O}NS)^2$	-0.008	0.0001	5.07%	6.47%
	${}^{\Sigma}\hat{L}IFT$	42.25	0.1835	2.37%	3.02%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	34.17	0.0980	0.88%	1.12%
	${}^{\Sigma L}MALL$	31.34	0.0591	1.96%	2.50%
	${}^{\Sigma L}NEAR$	14.62	0.0330	8.04%	10.26%
	${}^{\Sigma L}FAR$	12.25	0.0475	3.97%	5.07%
	${}^{\Sigma L}OPP$	11.53	0.0674	1.51%	1.93%
	${}^{\Sigma L}AT$	17.28	0.0792	0.81%	1.03%
Total Serviced Bus Stop Locations Near Traffic Signals (ΣLs)	${}^{\Sigma Ls}TC$	10.90	0.1638	0.60%	0.76%
	${}^{\Sigma Ls}NEAR$	8.56	0.0422	7.26%	9.27%
	${}^{\Sigma Ls}FAR$	1.50	0.0595	3.66%	4.67%
	${}^{\Sigma Ls}OPP$	6.45	0.1287	0.68%	0.86%
	${}^{\Sigma Ls}AT$	2.26	0.1333	0.61%	0.78%
High-Frequency RTE	${}^{\Sigma}VEH \times FREQ$	0.70	0.0618	5.54%	7.08%
Weekdays	${}^{\Sigma}VEH \times W_1^{AM}$	-6.39	0.0773	0.52%	0.66%
	${}^{\Sigma}VEH \times W_1^{PM}$	3.11	0.0683	1.59%	2.03%
Weekends	${}^{\Sigma}VEH \times W_0^{P}$	4.72	0.0939	0.21%	0.27%
Total Vehicle Interactions at Bus Stops Between Different RTE (ΣId)	${}^{\Sigma Id}LEAD$	10.98	0.2293	1.00%	1.28%
	${}^{\Sigma Id}TAIL$	11.18	0.2191	0.89%	1.14%
	${}^{\Sigma Id}WAIT$	52.74	0.6115	0.40%	0.51%
	${}^{\Sigma Id}JUMP$	4.26	0.5945	0.22%	0.28%
Total Vehicle Interactions at Bus Stops Within the Same RTE (ΣIs)	${}^{\Sigma Is}INT$	9.23	0.2453	0.61%	0.78%
	$n = 4,525,801$ Adjusted R-Squared = 78.36% p-value $\ll 0.001$ or all variables				

Table C-3 — $\{{}^{T}\widehat{B}AY_i : t \in \mathbb{t}\}$ aggregated linear regression model using composite frequency and time variables based on ${}^{\Sigma}MILES_t$. In-text summary and comparison given by Table 5-18 on page 148.

Variable Type	**Variable**	**Coefficient**	**Std Error**	**Contrib.**	**Rel-Imp**
Calculated Intercept	Intercept	-19.96	0.1792		
Number of Vehicles	${}^{\Sigma}VEH$	20.10	0.0834	8.78%	11.20%
Total Passenger Movements	${}^{\Sigma}\hat{O}NS$	5.57	0.0121	12.29%	15.69%
	${}^{\Sigma}\hat{O}FFS$	2.04	0.0080	9.09%	11.59%
	$({}^{\Sigma}\hat{O}NS)^2$	-0.008	0.0001	5.10%	6.50%
	${}^{\Sigma}\hat{L}IFT$	42.34	0.1834	2.36%	3.01%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	33.84	0.0990	0.87%	1.11%
	${}^{\Sigma L}MALL$	31.07	0.0592	2.00%	2.55%
	${}^{\Sigma L}NEAR$	14.44	0.0335	8.03%	10.25%
	${}^{\Sigma L}FAR$	12.02	0.0484	3.90%	4.97%
	${}^{\Sigma L}OPP$	11.23	0.0687	1.42%	1.82%
	${}^{\Sigma L}AT$	17.00	0.0808	0.76%	0.98%
Total Serviced Bus Stop Locations Near Traffic Signals (ΣLs)	${}^{\Sigma Ls}TC$	10.89	0.1641	0.58%	0.74%
	${}^{\Sigma Ls}NEAR$	8.60	0.0421	7.26%	9.27%
	${}^{\Sigma Ls}FAR$	1.53	0.0594	3.58%	4.57%
	${}^{\Sigma Ls}OPP$	6.60	0.1289	0.67%	0.86%
	${}^{\Sigma Ls}AT$	2.09	0.1331	0.59%	0.75%
High-Frequency RTE	${}^{\Sigma}MILES \times FREQ$	0.84	0.0329	5.71%	7.28%
Weekdays	${}^{\Sigma}MILES \times W_1^{AM}$	-3.38	0.0419	0.49%	0.62%
	${}^{\Sigma}MILES \times W_1^{PM}$	1.28	0.0368	1.41%	1.80%
Weekends	${}^{\Sigma}MILES \times W_0^{P}$	3.11	0.0544	0.27%	0.34%
Total Vehicle Interactions at Bus Stops Between Different RTE (ΣId)	${}^{\Sigma Id}LEAD$	11.16	0.2290	1.03%	1.32%
	${}^{\Sigma Id}TAIL$	11.40	0.2189	0.92%	1.17%
	${}^{\Sigma Id}WAIT$	52.95	0.6116	0.41%	0.52%
	${}^{\Sigma Id}JUMP$	4.66	0.5945	0.22%	0.28%
Total Vehicle Interactions at Bus Stops Within the Same RTE (ΣIs)	${}^{\Sigma Is}INT$	9.33	0.2452	0.62%	0.79%
	$n = 4{,}525{,}801$		Adjusted R-Squared = 78.35%		
	p-value $\ll 0.001$ or all variables				

Table C-4 — $\{^{\Sigma T}\widetilde{D}STB_{t} : t \in \mathbb{t}\}$ aggregated linear regression model using composite frequency and time variables based on ${}^{\Sigma}VEH_{t}$. In-text summary and comparison given by Table 5-21 on page 152.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	2.43	0.1780		
Number of Vehicles	${}^{\Sigma}VEH$	30.97	0.0929	7.10%	22.85%
Total Distance in Miles	${}^{\Sigma}MILES$	5.23	0.0335	3.68%	11.84%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	9.39	0.0796	0.69%	2.22%
	${}^{\Sigma L}MALL$	2.234	0.0437	0.61%	1.95%
	${}^{\Sigma L}NEAR$	-0.335	0.0216	0.99%	3.20%
	${}^{\Sigma L}FAR$	0.40	0.0302	1.37%	4.40%
	${}^{\Sigma L}OPP$	1.36	0.0543	0.29%	0.94%
	${}^{\Sigma L}AT$	2.44	0.0646	0.55%	1.76%
Total Non-Serviced Bus Stop Locations $\binom{\Sigma L}{thru}$	${}^{\Sigma L}_{thru}TC$	-1.15	0.2154	0.05%	0.15%
	${}^{\Sigma L}_{thru}NEAR$	-1.49	0.0132	0.58%	1.88%
	${}^{\Sigma L}_{thru}FAR$	-2.99	0.0232	0.37%	1.18%
	${}^{\Sigma L}_{thru}OPP$	-0.87	0.0298	0.14%	0.46%
	${}^{\Sigma L}_{thru}AT$	0.56	0.0396	0.21%	0.66%
Total Scheduled Bus Stop Locations near Traffic Signals $\binom{\Sigma Ls}{skd}$	${}^{\Sigma Ls}_{skd}TC$	6.09	0.1313	0.44%	1.41%
	${}^{\Sigma Ls}_{skd}NEAR$	1.35	0.0197	1.01%	3.26%
	${}^{\Sigma Ls}_{skd}FAR$	4.98	0.0297	1.92%	6.18%
	${}^{\Sigma Ls}_{skd}OPP$	-3.37	0.0604	0.10%	0.33%
	${}^{\Sigma Ls}_{skd}AT$	4.09	0.0727	0.48%	1.55%
High-Frequency RTE	${}^{\Sigma}VEH \times FREQ$	-9.05	0.0521	1.79%	5.76%
Weekdays	${}^{\Sigma}VEH \times W_{1}^{AM}$	3.77	0.0629	0.52%	1.68%
	${}^{\Sigma}VEH \times W_{1}^{PM}$	18.86	0.0554	5.11%	16.45%
Weekends	${}^{\Sigma}VEH \times W_{0}^{P}$	4.44	0.0768	0.07%	0.23%
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	${}^{\Sigma Id}INT$	13.32	0.1012	1.99%	6.39%
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	${}^{\Sigma Is}INT$	26.16	0.1922	1.02%	3.27%
	$n = 3{,}684{,}962$		Adjusted R-Squared $= 31.09\%$		
	p-value $\ll 0.001$ or all variables				

Table C-5 — $\{{}^{\Sigma T}\widetilde{D}STB_{t} : t \in \mathbb{t}\}$ aggregated linear regression model using composite frequency and time variables based on ${}^{\Sigma}MILES_{t}$. In-text summary and comparison given by Table 5-22 on page 152.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	4.57	0.1716		
Number of Vehicles	${}^{\Sigma}VEH$	29.69	0.0718	7.81%	25.18%
Total Distance in Miles	${}^{\Sigma}MILES$	3.64	0.0368	3.44%	11.09%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	10.63	0.0790	0.84%	2.71%
	${}^{\Sigma L}MALL$	2.269	0.0433	0.74%	2.40%
	${}^{\Sigma L}FAR$	0.658	0.0303	1.79%	5.78%
	${}^{\Sigma L}OPP$	2.01	0.0545	0.36%	1.16%
	${}^{\Sigma L}AT$	2.72	0.0642	0.62%	2.01%
Total Non-Serviced Bus Stop Locations $\left(\begin{smallmatrix}\Sigma L\\ thru\end{smallmatrix}\right)$	${}^{\Sigma L}_{thru}TC$	-2.48	0.2151	0.06%	0.18%
	${}^{\Sigma L}_{thru}NEAR$	-1.24	0.0132	0.54%	1.74%
	${}^{\Sigma L}_{thru}FAR$	-2.77	0.0228	0.35%	1.12%
	${}^{\Sigma L}_{thru}OPP$	-0.47	0.0298	0.13%	0.41%
	${}^{\Sigma L}_{thru}AT$	1.18	0.0395	0.20%	0.65%
Total Scheduled Bus Stop Locations near Traffic Signals $\left(\begin{smallmatrix}\Sigma Ls\\ skd\end{smallmatrix}\right)$	${}^{\Sigma Ls}_{skd}TC$	4.85	0.1313	0.45%	1.46%
	${}^{\Sigma Ls}_{skd}NEAR$	1.33	0.0161	1.06%	3.43%
	${}^{\Sigma Ls}_{skd}FAR$	5.05	0.0296	2.00%	6.44%
	${}^{\Sigma Ls}_{skd}OPP$	-3.78	0.0604	0.10%	0.33%
	${}^{\Sigma Ls}_{skd}AT$	4.32	0.0727	0.47%	1.50%
High-Frequency RTE	${}^{\Sigma}MILES \times FREQ$	-4.84	0.0292	1.47%	4.75%
Weekdays	${}^{\Sigma}MILES \times W_{1}^{AM}$	2.98	0.0353	0.60%	1.93%
	${}^{\Sigma}MILES \times W_{1}^{PM}$	10.50	0.0311	4.73%	15.25%
Weekends	${}^{\Sigma}MILES \times W_{0}^{P}$	2.48	0.0442	0.06%	0.21%
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	${}^{\Sigma Id}INT$	14.80	0.1008	2.18%	7.02%
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	${}^{\Sigma Is}INT$	26.04	0.1922	1.01%	3.26%

$n = 3{,}684{,}302$ Adjusted R-Squared = 31.02%

p-value $\ll 0.001$ or all variables

Table C-6 — $\left\{{}^{\Sigma T}\widetilde{M}OVE_{t} : t \in \mathbb{t}\right\}$ aggregated linear regression model using composite frequency and time variables based on ${}^{\Sigma}VEH_{t}$. In-text summary and comparison given by Table 5-27 on page 158.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	-21.36	0.2060		
Number of Vehicles	${}^{\Sigma}VEH$	57.28	0.1149	13.43%	14.75%
Total Distance in Miles	${}^{\Sigma}MILES$	90.89	0.0438	26.73%	29.36%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	58.22	0.1004	1.26%	1.38%
	${}^{\Sigma L}MALL$	32.244	0.0584	0.88%	0.96%
	${}^{\Sigma L}NEAR$	16.761	0.0280	6.59%	7.24%
	${}^{\Sigma L}FAR$	13.34	0.0294	4.98%	5.47%
	${}^{\Sigma L}OPP$	7.79	0.0706	2.04%	2.24%
	${}^{\Sigma L}AT$	29.59	0.0832	1.95%	2.14%
Total Non-Serviced Bus Stop Locations $\binom{\Sigma L}{thru}$	${}^{\Sigma L}_{thru}NEAR$	8.16	0.0171	5.15%	5.66%
	${}^{\Sigma L}_{thru}FAR$	3.72	0.0280	3.26%	3.58%
	${}^{\Sigma L}_{thru}OPP$	2.06	0.0379	1.69%	1.86%
	${}^{\Sigma L}_{thru}AT$	4.06	0.0510	1.29%	1.42%
Total Scheduled Bus Stop Locations near Traffic Signals $\binom{\Sigma Ls}{skd}$	${}^{\Sigma Ls}_{skd}TC$	15.33	0.1737	1.16%	1.27%
	${}^{\Sigma Ls}_{skd}NEAR$	-2.03	0.0256	6.73%	7.39%
	${}^{\Sigma Ls}_{skd}OPP$	0.60	0.0787	1.05%	1.16%
	${}^{\Sigma Ls}_{skd}AT$	4.24	0.0962	1.33%	1.46%
High-Frequency RTE	${}^{\Sigma}VEH \times FREQ$	-9.85	0.0666	6.20%	6.81%
Weekdays	${}^{\Sigma}VEH \times W_{1}^{AM}$	4.99	0.0823	1.06%	1.17%
	${}^{\Sigma}VEH \times W_{1}^{PM}$	17.52	0.0726	2.21%	2.43%
Weekends	${}^{\Sigma}VEH \times W_{0}^{P}$	4.87	0.1006	0.19%	0.21%
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	${}^{\Sigma Id}INT$	17.39	0.1356	1.21%	1.33%
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	${}^{\Sigma Is}INT$	19.75	0.2590	0.65%	0.72%
	$n = 4{,}524{,}128$		Adjusted R-Squared = 91.04%		
	p-value $\ll 0.001$ or all variables				

Table C-7 — $\left\{{}^{\Sigma T}\widetilde{M}OVE_{t} : t \in \mathbb{t}\right\}$ aggregated linear regression model using composite frequency and time variables based on ${}^{\Sigma}MILES_{t}$. In-text summary and comparison given by Table 5-28 on page 158.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	Intercept	-21.69	0.1984		
Number of Vehicles	${}^{\Sigma}VEH$	55.09	0.0925	13.18%	14.46%
Total Distance in Miles	${}^{\Sigma}MILES$	89.14	0.0476	24.25%	26.60%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	60.41	0.0994	1.20%	1.31%
	${}^{\Sigma L}MALL$	33.182	0.0581	0.88%	0.97%
	${}^{\Sigma L}NEAR$	17.228	0.0278	6.36%	6.97%
	${}^{\Sigma L}FAR$	14.27	0.0295	4.67%	5.12%
	${}^{\Sigma L}OPP$	9.09	0.0651	1.88%	2.06%
	${}^{\Sigma L}AT$	30.17	0.0827	1.77%	1.94%
Total Non-Serviced Bus Stop Locations $\binom{\Sigma L}{thru}$	${}^{\Sigma L}_{thru}NEAR$	8.70	0.0171	4.89%	5.36%
	${}^{\Sigma L}_{thru}FAR$	4.21	0.0278	3.02%	3.31%
	${}^{\Sigma L}_{thru}OPP$	2.62	0.0373	1.51%	1.66%
	${}^{\Sigma L}_{thru}AT$	5.34	0.0508	1.13%	1.24%
Total Scheduled Bus Stop Locations near Traffic Signals $\binom{\Sigma Ls}{skd}$	${}^{\Sigma Ls}_{skd}TC$	14.18	0.1719	1.17%	1.29%
	${}^{\Sigma Ls}_{skd}NEAR$	-1.73	0.0253	7.04%	7.73%
	${}^{\Sigma Ls}_{skd}AT$	4.74	0.0955	1.31%	1.44%
High-Frequency RTE	${}^{\Sigma}MILES \times FREQ$	-8.34	0.0379	8.67%	9.51%
Weekdays	${}^{\Sigma}MILES \times W_1^{AM}$	7.51	0.0465	2.16%	2.36%
	${}^{\Sigma}MILES \times W_1^{PM}$	11.26	0.0410	3.77%	4.13%
Weekends	${}^{\Sigma}MILES \times W_0^{P}$	3.45	0.0582	0.41%	0.45%
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	${}^{\Sigma Id}INT$	17.51	0.1340	1.24%	1.36%
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	${}^{\Sigma Is}INT$	18.96	0.2569	0.66%	0.72%

$n = 4{,}524{,}128$ Adjusted R-Squared $= 91.17\%$

p-value $\ll 0.001$ or all variables

Table C-8 — $\left\{{}^{\Sigma T}\widetilde{T}RVL_{t} : t \in \mathbb{t}\right\}$ aggregated linear regression model using composite frequency and time variables based on ${}^{\Sigma}VEH_{t}$. In-text summary and comparison given by Table 5-31 on page 163.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	*Intercept*	-57.90	0.3557		
Number of Vehicles	${}^{\Sigma}VEH$	103.91	0.1979	10.03%	11.10%
Total Distance in Miles	${}^{\Sigma}MILES$	95.67	0.0751	13.69%	15.15%
Total Passenger Movements (Σ)	${}^{\Sigma}\hat{O}NS$	7.06	0.0226	7.31%	8.09%
	${}^{\Sigma}\hat{O}FFS$	4.36	0.0228	6.80%	7.52%
	${}^{\Sigma}\hat{O}NS^{2}$	-0.013	0.0001	2.94%	3.26%
	${}^{\Sigma}\hat{O}FFS^{2}$	-0.006	0.0001	2.76%	3.06%
	${}^{\Sigma}\hat{L}IFT$	40.49	0.3370	1.16%	1.29%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	96.24	0.1833	1.07%	1.19%
	${}^{\Sigma L}MALL$	62.87	0.1094	1.11%	1.23%
	${}^{\Sigma L}NEAR$	29.92	0.0595	5.67%	6.27%
	${}^{\Sigma L}FAR$	22.27	0.0767	3.94%	4.36%
	${}^{\Sigma L}OPP$	19.56	0.1167	1.48%	1.64%
	${}^{\Sigma L}AT$	45.12	0.1500	1.27%	1.41%
Total Non-Serviced Bus Stop Locations $\left({}^{\Sigma L}_{thru}\right)$	${}^{\Sigma L}_{thru}NEAR$	6.95	0.0299	2.71%	3.00%
	${}^{\Sigma L}_{thru}FAR$	1.75	0.0518	1.70%	1.88%
	${}^{\Sigma L}_{thru}OPP$	-0.34	0.0644	0.76%	0.84%
	${}^{\Sigma L}_{thru}AT$	5.97	0.0880	0.59%	0.65%
Total Scheduled Bus Stop Locations near Traffic Signals $\left({}^{\Sigma Ls}_{skd}\right)$	${}^{\Sigma Ls}_{skd}TC$	31.22	0.2966	0.71%	0.78%
	${}^{\Sigma Ls}_{skd}NEAR$	3.41	0.0442	4.14%	4.59%
	${}^{\Sigma Ls}_{skd}FAR$	6.86	0.0662	3.21%	3.55%
	${}^{\Sigma Ls}_{skd}AT$	10.02	0.1660	0.64%	0.71%
High-Frequency RTE	${}^{\Sigma}VEH \times FREQ$	-18.73	0.1145	4.64%	5.14%
Weekdays	${}^{\Sigma}VEH \times W_{1}^{AM}$	2.64	0.1431	0.67%	0.75%
	${}^{\Sigma}VEH \times W_{1}^{PM}$	38.43	0.1262	1.93%	2.13%
Weekends	${}^{\Sigma}VEH \times W_{0}^{P}$	14.03	0.1723	0.16%	0.18%
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	${}^{\Sigma Id}LEAD$	37.51	0.4207	3.43%	3.79%
	${}^{\Sigma Id}TAIL$	38.79	0.4023	3.18%	3.52%
	${}^{\Sigma Id}WAIT$	92.02	1.1219	1.20%	1.33%
	${}^{\Sigma Id}JUMP$	56.34	1.0908	0.89%	0.99%
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	${}^{\Sigma Is}INT$	52.99	0.4566	0.55%	0.61%

$n = 4{,}525{,}799$ Adjusted R-Squared $= 90.35\%$

p-value $\ll 0.001$ or all variables

Table C-9 — $\left\{{}^{\Sigma T}\widetilde{T}RVL_{t} : t \in \mathbb{t}\right\}$ aggregated linear regression model using composite frequency and time variables based on ${}^{\Sigma}MILES_{t}$. In-text summary and comparison given by Table 5-32 on page 163.

Variable Type	Variable	Coefficient	Std Error	Contrib.	Rel-Imp
Calculated Intercept	*Intercept*	-55.77	0.3452		
Number of Vehicles	${}^{\Sigma}VEH$	100.08	0.1602	10.12%	11.19%
Total Distance in Miles	${}^{\Sigma}MILES$	92.17	0.0824	12.60%	13.94%
Total Passenger Movements (Σ)	${}^{\Sigma}\hat{O}NS$	7.19	0.0225	7.26%	8.03%
	${}^{\Sigma}\hat{O}FFS$	4.57	0.0227	6.74%	7.45%
	${}^{\Sigma}\hat{O}NS^{2}$	-0.014	0.0001	2.94%	3.25%
	${}^{\Sigma}\hat{O}FFS^{2}$	-0.009	0.0001	2.76%	3.05%
	${}^{\Sigma}\hat{L}IFT$	43.03	0.3359	1.17%	1.29%
Total Serviced Bus Stop Locations (ΣL)	${}^{\Sigma L}TC$	98.50	0.1836	1.06%	1.17%
	${}^{\Sigma L}MALL$	63.55	0.1092	1.15%	1.27%
	${}^{\Sigma L}NEAR$	30.47	0.0595	5.60%	6.20%
	${}^{\Sigma L}FAR$	23.09	0.0771	3.83%	4.23%
	${}^{\Sigma L}OPP$	20.97	0.1168	1.39%	1.54%
	${}^{\Sigma L}AT$	45.53	0.1498	1.18%	1.31%
Total Non-Serviced Bus Stop Locations $\binom{\Sigma L}{thru}$	${}^{\Sigma L}_{thru}NEAR$	7.71	0.0300	2.62%	2.90%
	${}^{\Sigma L}_{thru}FAR$	2.43	0.0518	1.59%	1.75%
	${}^{\Sigma L}_{thru}OPP$	0.50	0.0643	0.70%	0.78%
	${}^{\Sigma L}_{thru}AT$	7.72	0.0882	0.53%	0.58%
Total Scheduled Bus Stop Locations near Traffic Signals $\binom{\Sigma Ls}{skd}$	${}^{\Sigma Ls}_{skd}TC$	29.26	0.2958	0.68%	0.75%
	${}^{\Sigma Ls}_{skd}NEAR$	3.66	0.0440	4.09%	4.53%
	${}^{\Sigma Ls}_{skd}FAR$	6.89	0.0661	3.10%	3.43%
	${}^{\Sigma Ls}_{skd}AT$	10.86	0.1655	0.63%	0.69%
High-Frequency RTE	${}^{\Sigma}MILES \times FREQ$	-12.49	0.0653	5.71%	6.32%
Weekdays	${}^{\Sigma}MILES \times W_{1}^{AM}$	7.93	0.0809	1.10%	1.22%
	${}^{\Sigma}MILES \times W_{1}^{PM}$	23.09	0.0716	2.53%	2.80%
Weekends	${}^{\Sigma}MILES \times W_{0}^{P}$	9.09	0.1002	0.29%	0.32%
Total Vehicle Interactions at Bus Stops Between Different $ROUTE$s (ΣId)	${}^{\Sigma Id}LEAD$	39.80	0.4189	3.36%	3.72%
	${}^{\Sigma Id}TAIL$	40.48	0.4007	3.13%	3.46%
	${}^{\Sigma Id}WAIT$	92.25	1.1188	1.13%	1.25%
	${}^{\Sigma Id}JUMP$	56.29	1.0879	0.87%	0.96%
Total Vehicle Interactions at Bus Stops Within the Same $ROUTE$ (ΣIs)	${}^{\Sigma Is}INT$	54.19	0.4551	0.56%	0.62%
	$n = 4{,}525{,}799$ Adjusted R-Squared $= 90.41\%$ *p-value* $\ll 0.001$ or all variables				

C.3. Speed Regression Models

Table C-10 — Average total travel speed adjusted R-squared for one-variable and two-variable models.

	1-Variable Models	2-Variable Models			
Variable		$^{\Sigma}VEH$	$^{\Sigma}\hat{O}NS$	$^{\Sigma}\hat{O}FFS$	$\sum[^{\Sigma L}VAR]$
$^{\Sigma}\hat{O}NS$	0.1003	0.1173	-NA-	0.1195	0.1089
$\sum[^{\Sigma L}VAR]$	0.0934	0.1022	0.1089	0.0993	-NA-
$^{\Sigma}VEH$	0.0892	-NA-	0.1173	0.1048	0.1022
$^{\Sigma}\hat{O}FFS$	0.0820	0.1048	0.1195	-NA-	0.0993
$^{\Sigma}VEH \times FREQ$	0.0534	0.0892	0.1054	0.0900	0.0944
$\sum[^{\Sigma Ls}_{skd}VAR]$	0.0449	0.0892	0.1004	0.0830	0.0989
$^{\Sigma Id}INT$	0.0407	0.1085	0.1132	0.0968	0.1129
$^{\Sigma}\hat{L}IFT$	0.0295	0.0977	0.1053	0.0888	0.0979
$^{\Sigma}VEH \times W_1^{PM}$	0.0281	0.0923	0.1046	0.0877	0.0977
$\sum[^{\Sigma L}_{thru}VAR]$	0.0078	0.2030	0.1440	0.1214	0.2001
$^{\Sigma}VEH \times W_0^{P}$	0.0056	0.0903	0.1027	0.0847	0.0950
$^{\Sigma}VEH \times W_1^{AM}$	0.0046	0.0899	0.1003	0.0822	0.0934
$^{\Sigma Is}INT$	0.0030	0.0892	0.1003	0.0826	0.0936

Table C-11 — Average moving speed adjusted R-squared for one-variable and two-variable models.

	1-Variable Models	2-Variable Models			
Variable		$^{\Sigma}VEH$	$^{\Sigma}\hat{O}NS$	$^{\Sigma}\hat{O}FFS$	$\sum[^{\Sigma L}VAR]$
$^{\Sigma}VEH$	0.0387	-NA-	0.0430	0.0422	0.0393
$^{\Sigma Id}INT$	0.0353	0.0607	0.0506	0.0502	0.0457
$^{\Sigma}\hat{O}NS$	0.0303	0.0430	-NA-	0.0387	0.0303
$^{\Sigma}\hat{O}FFS$	0.0293	0.0422	0.0387	-NA-	0.0294
$^{\Sigma}VEH \times FREQ$	0.0212	0.0387	0.0340	0.0331	0.0242
$\sum[^{\Sigma L}_{thru}VAR]$	0.0210	0.1483	0.0774	0.0765	0.0965
$\sum[^{\Sigma L}VAR]$	0.0195	0.0393	0.0303	0.0294	-NA-
$^{\Sigma}VEH \times W_1^{PM}$	0.0174	0.0423	0.0363	0.0354	0.0273
$^{\Sigma}\hat{L}IFT$	0.0071	0.0396	0.0311	0.0302	0.0208
$\sum[^{\Sigma Ls}_{skd}VAR]$	0.0031	0.0515	0.0360	0.0351	0.0311
$^{\Sigma}VEH \times W_1^{AM}$	0.0029	0.0388	0.0306	0.0297	0.0199
$^{\Sigma Is}INT$	0.0018	0.0387	0.0304	0.0293	0.0196
$^{\Sigma}VEH \times W_0^{P}$	0.0018	0.0389	0.0311	0.0301	0.0201

Table C-12 — Percent change to adjusted R-squared for average total travel speed one-variable (column) models adding row-variables.

	Percent Change				
Variable	${}^{\Sigma}VEH$	${}^{\Sigma}\hat{O}NS$	${}^{\Sigma}\hat{O}FFS$	$\sum[{}^{\Sigma L}VAR]$	**Average**
$\sum\left[{}^{\Sigma L}_{thru}VAR\right]$	128%	44%	48%	114%	83%
${}^{\Sigma}\hat{O}NS$	32%	-NA-	46%	17%	23%
${}^{\Sigma Id}INT$	22%	13%	18%	21%	18%
${}^{\Sigma}VEH$	-NA-	17%	28%	9%	14%
$\sum[{}^{\Sigma L}VAR]$	15%	9%	21%	-NA-	11%
${}^{\Sigma}\hat{O}FFS$	18%	19%	-NA-	6%	11%
${}^{\Sigma}\hat{L}IFT$	10%	5%	8%	5%	7%
${}^{\Sigma}VEH \times W_1^{PM}$	3%	4%	7%	5%	5%
${}^{\Sigma}VEH \times FREQ$	<0.1%	5%	10%	1%	4%
${}^{\Sigma}VEH \times W_0^{P}$	1%	2%	3%	2%	2%
$\sum\left[{}^{\Sigma Ls}_{skd}VAR\right]$	<0.1%	<0.1%	1%	6%	2%
${}^{\Sigma}VEH \times W_1^{AM}$	<1%	<0.1%	<1%	<0.1%	0%
${}^{\Sigma Is}INT$	<0.1%	<0.01%	<1%	<1%	0%

Table C-13 — Change to adjusted R-squared for average total travel speed one-variable (column) models adding row-variables for non-serviced stops by types.

	Total Change				Percent Change				
Variable	${}^{\Sigma}VEH$	${}^{\Sigma}\hat{O}NS$	${}^{\Sigma}\hat{O}FFS$	$\sum[{}^{\Sigma L}V]$	${}^{\Sigma}VEH$	${}^{\Sigma}\hat{O}NS$	${}^{\Sigma}\hat{O}FFS$	$\sum[{}^{\Sigma L}V]$	**Avg**
$\sum\left[{}^{\Sigma L}_{thru}VAR\right]$	0.1138	0.0438	0.0394	0.1067	128%	44%	48%	114%	83%
${}^{\Sigma L}_{thru}FAR$	0.0669	0.0520	0.0453	0.0687	75%	52%	55%	74%	64%
${}^{\Sigma L}_{thru}OPP$	0.0765	0.0380	0.0385	0.0733	86%	38%	47%	78%	62%
${}^{\Sigma L}_{thru}AT$	0.0422	0.0275	0.0267	0.0356	47%	27%	33%	38%	36%
${}^{\Sigma L}_{thru}NEAR$	0.0201	0.0047	0.0033	0.0204	23%	5%	4%	22%	13%
${}^{\Sigma L}_{thru}TC$	0.0006	0.0033	0.0015	0.0025	<1%	3%	2%	3%	2%

APPENDIX D — LIST OF APPENDIX TABLES AND FIGURES

D.1. Appendix Tables

D.2. Appendix Figures