

3-2016

# Estimating Uncertainty of Bus Arrival Times and Passenger Occupancies

Vikash V. Gayah

*Pennsylvania State University - Main Campus*

Zhengyao Yu

*Pennsylvania State University - Main Campus*

Jonathan S. Wood

*Pennsylvania State University - Main Campus*

Follow this and additional works at: [http://scholarworks.sjsu.edu/mti\\_publications](http://scholarworks.sjsu.edu/mti_publications)



Part of the [Transportation Commons](#)

---

## Recommended Citation

Vikash V. Gayah, Zhengyao Yu, and Jonathan S. Wood. "Estimating Uncertainty of Bus Arrival Times and Passenger Occupancies" *Mineta Transportation Institute Publications* (2016).

This Report is brought to you for free and open access by SJSU ScholarWorks. It has been accepted for inclusion in Mineta Transportation Institute Publications by an authorized administrator of SJSU ScholarWorks. For more information, please contact [scholarworks@sjsu.edu](mailto:scholarworks@sjsu.edu).

# Estimating Uncertainty of Bus Arrival Times and Passenger Occupancies



MNTRC Report 12-56



# MINETA TRANSPORTATION INSTITUTE

## LEAD UNIVERSITY OF MNTRC

The Mineta Transportation Institute (MTI) was established by Congress in 1991 as part of the Intermodal Surface Transportation Equity Act (ISTEA) and was reauthorized under the Transportation Equity Act for the 21st century (TEA-21). MTI then successfully competed to be named a Tier I Center in 2002 and 2006 in the Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU). Most recently, MTI successfully competed in the Surface Transportation Extension Act of 2011 to be named a Tier I Transit-Focused University Transportation Center. The Institute is funded by Congress through the United States Department of Transportation's Office of the Assistant Secretary for Research and Technology (OST-R), University Transportation Centers Program, the California Department of Transportation (Caltrans), and by private grants and donations.

The Institute receives oversight from an internationally respected Board of Trustees whose members represent all major surface transportation modes. MTI's focus on policy and management resulted from a Board assessment of the industry's unmet needs and led directly to the choice of the San José State University College of Business as the Institute's home. The Board provides policy direction, assists with needs assessment, and connects the Institute and its programs with the international transportation community.

MTI's transportation policy work is centered on three primary responsibilities:

### Research

MTI works to provide policy-oriented research for all levels of government and the private sector to foster the development of optimum surface transportation systems. Research areas include: transportation security; planning and policy development; interrelationships among transportation, land use, and the environment; transportation finance; and collaborative labor-management relations. Certified Research Associates conduct the research. Certification requires an advanced degree, generally a Ph.D., a record of academic publications, and professional references. Research projects culminate in a peer-reviewed publication, available both in hardcopy and on TransWeb, the MTI website (<http://transweb.sjsu.edu>).

### Education

The educational goal of the Institute is to provide graduate-level education to students seeking a career in the development and operation of surface transportation programs. MTI, through San José State University, offers an AACSB-accredited Master of Science in Transportation Management and a graduate Certificate in Transportation Management that serve to prepare the nation's transportation managers for the 21st century. The master's degree is the highest conferred by the California State Univer-

sity system. With the active assistance of the California Department of Transportation, MTI delivers its classes over a state-of-the-art videoconference network throughout the state of California and via webcasting beyond, allowing working transportation professionals to pursue an advanced degree regardless of their location. To meet the needs of employers seeking a diverse workforce, MTI's education program promotes enrollment to under-represented groups.

### Information and Technology Transfer

MTI promotes the availability of completed research to professional organizations and journals and works to integrate the research findings into the graduate education program. In addition to publishing the studies, the Institute also sponsors symposia to disseminate research results to transportation professionals and encourages Research Associates to present their findings at conferences. The World in Motion, MTI's quarterly newsletter, covers innovation in the Institute's research and education programs. MTI's extensive collection of transportation-related publications is integrated into San José State University's world-class Martin Luther King, Jr. Library.

---

### DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation, University Transportation Centers Program and the California Department of Transportation, in the interest of information exchange. This report does not necessarily reflect the official views or policies of the U.S. government, State of California, or the Mineta Transportation Institute, who assume no liability for the contents or use thereof. This report does not constitute a standard specification, design standard, or regulation.

REPORT 12-56

# **ESTIMATING UNCERTAINTY OF BUS ARRIVAL TIMES AND PASSENGER OCCUPANCIES**

Vikash V. Gayah, Ph.D.  
Zhengyao Yu  
Jonathan S. Wood

March 2016

A publication of  
**Mineta National Transit  
Research Consortium**

College of Business  
San José State University  
San José, CA 95192-0219

# TECHNICAL REPORT DOCUMENTATION PAGE

<b>1. Report No.</b> CA-MNTRC-14-1246	<b>2. Government Accession No.</b>	<b>3. Recipient's Catalog No.</b>	
<b>4. Title and Subtitle</b> Estimating Uncertainty of Bus Arrival Times and Passenger Occupancies		<b>5. Report Date</b> March 2016	
		<b>6. Performing Organization Code</b>	
<b>7. Authors</b> Vikash V. Gayah, Ph.D., Zhengyao Yu and Jonathan S. Wood		<b>8. Performing Organization Report</b> MNTRC Report 12-56	
<b>9. Performing Organization Name and Address</b> Mineta National Transit Research Consortium College of Business San José State University San José, CA 95192-0219		<b>10. Work Unit No.</b>	
		<b>11. Contract or Grant No.</b> DTRT12-G-UTC21	
<b>12. Sponsoring Agency Name and Address</b> <div style="display: flex; justify-content: space-between;"> <div style="width: 45%;">           U.S. Department of Transportation            Office of the Assistant Secretary for            Research and Technology            University Transportation Centers Program            1200 New Jersey Avenue, SE            Washington, DC 20590         </div> <div style="width: 45%;">           The Thomas D. Larson Pennsylvania            Transportation Institute            Pennsylvania State University            201 Transportation Research Building            University Park, PA 16802-4710         </div> </div>		<b>13. Type of Report and Period Covered</b> Final Report	
		<b>14. Sponsoring Agency Code</b>	
<b>15. Supplemental Notes</b>			
<b>16. Abstract</b> <p>Travel time reliability and the availability of seating and boarding space are important indicators of bus service quality and strongly influence users' satisfaction and attitudes towards bus transit systems. With Automated Vehicle Location (AVL) and Automated Passenger Counter (APC) units becoming common on buses, some agencies have begun to provide real-time bus location and passenger occupancy information as a means to improve perceived transit reliability. Travel time prediction models have also been established based on AVL and APC data. However, existing travel time prediction models fail to provide an indication of the uncertainty associated with these estimates. This can cause a false sense of precision, which can lead to experiences associated with unreliable service. Furthermore, no existing models are available to predict individual bus occupancies at downstream stops to help travelers understand if there will be space available to board.</p> <p>The purpose of this project was to develop modeling frameworks to predict travel times (and associated uncertainties) as well as individual bus passenger occupancies. For travel times, accelerated failure-time survival models were used to predict the entire distribution of travel times expected. The survival models were found to be just as accurate as models developed using traditional linear regression techniques. However, the survival models were found to have smaller variances associated with predictions. For passenger occupancies, linear and count regression models were compared. The linear regression models were found to outperform count regression models, perhaps due to the additive nature of the passenger boarding process. Various modeling frameworks were tested and the best frameworks were identified for predictions at near stops (within five stops downstream) and far stops (further than eight stops). Overall, these results can be integrated into existing real-time transit information systems to improve the quality of information provided to passengers.</p>			
<b>17. Key Words</b> Bus travel time models; Bus passenger occupancy models; Survival models; Linear regression; Count regression		<b>18. Distribution Statement</b> No restrictions. This document is available to the public through The National Technical Information Service, Springfield, VA 22161	
<b>19. Security Classif. (of this report)</b> Unclassified	<b>20. Security Classif. (of this page)</b> Unclassified	<b>21. No. of Pages</b> 69	<b>22. Price</b> \$15.00

Copyright © 2016  
by **Mineta National Transit Research Consortium**  
All rights reserved

Library of Congress Catalog Card Number:  
2016934873

**To order this publication, please contact:**

Mineta National Transit Research Consortium  
College of Business  
San José State University  
San José, CA 95192-0219

Tel: (408) 924-7560  
Fax: (408) 924-7565  
Email: [mineta-institute@sjsu.edu](mailto:mineta-institute@sjsu.edu)

[transweb.sjsu.edu/mntrc/index.html](http://transweb.sjsu.edu/mntrc/index.html)

## **ACKNOWLEDGMENTS**

The authors thank the Centre Area Transportation Authority (CATA) for providing the data used in this project. Particular thanks to Hugh Mose for his willingness to collaborate and Ryan Harshbarger for his time and effort in securing the data. The authors also thank the Mineta National Transit Research Consortium for their support of this project, particularly Executive Director Karen Philbrick, Ph.D.; Publication Support Coordinator Joseph Mercado; and Editor and Webmaster Frances Cherman.

---

## TABLE OF CONTENTS

<b>Executive Summary</b>	<b>1</b>
<b>I. Introduction</b>	<b>4</b>
Motivation	4
Research Objectives	8
Organization	9
<b>II. Review of Previous Modeling Methods</b>	<b>10</b>
Models of Bus Travel Times	10
Models of Bus Passenger Demand	13
<b>III. Data Sources for Statistical Modeling</b>	<b>15</b>
Transit Data	15
Weather Data	17
Time of Day Split	17
Data Cleaning	18
<b>IV. Statistical Modeling Methods</b>	<b>22</b>
Linear Regression Model	22
Negative Binomial Regression Model	23
Accelerated Failure Time Survival Model	24
Quantile Regression Models	25
<b>V. Modeling Travel Time Uncertainty</b>	<b>27</b>
Regression Results	29
Model Testing and Validation	31
Uncertainty Comparisons	33
Factors Most Impacting Uncertainty	35
<b>VI. Modeling Passenger Occupancies</b>	<b>37</b>
Modeling Frameworks	39
Regression Results	41
Model Testing	43
Uncertainty Comparisons	46
<b>VII. Concluding Remarks</b>	<b>51</b>
<b>Appendix A: Modeling Outputs and Technical Details</b>	<b>53</b>



<b>Glossary of Abbreviations</b>	<b>62</b>
<b>Bibliography</b>	<b>63</b>
<b>About the Authors</b>	<b>68</b>
<b>Peer Review</b>	<b>69</b>

---

## LIST OF FIGURES

1. Historical Passenger Occupancy Information from the Swiss Federal Railway	5
2. Real-time Bus Location and Passenger Occupancies from CATA Application	6
3. (a) Real-time Bus Arrival Prediction through NextBus; and, (b) Real-time Bus Arrival Prediction through OneBusAway	7
4. Route Illustration for the Blue Loop	15
5. Average Predicted Travel Time and Relative Errors as a Function of Time_Period	32
6. Comparison of Confidence Interval with Fraction of Observations that are Observed within Confidence Interval	34
7. Size of Confidence Interval for Given Confidence Level	34
8. Graphical Depiction of the Three Modeling Frameworks Considered to Estimate Passenger Occupancies	39
9. Comparison of Model Accuracy Across Modeling Frameworks Both Without and With Interaction Terms	44
10. Accuracy of Best Passenger Occupancy Models as a Function of How Far Away Prediction is Being Made	45
11. Size of Confidence Interval for Given Confidence Level	47
12. Comparison of Confidence Interval with Fraction of Observations that are Observed within Confidence Interval	48

---

## LIST OF TABLES

1. Summary of Bus Arrival Time Prediction Literature	10
2. Summary of Bus Demand Prediction Literature	14
3. Definition of Time_Period Variable	18
4. Summary Statistics for Blue Loop Data Used in Model Development Before and After Data Cleaning	21
5. Summary Statistics for Data Used in Travel Time Models Representing Observed Travel Times from Stop 9 to Stop 15	28
6. Summary Measures for Travel Time Regression Models without Interactions	29
7. Summary Measures for Travel Time Regression Models Including Interaction Terms	31
8. Confidence Intervals for Travel Time Estimates Using Linear and Survival Regression Models	33
9. Elasticities Describing Influence on Travel Time Variance	35
10. Summary Statistics for Data Used in Passenger Occupancy Models	38
11. Summary Measures for Passenger Occupancy Regression Models without Interactions	41
12. Summary Measures for Passenger Occupancy Models Including Interaction Terms	42
13. Confidence Intervals for Passenger Occupancy Estimates Using Linear and Quantile Regression Models	47
14. Sensitivity of Confidence Interval Size to Various Independent Variables for Passenger Occupancy Model	49
15. Summary of Major Findings	52

---

## EXECUTIVE SUMMARY

Public transportation users identify reliability as a key measure of the quality of transit service and major determinant of transit use. Improving the reliability of a transit system can have numerous potential benefits, such as increased transit ridership, decreased congestion (which further improves transit reliability), and decreased negative externalities like greenhouse gas and other emissions. Unfortunately, bus transit systems are highly unstable, making it difficult to maintain reliable schedules for riders to use. To mitigate this, real-time information on the current state of the bus transit system can be used to update bus transit schedules (e.g., bus arrival times to each stop), increasing the *perceived* reliability of the system from a user perspective. However, a potential drawback of these real-time traveler information systems is that they can provide a false sense of precision. That is, users expecting a certain arrival time based on real-time information can develop increased negative feelings about the transit system if the bus is earlier or later than expected. A better approach might be to also provide users an indication of the uncertainty of these predictions. This would help to temper expectations, help users plan their trips more effectively, and minimize the occurrence of negative experiences. However, very little work has been done to model and/or quantify the uncertainty of real-time bus transit information.

Two pieces of real-time bus transit information were specifically considered in this project: bus arrival times and bus passenger occupancies upon arriving at each stop. Bus arrival time has been well studied in the transportation research literature and is generally provided to passengers through a variety of services. However, these existing efforts ignore the uncertainty associated with these estimates, knowledge of which can improve transit users' trip planning. Bus passenger occupancy appears to be ignored in the research literature entirely, although some agencies provide current bus passenger occupancies to passengers. However, anecdotal evidence suggests that passengers would value having predictions of passenger occupancy at downstream stops. This type of information can help to minimize the number of negative experiences (in this case, arriving to a full bus and being unable to board) that might reduce confidence in bus transit.

In light of this information, the present project describes new modeling techniques that can be used to provide estimates of bus travel times and expected occupancies of buses at downstream locations, as well as the uncertainty associated with these estimates. The report is created specifically for real-time bus transit information providers and data analysis teams within transit agencies. The report is intended to document the use of new techniques that can be used to improve the quality of information that is provided to transit users.

To model travel times, linear regression models—which appear to be a common practice in the literature—are compared with a newly proposed technique: accelerated failure time (AFT) survival models. AFT survival models (or more simply, survival or duration models) are used to predict the time remaining until an event occurs. In the transportation field, they have been used extensively to model time-to-failure of infrastructure elements. In the present case, the event considered is the arrival of the bus to a downstream stop. Survival models are ideal for this project because they predict the distribution of the dependent variable (in this case, expected travel times), which can provide the mean value as well as the variance of expected values around the mean. For passenger occupancy estimates,

two modeling approaches are proposed: linear regression and count regression models. The former is found to more accurately predict passenger occupancies than the latter, likely due to the additive nature of passengers entering and exiting buses. Quantile regression models were developed to describe uncertainty in the passenger occupancy estimates. These models are able to directly estimate any desired confidence interval for expected passenger occupancy and can be used to reveal characteristics that most impact the size of these confidence intervals.

These techniques are demonstrated using archived data from automatic vehicle location and automated passenger counter systems for a campus bus route in State College, PA. An exhaustive dataset of 12 months of data that included a total of over 500,000 unique observations were available for this effort. The dataset was appended with weather data from local and national sources—specifically, precipitation, snow depth and temperature information—as weather is known to impact bus operations and passenger demand.

Travel time models were estimated for a single stop-pair along the route. AFTs models were compared with linear regression models, which is a commonly used method for estimating bus travel times. The results suggest that the survival models were as accurate as (if not more accurate than) the linear regression models in terms of point estimates. However, the survival models better described the dataset, including the distribution of dependent variable and the uncertainty associated with the travel time estimates. These models revealed key insights about how expected travel times change across various time periods and how other variables impact travel time estimates. For example, the current passenger count on the bus was found to increase expected travel time, as this would increase the dwell time expected at intermediate stops due to longer unloading times. The survival model also revealed parameters that most influenced travel time uncertainty. This included the recorded travel time from the previous bus traveling through the segment and the number of passengers currently onboard the bus.

Three different modeling frameworks were considered to develop a single model to estimate passenger occupancies at any stop along the entire route. The results suggest that the “next-stop” model, which predicts passenger occupancies after the bus passes the next downstream stop, is most accurate for short-term passenger occupancy predictions within a few stops downstream of a bus’s current location. For longer-term predictions, the “segment-based” framework is found to be most accurate. The uncertainty analysis reveals that smaller bus headways are associated with more uncertainty in the passenger occupancy estimates. The presence of precipitation and lower temperatures increases passenger occupancy uncertainty, while snow reduces uncertainty. The estimated models also reveal which time periods and segments along the route have the most uncertainty in passenger occupancies. A summary of major findings is provided in tabular form below.

---

Outcome (predicted)	Conclusions from modeling activities
Travel time	<ul style="list-style-type: none"><li>• AFT survival models outperform linear regression models</li><li>• Uncertainty in travel time increases with mean travel time</li><li>• Travel time of previous bus and current onboard passenger occupancy associated with more uncertain travel times</li><li>• Weather related variables have little impact on travel time uncertainty</li></ul>
Passenger occupancy	<ul style="list-style-type: none"><li>• Linear regression models outperform count regression models</li><li>• “Next-stop” modeling framework most accurate for predictions 1-5 stops away</li><li>• “Segment-based” modeling framework most accurate for predictions &gt;5 stops away</li><li>• Quantile regression model accurately describes uncertainty associated with estimates</li><li>• Smaller bus headways found to have more uncertainty passenger occupancies than larger bus headways</li><li>• Precipitation and lower temperatures increase uncertainty, while snow reduced uncertainty in passenger occupancies</li></ul>

---

Overall, these methods show great promise in describing bus transit data, as well as predicting travel times and passenger counts. They are able to accurately predict bus travel times and passenger occupancies of individual vehicles and provide reliable indications of the uncertainties associated with these estimates. In an environment that is information-rich and in which transit users seek the most high-quality information about the current state of the transit network, providing these results to passengers might help to improve their decision-making and increase their confidence in the reliability of real-time transit information systems. These models can also benefit transit service providers. For example, models of passenger occupancy can predict when buses are expected to be full so that additional capacity can be provided in real time. Travel time uncertainty can also be used to optimize staffing decisions and plan driver shift changes.

---

## I. INTRODUCTION

### MOTIVATION

The need to maximize the number of passenger trips carried by public transportation modes increases as the complexity and magnitude of urban traffic congestion continues to grow. Many strategies have been proposed and implemented to entice commuters away from the private automobile mode and onto public transportation. These include providing priority for transit vehicles at individual intersections (Christofa and Skabardonis, 2011; Conrad et al., 1998; Hunt et al., 1982; Skabardonis, 2000; Xuan et al., 2012; Xuan et al., 2010), along specific links or routes (Eichler and Daganzo, 2006; Guler and Cassidy, 2012; Guler and Menendez, 2014; Viegas and Lu, 2001, 2004; Viegas et al., 2007), or across entire networks (Daganzo et al., 2012; Gonzales and Daganzo, 2012). Other strategies attempt to make public transportation more economically attractive by charging automobile use (e.g., congestion pricing) and providing transit subsidies (Anas, 1981; Ben-Akiva and Boccara, 1995; Glazer and Niskanen, 2000; Sherman, 1972).













Improving the reliability of a transit system is another way to increase ridership. Users of public transportation systems typically identify reliability—measured by the predictability of travel times and consistent availability of space for passengers to board—as a key measure of the quality of transit service and a major determinant of transit use (Bates et al., 2001; Brownstone and Small, 2005; Golob et al., 1972; Prashker, 1979). Improving the reliability of a transit system increases transit passengers' satisfaction and improves ridership, leading to reduced traffic congestion (which can further improve transit reliability) and decreased negative externalities like greenhouse gas and other harmful emissions. Improved reliability can also reduce the type of experiences that have been shown to elicit negative feelings in passengers toward transit systems, such as excessively long waiting/transfer times or not being able to board a vehicle when it arrives due to overcrowding (Carrel et al., 2013). Therefore, maintaining or improving reliability is vital to maintaining and increasing the competitiveness of transit systems.

Unfortunately, transit systems—and especially bus transit systems—are inherently unreliable. Research has shown that buses traveling along a route have a natural tendency towards bunching or pairing, making it difficult for bus transit agencies to maintain reliable arrival schedules (Newell, 1974; Newell and Potts, 1964). The mechanism that causes this instability is the passenger arrival and service process. As described in Newell and Potts (1964), the time that a bus spends serving passengers at a stop generally increases proportionately with the time between the current and preceding bus arrivals to that stop—the longer the time between two consecutive buses, the more people generally arrive to wait. For this reason, a bus arriving late to a stop would spend more time than expected serving passengers, causing it to arrive even later to the next stop as a result. The reverse is true for a bus arriving early to a stop. This positive feedback loop causes buses to eventually become paired together as an early bus will “catch up” to the late bus ahead of it. Several control strategies have been used to address this instability in practice, including: adding additional slack time to a schedule to absorb variability in travel times and passenger demands (Daganzo, 2009; Xuan et al., 2011), running buses at a speed below their maximum speed to allow late buses to catch up (Daganzo and Pilachowski, 2011), having late buses skip stops to catch up (Sun

and Hickman, 2005), and limiting the number of passengers that are allowed to board a bus at any stop (Delgado et al., 2009). Adding additional slack time and running buses at speeds below their maximum speed can be effective at reducing bunching. However, they have some drawbacks: 1) they are difficult to implement in practice; 2) they reduce overall commercial speeds; 3) they may still result in unreliable service, even if executed perfectly; and, 4) they are vulnerable to large service disruptions. Having late buses skip stops and limiting the number of passengers allowed to board at stops are also undesirable because they directly increase the occurrence of the types of negative experiences that lead to decreased transit ridership. Transit priority strategies can also reduce the variability of travel times to reduce the probability that a vehicle becomes early or late in the first place. Examples include transit signal priority (Christofa and Skabardonis, 2011; Smith et al., 2005), dedicated bus or queue jumping lanes (Guler and Cassidy, 2012; Nowlin and Fitzpatrick, 1997; Viegas and Lu, 2001, 2004; Viegas et al., 2007), and the presence of additional signals used for buses to skip queues at intersections (Guler and Menendez, 2014; Wu and Hounsell, 1998). However, these do not address the feedback loop that leads to bunching once a vehicle becomes early or late.

## Improving Perceived Reliability Through Information Provision

Instead of trying to fix the inherent instability directly, transit agencies also attempt to address transit system unreliability through the provision of high quality information to users about the current state of the system (e.g., the current location of buses and expected arrival times to each stop). This can improve the *perceived reliability* of the system, allowing users to make better decisions during the trip planning process. For example, the Swiss Federal Railways provide users with an indication of seating availability on a train based on historical information when purchasing a ticket (Figure 1). This allows passengers to account for seating availability in their decision-making process.

	Station/Stop	Time	Duration	Chg.	Travel with	Information
Connections for Mo, 22.06.15						
1	 Lausanne	dep 04:45	2:17	1	IR, IC	1.   2.  
	 Zürich HB	arr 07:02				
2	 Lausanne	dep 05:39	2:17	0	ICN	1.    2.  
	 Zürich HB	arr 07:56				
3	 Lausanne	dep 05:45	2:13	1	IR, IC	1.    2.  
	 Zürich HB	arr 07:58				

**Figure 1. Historical Passenger Occupancy Information from the Swiss Federal Railway**

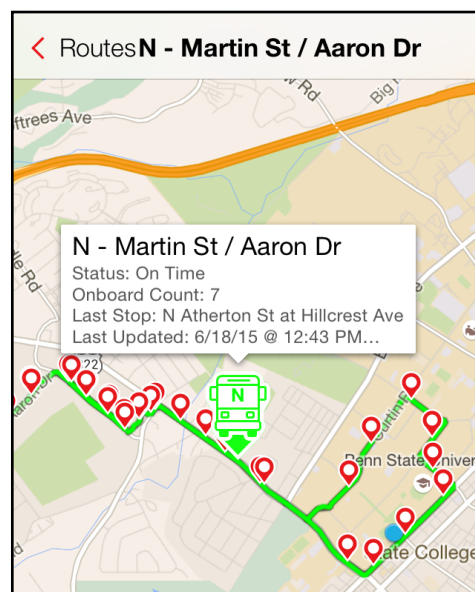
Black passenger icons indicate trains with historical seating availability; red passenger icons indicate trains with very low historical seating availability.

Of course, historical information alone may not be sufficient for transit systems that are highly unstable and dynamic, such as bus systems (as will be shown here). Instead, real-time data/information can be obtained by agencies and provided to passengers. Sources of real-time information include Automated Vehicle Location (AVL) systems for current bus locations and Automated Passenger Counter (APC) systems for bus passenger



occupancies. Some types of real-time information can also be crowd-sourced directly from transit passengers; e.g., the mobile app Tiramisu obtains seating availability from current passengers. This real-time information can then be provided directly to users as a snapshot describing current transit system operations.

Real-time transit information can be disseminated to passengers using a variety of methods. One way to provide this information is through real-time maps that show the current location of all buses and the current passenger occupancies of each. Figure 2 shows an example taken from the Centre Area Transportation Authority mobile application. Maps like these are useful for regular commuters or passengers with a high degree of familiarity with the transit system, as they can predict how the system will evolve in the near future. However, because they place the burden of prediction onto the users, they might be problematic for visitors or those not familiar with the transit system. These users are not likely to be able to make these predictions on their own, as they might not know which segments of the route serve the most passengers or are the most likely to become congested. If new transit users are unable to predict arrival times or future bus passenger occupancies, they are likely to encounter the types of negative experiences that preclude them from using transit again in the future. Additionally, these types of map-based systems cannot be integrated with real-time trip scheduling services to plan optimal routes or paths in a system using real-time information (Jariyasunant et al., 2011).

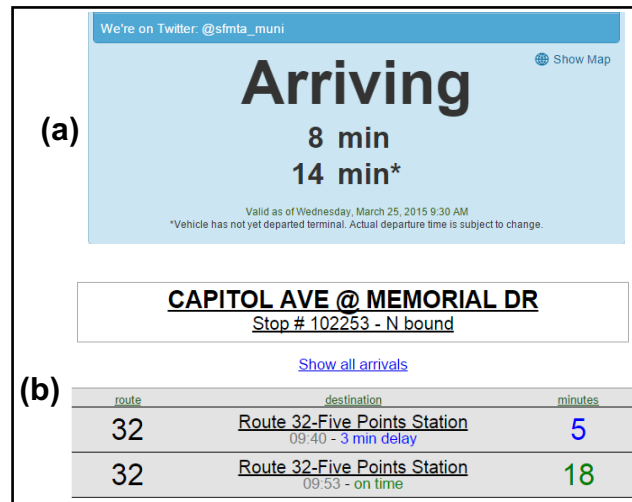


**Figure 2. Real-time Bus Location and Passenger Occupancies from CATA Application ([www.catabus.com](http://www.catabus.com))**

## Predictions of System Evolution

To avoid these drawbacks, real-time estimates of how the system will evolve—e.g., anticipated bus arrival times to each stop—can be provided directly to potential transit users. Doing so takes the burden of prediction away from the user and places it on the transit agency or real-time service provider, which is better-suited to make these predictions. Examples of independent services that provide travel time predictions include NextBus

([www.nextbus.org](http://www.nextbus.org)) and OneBusAway (<http://onebusaway.org>) (Figure 3a and Figure 3b). These estimates can also be used as input for trip scheduling services (such as Google Maps) to enable real-time route planning. Statistical models are developed to estimate bus arrival times that rely on AVL data and historical information of travel times along routes. Models of bus arrival times have been discussed at length in the literature, and various modeling approaches have been explored.



**Figure 3. (a) Real-time Bus Arrival Prediction through NextBus ([www.nextbus.com](http://www.nextbus.com)); and, (b) Real-time Bus Arrival Prediction through OneBusAway ([www.onebusaway.org](http://www.onebusaway.org))**

However, services such as Nextbus and OneBusAway provide only a single numerical point estimate of the time until the next bus arrival and do not include any indication of the potential error associated with that estimate. Single point estimates are potentially troublesome because they could provide users with a false sense of precision about when the bus will arrive—the user will expect the bus to arrive at the predicted time and will plan around that value. If a user experiences a long wait or misses a bus based on this estimate, the event could generate increased negative feelings toward transit, eventually leading to decreased transit use. Furthermore, any deviations from the predicted arrival times, which are likely to occur, could be viewed as an indication of unreliable or poor transit service and information. For example, Figure 3a shows that the next bus will arrive to the stop in eight minutes. In practice, there is some variability from this estimate when compared with the actual arrival time, and the amount of variability can significantly impact the expectations and behaviors of transit users. One can imagine that their trip planning would be different if the estimate were minutes compared with minutes. Furthermore, user satisfaction with the service would be different if it took the bus 12 minutes to arrive in the second case than it would in the first.

There are many factors that might introduce variability into bus arrival times and cause variability from these point estimates. Examples include traffic control at intersections, bus demands at intermediate stops, and interruptions from other modes—such as pedestrians at crosswalks or congestion from automobiles (Mazloumi et al., 2009). If these factors are non-existent or cause only very minor variations along a route, point estimates alone

might be adequate. However, if these factors are prevalent, the estimates could change substantially as the system evolves, and real arrivals could differ greatly from initial predictions. The magnitude of the variability in arrival times would also change significantly based on temporal factors, such as time of day, day of the week, and the locations of other buses. Another factor that might cause variability in expected bus arrival times is the number of passengers currently on the bus because these passengers may have to alight from the vehicle at intermediate stops, which could increase both the number of stops and dwell times. Few studies incorporate APC data into point estimates (Shalaby and Farhan, 2004), and none quantify how these factors simultaneously influence variability of arrival time estimates.

Additionally, while transit agencies tend to focus their efforts on providing predictions of bus arrival times, little to no attention is paid to predictions of bus passenger occupancies. APC data is sometimes provided along with vehicle locations on real-time maps, but this requires that users estimate on their own how full the bus will be at subsequent stops. In systems with one-way travel patterns, like the morning commute toward a central business district, this information might be sufficient for everyday commuters. However, different travel patterns would make it difficult for users to determine how full the bus will be at various points along the route. To the authors' knowledge, no modeling attempts have been made to estimate bus passenger occupancies (as well as their uncertainty) based on real-time transit information. These models could be integrated within larger trip planning models to incorporate the likelihood that a specific trip can be made feasibly and to inform passengers whether or not a bus is likely to be full. Providing this information to transit users can help reduce the occurrence of the negative experiences that are likely to cause negative feelings about transit and reduce ridership.

## **RESEARCH OBJECTIVES**

In light of the existing gaps in the current state of practice and the research literature, two research objectives were identified for this project. The first was to develop a modeling framework that can be used to simultaneously estimate bus travel times and the uncertainty associated with these estimates. Providing this information to transit users can help them make more informed decisions in the trip planning process. The second is to explore the feasibility of estimating passenger occupancies of individual buses (and the associated uncertainty of these estimates) using real-time information available from APC and AVL data. Such information can help inform transit passengers of potential issues with overcrowding that might lead to skipped stops or boarding limits at specific stops, which could help mitigate negative experiences associated with these events. It can also provide transit agencies with an indication of when and where additional bus capacity is required in situations where demand is highly variable.

As a proof of concept, these modeling frameworks are developed and applied using field data from the bus system in State College, PA. The benefits of using these models are assessed by comparing their outputs with estimates based on historical trends, and the results are very promising. The travel time models are also compared with traditional modeling approaches that provide only a point estimate of travel time to determine if the additional complexity of modeling the uncertainty is warranted. Through this case

study application, significant predictors of bus arrival time and bus passenger occupancy uncertainty are confirmed and unveiled.

## **ORGANIZATION**

The remainder of this report is organized as follows. Chapter II provides a review of the existing modeling approaches for bus travel times and bus transit demands (as a surrogate for passenger occupancies). Chapter III describes the data used in this project to develop the proposed models. Chapter IV provides a description of the statistical modeling approaches that are used as a part of this project. Chapter V describes the efforts to model bus travel times, while Chapter VI describes the passenger occupancy models. Finally, Chapter VII summarizes the major points and provides concluding remarks.

## II. REVIEW OF PREVIOUS MODELING METHODS

A review of the relevant literature was performed to identify existing efforts to estimate real-time bus travel times and passenger occupancies of individual buses. The remainder of this section describes these existing efforts, the modeling methods used, and their drawbacks.

### MODELS OF BUS TRAVEL TIMES

Various methods have been proposed to develop models of bus travel times in the research literature. In general, three major modeling tools have been used:

- Kalman-Filters, KF
- Regression models
- Artificial Neural Networks, ANN

Hybrid prediction models also exist that combine two of these approaches, such as ANN and KF. Table 1 provides an overall summary of the individual studies that were found in the literature, along with the methods and independent variables considered.

**Table 1. Summary of Bus Arrival Time Prediction Literature**

Modeling approach	Reference	Input	Notes
Kalman Filter	(Cathey and Dailey, 2003)	Real-time vehicle location	Gave a general prescription for the prediction of transit vehicle arrival/departure based on Kalman filter
	(Shalaby and Farhan, 2004)	Travel time and passenger arrival rate from historical data and previous bus	Calculated running time and dwell time separately by predicting the onboarding passenger numbers at stops and assuming a constant time for each of them
	(Vanajakshi et al., 2009)	Real-time spot speed and location	Used algorithm based on a model discretized over space to account for the heterogeneity in road conditions
Regression of Bus Travel Speeds	(Sun et al., 2007)	Spot speed from historical data and current bus, location of current bus	Raised an dynamic algorithm that calculate the mean travel speed to downstream segments according to the location of the vehicle
	(Chen et al., 2011)		Compared between two frameworks of modeling the travel segments and found the "section-based" models better

Modeling approach	Reference	Input	Notes
Regression of Bus Travel Times	(Alfa et al., 1988)	Time of day, number of bus stops, number of stop signs, number of traffic lights, and distance	Compared linear and non-linear models, taking bus stop number, stop sign number, traffic light number, and distance as independent variables; the linear model was found to be the most appreciate
	(Frechette and Khan, 1998)	Average flow, turning ratio, stop vehicle ratio, traffic light number, heavy vehicle ratio, and transit flow	Applied Bayesian regression to model unit travel time of automobiles, taking independent variables including average flow, turning ratio, stop vehicle ratio, traffic light number, heavy vehicle ratio, and transit flow
	(Patnaik et al., 2004)	Distance, total dwell time, number of stops, and time of day	Considered lots of factors and finally took distance, total dwell time, number of stops, and time of the day as independent variables
Artificial Neural Networks	(Chien et al., 2002)	Distance, traffic volume, speed, delay, queue time, passenger demands, and number of intersections	Tried a hybrid ANN model integrating link-based and stop-based ANN models and developed an adaptive algorithm that adjust the prediction results based on previous prediction errors
	(Chen et al., 2007)	Weather, dwell time, day of week, time of day, and trip pattern	Developed ANN models of different structures and found dwell time, time of day as the most important factors
Artificial Neural Networks with Kalman Filter	(Chen et al., 2004)	Weather, day of week, time of day, and segment	Developed a dynamic algorithm based on Kalman filter that combines the most recent bus-arrival information together with the estimated travel times generated by the ANN model
Historical Data, Regression, and Artificial Neural Networks	(Jeong and Rilett, 2005)	Historical: Mean travel time and mean dwell time Regression: Distance and schedule adherence ANN: Arrival time, dwell time, and schedule adherence at the current stop	ANN model was found to give best predictions; and an ANN model that predicts with previously calibrated parameters was found to be no worse than one calibrates the parameters in real-time

Note that the input variables have been found to be statistically significant in most modeling efforts.

## Kalman-Filter Models

Kalman-Filter (KF) models provide estimates of future events based on knowledge of historical information and the most recent real-time observations. Future estimates are made using a linear recursive update algorithm that essentially adjusts the historical average based on the more recently observed data using optimal weights assigned to each. More information on Kalman Filters can be found in (Haykin, 2004).

For bus travel times, two general approaches have been developed. The first uses Kalman Filters to estimate travel times directly using historical travel times through a segment and current (real-time) travel times provided by AVL systems. The second uses Kalman Filters to estimate bus travel speeds based on historical speeds and real-time speed estimates. The travel time is then calculated using the estimated bus travel speed and travel distance.

In this approach, any environmental impacts—such as weather or the presence of special events—are assumed to be accounted for by the most recent observed travel times and thus indirectly incorporated into the predictions of future travel times. For example, weather effects that might cause larger travel times are accounted for in the observed larger travel times of the most recent vehicle.

These models are fairly simple to estimate and apply in practice, which makes them advantageous for practical implementation. However, these models assume that travel times are fairly stable in time, such that the travel time of the previous bus traveling through a segment can indicate the travel time for the next bus traveling through the segment. This approach may not be valid if buses experience any instability that might cause bunching—an inherent problem in bus systems—as bunching can result in widely different travel times for consecutive buses. Furthermore, this approach provides only a point estimate of travel time and does not indicate the potential uncertainty that is associated with this estimate (i.e., the accuracy of the estimate is not simultaneously provided).

## **Linear Regression Models**

Regression models are also used to identify relationships between a specific variable of interest and various other independent variables. Regression models are useful because they can account for the impact of changing any specific independent variable while hold all other independent variables constant. They can reveal the relationship between individual variables and also joint relationships when one or more variables interact. More information on linear regression models can be found in (Neter et al., 1996).

Regression-based models of bus travel times generally use travel time as the dependent variable, although speed-based models exist that seek to predict bus travel speed along a segment. The most common type of regression models used are linear regression models, which assume an additive relationship between the independent variable and bus travel times or speeds. A range of independent variables have been considered in regression models of travel time. These include traffic variables (e.g., traffic flow, bus flow, proportion of heavy vehicles), route characteristics (e.g., travel distance, average dwell time, number of stops), and environmental variables (e.g., time of day, day of week, precipitation, snow depth, and temperature). Environmental variables are explicitly included in the model, although some regression efforts also include the travel time/speed of the previous bus to also help capture these impacts.

Regression models are also fairly simple to estimate and implement for bus travel time prediction, which means they can be readily used by transit agencies and real-time transit information providers. However, like the KF models, existing regression-based approaches provide only a single numerical value of the travel time estimate (i.e., a point estimate). The uncertainty associated with these estimates can be obtained from the linear model results (based on basic regression assumptions), but this uncertainty must be assumed to be constant for all travel time estimates; i.e., the accuracy associated with each travel time estimate does not change based on other factors.

---

## Artificial Neural Networks

Artificial neural networks are a powerful tool for modeling data where the relationships between dependent and independent variables are not entirely clear or do not follow a linear (or other basic) functional form. These models are trained using an extensive dataset, during which statistical relationships between dependent and independent variables are identified. More information on ANNs can be found in (Haykin, 2004).

In general, the same basic independent variables are used in ANNs as in regression models. However, the predictions tend to be more accurate in ANN models than other modeling approaches because ANNs can model very complicated relationships without assuming the functional relationships between predictor and dependent variables; e.g., one study found that ANN models of bus travel times were more accurate than linear regression models estimated from the same dataset (Jeong and Rilett, 2005).

However, the trained ANN models themselves are generally “black boxes” in that they do not reveal the nature of the relationships that are uncovered. ANNs also require more data to be estimated and are not likely to be as transferable. Thus, ANNs are not generally insightful and are used primarily for prediction purposes. While ANNs may be capable of providing an indication of estimate uncertainty, no efforts have been made in the research literature to use ANNs to examine bus travel time estimation uncertainty.

## MODELS OF BUS PASSENGER DEMAND

In comparison with bus travel time prediction models, little attention has been paid to the prediction of transit vehicle occupancies. A significant amount of literature exists on transit demand prediction models (Table 2). However, these demand models are macroscopic models that focus on network-level predictions and are used to design and evaluate transit networks. A few demand prediction models exist at the route- and stop-level, but these predict passenger flows (e.g., expected number of passengers per hour) and do not consider the occupancies of individual vehicles. The only relevant research work found uses Kalman Filter techniques to predict boarding numbers at individual stops (Shalaby and Farhan, 2004). These stop-level boarding numbers are used as a predictor to estimate dwell times at these stops. However, this model considers only the number of passengers boarding at specific stops; the numbers of passengers alighting at each stop—which are necessary to predict bus occupancies—are not considered. Thus, it appears that estimates of real-time passenger occupancies are lacking in the research literature.



**Table 2. Summary of Bus Demand Prediction Literature**

Reference	Level of study	Summary
(Paulley et al., 2006)	Network-level	Studied the influence of fares, quality of service, income, and car ownership on public transit demands to produce an up-to-date guidance manual for UK
(Ryan and Frank, 2009)		Found a small but significant, positive relationship between the walkability of the environment and transit ridership
(Lee et al., 2013)		Applied stops aggregations based on distance and stop names to analyze the relationship between the public transit demands at specific times of the day and the associated land uses that may strongly influence the timing of that demand
(García-Ferrer et al., 2004)	Route-level	Established a ridership forecasting model that considers calendar effects, changing supply service, changing seasonality, and outliers
(Kerkman et al., 2015)	Stop-level	Applied two cross-sectional multiple regression models and summarized the important factors affecting ridership: potential demand, transit supply, and match between transit supply and demand

### III. DATA SOURCES FOR STATISTICAL MODELING

This project examined novel methods to estimate bus travel times simultaneously with travel time uncertainty and bus passenger occupancies. The model dataset used in this project to create these models combines data from three sources. The primary source was the Centre Area Transportation Authority, which provided archived real-time data on transit vehicle operations for one of its busiest transit lines. These data were supplemented with weather data that were obtained from both the Pennsylvania State Climatologist and National Climatic Data Center. The remainder of this section will provide more details on these data, as well as a discussion of data cleaning and preparation.

#### TRANSIT DATA

The Centre Area Transportation Authority (CATA) is the primary transit service provider in Centre County, PA. CATA buses provide service within five local municipalities—the State College Borough and the four surrounding townships: Patton, Ferguson, Harris, and College. Buses in the CATA fleet are equipped with both Automatic Vehicle Location (AVL) and Automatic Passenger Counter (APC) systems, which provide real-time information as the bus travels along its routes. These data are collected and archived by CATA to assess operational performance and the quality of the transit service provided.

During the course of this project, CATA provided the research team with AVL and APC data for the Blue Loop, a 4.1-mile (6.6-km) clockwise route with 15 stops serving the Pennsylvania State University (PSU) main campus. The Blue Loop is one of the busiest bus routes in State College, PA. It serves most areas inside PSU, connecting residence halls, instruction halls, bus transit center, student commuter parking lot, and downtown State College (Figure 4). Because of the high travel demands within the campus, two to four buses serve the relatively short route during daylight hours. The buses operate on a headway-based scheme with scheduled headways between 5-12 minutes during this time. Stop 6 (Jordan East Parking) serves as a headway checkpoint where buses are held to maintain consistent headways.

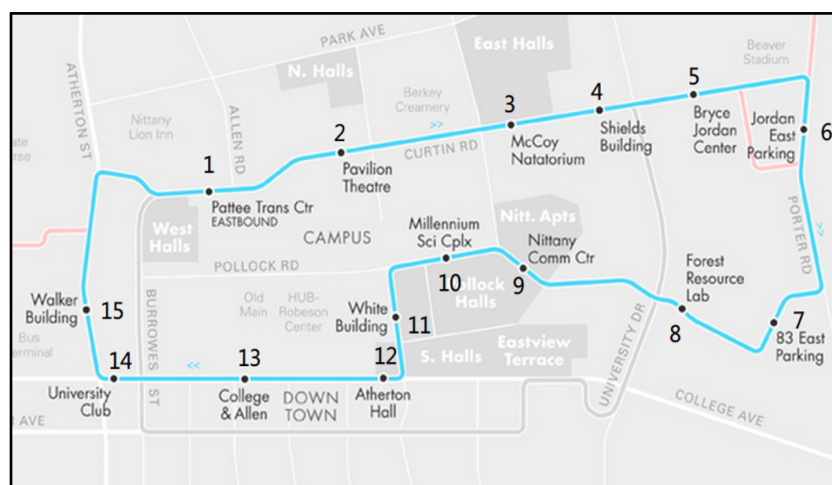


Figure 4. Route Illustration for the Blue Loop (source: [www.catabus.com](http://www.catabus.com))

The Blue Loop (BL) is a very busy route and was specifically selected from all of the available CATA routes. It served as an ideal case study to model travel time and passenger occupancy uncertainty due to various factors that lead to unpredictable operations. For one, the high-frequency and campus setting results in large demand—and thus, dwell time—variations at each stop. This is due primarily to demand being driven by the university class schedule, which causes significant demand variations throughout the day at highly frequent intervals. The route also has a mixture of intersection control strategies—stop signs and traffic signals—that lead to unpredictable waiting times at intersections. Lastly, the high pedestrian activity and large number of pedestrian crossings result in buses potentially experiencing significant delays during their route at several locations. Thus, the BL provides a worst-case scenario that should provide poor overall predictive ability compared with other routes. The methods developed in this paper will likely provide more precise predictions when applied to non-campus routes with less variable demand patterns, more consistent intersection control, and fewer bus-pedestrian interactions.

Data along the route were collected as follows. All 15 stops along the route are defined using geocoded “stop-zones” around the bus stop area. Buses are identified and generate a data report as they enter the stop-zone (a “pull-in” maneuver) and exit the stop-zone (a “pull-out” maneuver). A separate “stop report” is created for each pull-in and pull-out pair that occurs at a particular stop at a particular time. This stop report contains the following information:

- Vehicle ID: unique identifier for each vehicle along the route
- Stop ID: identifier of the stop that the report was received from
- Time: time the vehicle reported pulling out of the stop zone
- Dwell time: the difference between the pull-out and pull-in times; i.e., the time the vehicle spent within the stop zone [seconds or sec]
- Scheduled headway: the headway vehicles seek to maintain [sec]
- Headway deviation: difference between the actual headway and scheduled headway [sec]
- Onboard occupancy: number of passengers onboard when the bus leaves the current stop [passengers or pax]
- Boarding count: number of passengers boarding the vehicle at the current stop [pax]
- Alighting count: number of vehicles alighting the vehicle at the current stop [pax]

A total of 537,192 unique stop reports were obtained from January 2013 to April 2014. Observations from May, June, July, and December were excluded, as these represented months with atypical schedules due to the university not being in session.

## WEATHER DATA

Weather information was appended to the archived transit data because weather has been found to have a statistically significant impact on transit operations and bus ridership in previous research efforts (Table 1). Temperature [°F] and precipitation [mm] data were available from Pennsylvania State Climatologist on an hourly basis (Pennsylvania State Climatologist, 2015). This information was appended to the transit data based on the time variable described above. Snowfall [mm] and snow depth on ground [mm] data were available from National Climatic Data Center each day and were appended based on the date of each data observation (National Oceanic and Atmospheric Administration, 2015).

## TIME OF DAY SPLIT

The Blue Loop runs from 4:45 a.m. to 12:30 a.m. on weekdays and from 9:00 a.m. to 12:30 a.m. on weekends. Only weekday data from 8:00 a.m. to 9:00 p.m. were used for the analyses performed as a part of this report. The selected time span covered the busiest periods for Blue Loop service when the demand is highest and the system is most unstable. Weekend data were excluded because travel demand patterns are significantly different on weekends. Most weekends experience significantly lower travel demands; however, a subset of weekends experience large demand peaks and service changes due to sporting events or other special events at the Penn State University. Although the Blue Loop runs nearly continuously, nighttime hours were excluded due to the very low travel demands.

As mentioned previously, buses on the Blue Loop experience significant travel time and demand fluctuations regularly during the weekday hours due to class schedules. A set of dummy variables named `Time_Period_x` was created to indicate the combination of day in the week and time in the day based on visual checks of fluctuations in onboard passenger number and travel time respectively. Here, the `x` represents a specific time period as defined in Table 3. These checks indicated that the fluctuation patterns of travel times and passenger occupancies followed a repeatable pattern that was in line with the class schedule, which are the same on Monday-Wednesday-Friday and on Tuesday-Thursday. On Mondays, Wednesdays, and Fridays, 50-minute classes are separated by 15-minute breaks all day long. On Tuesdays and Thursdays, most classes are 75 minutes long and are separated by 15-minute breaks, while a few 50-minute classes still exist. Therefore, this time period was created using only knowledge of Penn State University class schedules, which is the largest driver of bus travel demands for this campus bus route. The new variable set contained 27 categories: every class slot was regarded as a low-demand period, while intervals between classes represented high-demand periods. Time periods on weekdays with the same class schedules were combined. Table 3 shows how the `Time_Period` was divided.

**Table 3. Definition of Time\_Period Variable**

Time_Period	Weekday	Starts	Ends
1	MWF	8:00	9:05
2		9:05	9:55
3		9:55	10:10
4		10:10	11:00
5		11:00	11:15
6		11:15	12:05
7		12:05	12:20
8		12:20	13:10
9		13:10	13:25
10		13:25	14:15
11		14:15	14:30
12		14:30	15:20
13		15:20	15:35
14		15:35	16:25
15		16:25	19:00
16	TR	8:00	8:30
17		8:30	9:15
18		9:15	9:45
19		9:45	11:00
20		11:00	11:15
21		11:15	12:30
22		12:30	13:00
23		13:00	14:15
24		14:15	14:30
25		14:30	15:45
26		15:45	16:15
27		16:15	19:00

MWF represents Monday, Wednesday Friday; TR represents Tuesday and Thursday.

## DATA CLEANING

A total of 345,153 unique stop report observations from the time periods of interest were available for modeling. These data were then examined and filtered to eliminate any erroneous or potentially inaccurate information, as described below.

### Loss of Information

Several stop report observations had unrealistic values of scheduled headway (e.g., a scheduled headway of 0) or dwell time at a stop (e.g., a negative dwell time). Based on conversations with CATA staff and the authors' knowledge of the dataset and data collection process, it was determined that these erroneous values most likely occurred due to loss of communications between the bus radio system and the CATA headquarters where data are received and stored. These errors accounted for 0.6% of the data, and these observations were removed from the dataset.

## Headway and Headway Deviation Issues

In general, buses on the Blue Loop maintain scheduled headways of 300 or 600 seconds (5 minutes or 10 minutes, respectively). When transitioning between these headways, often an intermediate scheduled headway of 480 seconds (8 minutes) is used. When this scheduled headway is used, deviations from scheduled headway often become very large. This occurs because actual headways only change gradually while the transition in scheduled headway occurs abruptly. Since the deviation from scheduled headway is the difference between actual and scheduled headway, the transition period does not accurately reflect actual schedule deviations. For this reason, all observations with a 480-second headway, which represented 0.9% of all observations, were removed from the dataset.

Similarly, some observations had headways deviations that were deemed unrealistic. This included all deviations from scheduled headway that were greater than 1200 seconds or less than -600 seconds. The larger positive values indicated a bus that was more than 20 minutes late compared to its scheduled headway, and the large negative values indicated a bus that was more than 10 minutes earlier than its scheduled headway. These values were considered unreasonable outliers for a bus routes that had an average cycle time of just 20 minutes. About 0.8% of the observations had these unrealistic values and were removed from the dataset.

## Operation Interruption

Interruptions in bus operations along the route occurred for various reasons. These included: driver breaks, driver changes, mechanical breakdowns, and emergencies. During an interruption, bus operations may lose continuity if the bus stops for a long period. In such a case, passengers are likely to leave the vehicle and instead travel on foot. Thus, travel times would not be reliable, and information on operations before the break would not necessarily be useful for predictions of occupancy after the break. The data were scanned to identify any running times between consecutive stops that exceeded 10 minutes. Once this limit was exceeded, it was assumed a long break in operations occurred, and these observations were removed from the dataset. As a result, 0.4% of the observations were removed.

## Unreasonably High Passenger Occupancies

The most significant data cleaning issue concerned the presence of unrealistically high onboard passenger occupancies. Approximately 5.9% of the data had onboard passenger counts larger than 80, with 25% of these having a value of 255 (the maximum value possible within this data field). These are not realistic considering that CATA considers 80 passengers to be the maximum realistic occupancy of buses traveling on the Blue Loop. Conversations with CATA staff revealed that these unrealistic values are the result of malfunctioning APC systems. Specifically, the APC systems that count alighting passengers often malfunction and underestimate the number of people alighting from the bus. This results in overestimates of onboard passenger occupancies because bus occupancies are determined by counts of passengers boarding at each stop and subtracting counts of alighting passengers. These errors also accumulate throughout the day and become

larger the longer the bus is in service. These errors are eliminated only when the bus returns to the depot and the APC systems are reset.

No reliable methods exist to differentiate accurate onboard counts from problematic counts within any given day; instead, it is known only that counts larger than 80 are likely the result of a malfunctioning APC unit. To overcome this, all buses were identified that had onboard passenger counts greater than the maximum reasonable value (80) at least once throughout its time in service. Because these represented known instances of erroneous APC readings, all observations associated with the same vehicle ID for the same day were then removed, given that all of those counts were subject to potential errors. In total, a large fraction (about 26%) of the observations were removed due to unreasonable passenger counts. Examination of this data revealed that data removed in this way were more likely to come from a small subset of buses (identified by their unique bus ID), which confirmed that these were caused by the malfunctioning APC units.

### **Summary Statistics**

After the data cleaning process, a total of 230,222 observations remained from the original 345,153 observations available. Although a large fraction of data were removed, most of this was due to the malfunctioning APC systems. The removal of this data is not likely to introduce any systematic bias in the data. Table 4 provides basic summary statistics for the pertinent variables for datasets both pre-cleaning and post-cleaning. The cleaning process helped to reduce the range of values observed for the variables, which is evidenced by the smaller difference between maximum and minimum values and the smaller standard deviations. It should be noted that this reduction was purely coincidental for the weather-related variables because their quality was not considered in the data cleaning process. Eighty percent of this data was used for the model fitting/estimation process, while the remaining twenty percent was reserved for validation of the model results.

**Table 4. Summary Statistics for Blue Loop Data Used in Model Development Before and After Data Cleaning**

Variable	Before cleaning					After Cleaning				
	N	Mean	Std. Dev.	Min.	Max.	N	Mean	Std. Dev.	Min.	Max.
Headway_Deviation		129.35	556.97	-39366	2681		136.86	241.32	-600	1200
Onboard		27.20	40.97	0	255		14.20	13.81	0	80
Temperature		41.67	18.22	-9.4	91.4		48.91	18.89	6.1	90.0
Precipitation		5.06	18.99	0	401		6.61	16.55	0	401
SnowDepth		25.26	61.47	0	381		25.36	63.69	0	381
Variable		Percentage with value of 1					Percentage with value of 1			
Scheduled_Headway 0	345,153		0.01			230,222		-		
Scheduled_Headway 300_seconds			0.59					0.59		
Scheduled_Headway 360_seoncds			0.27					0.27		
Scheduled_Headway 480_seconds			0.01					-		
Scheduled_Headway 600_seconds			0.11					0.12		
Scheduled_Headway 1200			0.01					0.02		



## IV. STATISTICAL MODELING METHODS

Various statistical modeling techniques were employed in this study to develop models for bus travel time and bus passenger occupancy. These different techniques facilitated the modeling of expected outcomes (which is traditionally done for transit data) as well as the distribution and uncertainty associated with these outcomes. The remainder of this section will describe these methods and provide some background on their advantages and limitations. For readers who are not interested in some of these technical details (or for those who are familiar with these models), this section may be skipped without loss of continuity.

### LINEAR REGRESSION MODEL

The linear regression model is one of the most basic—and most prominently used—statistical regression techniques. As discussed in the introductory section, linear regression has been previously applied when modeling bus travel times and for predicting transit demands, although not at the granularity of individual bus occupancies.

Linear models adopt the following form:

$$y_i = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon_i,$$

in which  $y$  is the dependent variable to be predicted,  $i$  is an index of the observation number,  $\bar{x}_i$  is a vector of  $J$  independent explanatory variables used to predict  $y$ , and  $\beta_j$  are the set of model coefficients to be estimated. The last term,  $\varepsilon_i$ , is an error term associated with each observation.

In a linear model, the effects of all variables are assumed to be additive (Rosenbaum, 2002). That is, the change in any independent variable  $x_j$  of one unit corresponds with an additive change in the prediction of the dependent variable by an amount equal to the associated model coefficient,  $\beta_j$ . These model coefficients are typically estimated using the ordinary least squares procedure, which selects the coefficients that minimized errors between actual observations and those predicted by the model.

Error terms,  $\varepsilon_i$ , in the linear regression model are assumed to be independently and identically distributed random observations with zero mean and constant variance (Gaussian distributed). Thus, the variance of these error terms is assumed to be independent of the dependent variables (i.e., uncorrelated). Violation of the constant variance assumption, commonly known as heteroscedasticity, results in biased estimates of the standard errors associated with each of the model coefficients. In practice, these violations are not unexpected, and this limits the predictability of the outcome's uncertainty using the results obtained from linear regression models. Specific to the work in this report, the linear models start with the assumption that the uncertainty associated with the prediction is the same for all predictions made (i.e., constant variance). These models can provide the expected value (point estimate) of the independent variable for a given set of independent variables:

$$E(y_i|\bar{x}_i) = \beta_0 + \sum_j \beta_j x_{ij}.$$

The uncertainty associated with this estimation is assumed to be constant and independent of the explanatory variables. This precludes the simultaneous estimation of the expected value and level of uncertainty for a given set of independent variables. To overcome this limitation, accelerated failure time survival models and quantile regression models will be used.

## NEGATIVE BINOMIAL REGRESSION MODEL

The negative binomial regression model is a type of count regression model that is used to model dependent variables with values restricted to non-negative integers. This type of count regression model has been used extensively in the modeling of transportation data, particularly crash data (Lord and Mannering, 2010; Lord et al., 2005; Poch and Mannering, 1996; Shankar et al., 1995).

To ensure the count restriction is maintained, count models assume that independent variables have multiplicative effects; i.e., the change in a single independent variable is associated with a multiplicative change in the estimate of the dependent variable (Hilbe, 2011). The specific functional form adopted in the negative binomial model is:

$$y_i = e^{\beta_0 + \sum_j \beta_j x_{ij} + \varepsilon_i},$$

where the exponent of the error terms are gamma distributed with mean  $\exp(\varepsilon_i) = 1$  and scale parameter  $\frac{1}{\alpha}$ . Using this error term, the negative binomial model assumes observed data can be described using the following density function:

$$P(y = y_i | \bar{x}_i) = \frac{\Gamma(y_i + \alpha^{-1})}{y_i! \Gamma(\alpha^{-1})} \left( \frac{\alpha^{-1}}{\alpha^{-1} + \mu_i} \right)^{\alpha^{-1}} \left( \frac{\mu_i}{\alpha^{-1} + \mu_i} \right)^{y_i}, y_i = 0, 1, 2 \dots$$

where  $\Gamma(x)$  represents the gamma function, and  $\mu_i$  is the mean value that describes the  $E(y_i)$ . The parameter coefficients are generally estimated using the maximum likelihood technique that seeks to select the parameters that maximize the likelihood that this distribution describes the observed count data.

Often, especially when using transportation data such as crash frequencies or passenger occupancies, a large number of zero values is observed. For example, many roadway segments experience zero crashes. With respect to transit, many buses run empty (i.e., zero passengers) during off-peak times. In this case, a zero-inflated negative binomial model (ZINB) can be used to account for these large numbers of zero observations. The ZINB uses a separate model to account for excess observations of zero counts (compared with the number that would likely be observed using the negative binomial distribution). A binary logit model is used to determine the likelihood of excess zero-count observations. Different independent variables can be used in this model than in the count prediction model. In essence, this increases the probability of observing a value of zero. Using  $F_i$  as the probability of the observation have a zero value from the binary logit model, the probability distribution function of the ZINB becomes:

$$P(y = 0|\bar{x}_i) = F_i + (1 - F_i)(1 + \alpha\mu_i)^{-\alpha^{-1}} \text{ and}$$

$$P(y_i = y_i|\bar{x}_i) = (1 - F_i) \frac{\Gamma(y_i + \alpha^{-1})}{y_i! \Gamma(\alpha^{-1})} \left( \frac{\alpha^{-1}}{\alpha^{-1} + \mu_i} \right)^{\alpha^{-1}} \left( \frac{\mu_i}{\alpha^{-1} + \mu_i} \right)^{y_i}$$

The predicted values using the ZINB becomes:

$$E(y_i|\bar{x}_i) = (1 - F_i) \times e^{\beta_0 + \sum_j \beta_j x_{ij}}$$

To test for zero-inflation and determine if the ZINB should be used, a Vuong test is used (Vuong, 1989; Washington et al., 2010). The Vuong test is essentially a t-test between two models that helps to determine which model is preferred. The Vuong test statistic is calculated as:

$$V = \frac{\bar{M}\sqrt{N}}{S_m}$$

where  $\bar{M}$  = the mean value of  $M$ ,  $M = \log\left(\frac{P(ZINB)}{P(NB)}\right)$ ,  $P(ZINB)$  = the probability of observing the outcome based on the ZINB model,  $P(NB)$  = the probability of observing the outcome based on the standard negative binomial model,  $S_m$  = the standard deviation of  $M$ , and  $N$  = the sample size. A test statistic value of larger than 1.96 indicates that the ZINB model is preferred over the standard negative binomial, a value of smaller than -1.96 indicates that the standard negative binomial is preferred, and a value between -1.96 and 1.96 is inconclusive. If the Vuong test indicates that the ZINB model is preferred, this provides justification for using the ZINB model in predicting the passenger occupancy of buses.

## ACCELERATED FAILURE TIME SURVIVAL MODEL

Survival models, also known as duration models, describe the time until a specific event occurs. These models have seen extensive use in the modeling of infrastructure deterioration, where the time-to-failure or required maintenance activity for a specific infrastructure element is considered as the dependent variable. However, as described in the review of relevant literature, survival models have yet to be used to describe transit data, even though the time required to travel between two stops fits within this general modeling framework.

An accelerated failure time survival model can be described using a survival function,  $S(t)$ , which provides the probability that the time until the event occurs exceeds a certain value,  $t$ :

$$S(t|\bar{x}_i) = Pr\{y_i > t|\bar{x}_i\} = Pr\left\{\beta_0 + \sum_j \beta_j x_{ij} + \varepsilon_i > \log t \mid \bar{x}_i\right\}$$

The accelerated failure time survival model is useful because the survival function provides the entire distribution of the dependent variable as a function of the independent variables  $\bar{x}$ . Obtaining such a distribution is not unique to accelerated failure time survival models—

the distribution of the dependent variable can be also obtained using linear and negative binomial models. However, the distributions obtained from the linear models change in the expected values only as the variance of the estimates is held constant.

Various distributions of the disturbance terms  $\varepsilon_i$  can be assumed, and each assumption provides a different form of the survival function and different distribution of the dependent variable  $y$ . Common error-term distributions include the extreme value, generalized extreme value, normal, and logistic distributions, which lead to the Weibull, generalized gamma, log-normal and log-logistic distributions for the dependent variable, respectively. These not only differ based on the functional form of the distribution assumed but also on the general behavior expected of the dependent variable. These can be described by the ratio of the probability density function to the survival function for the dependent variable, which is known more commonly as the hazard function. The Weibull distribution results in a monotonic hazard function, while the log-normal and log-logistic distributions result in a concave hazard function. The generalized Gamma distribution has the most flexible shape that can mimic the previous two hazard shapes or even create a convex U-shaped hazard function (Greene, 2011).

The model coefficients can be estimated in one of two ways. If data are censored (i.e., the time of some events is unknown), maximum likelihood estimation is applied. However, if the data are not censored, the model coefficient can be estimated using ordinary least squares.

## QUANTILE REGRESSION MODELS

Quantile regression is a modeling approach that predicts the  $\theta^{th}$  percentile of the dependent variable  $y$ , as opposed to the expected (mean) value (Koenker and Bassett Jr, 1978; Koenker and Hallock, 2001). This is useful when extreme outliers are likely to bias models of expected value or when information other than the mean value is needed. More pertinent to this project, quantile models can provide an indication of the range of values expected for a given set of conditions (subject to a certain degree of confidence defined by the quantiles chosen in the model). The size of this range can reveal how uncertain the estimate might be.

Linear quantile models have additive effects, like linear regression models, where the change in an independent variable results in an additive change in the estimate of the quantile value. The linear quantile models are estimated differently, however. Model coefficients are estimated by considering a weighted average of positive and negative error terms, as shown here:

$$\min \left( \sum_{i \in \{y_i \geq \beta_0 + \sum_j \beta_j x_{ij}\}} \theta \left| y_i - \beta_0 - \sum_j \beta_j x_{ij} \right| + \sum_{i \in \{y_i < \beta_0 + \sum_j \beta_j x_{ij}\}} (1 - \theta) \left| y_i - \beta_0 - \sum_j \beta_j x_{ij} \right| \right)$$

As shown, the estimate of the dependent variable is made using a linear model. The first term in the parentheses represents observations that would be underestimated by this linear model, and these errors are weighted by the percentile being considered. The second term in the parentheses represents observations that would be overestimated

by the linear model, and these errors are weighted by one, minus the percentile being considered. The coefficients  $\beta$  are selected that minimize this weighted average of the positive and negative error terms using maximum likelihood estimation techniques. The resulting model estimates the  $\theta^{th}$  percentile of the dependent variable as a function of the independent variable. There are no restrictions on the set of independent variables from one percentile to another; therefore, each percentile can have its own set of independent variables, and the functional form of the model for each percentile can be different.

## V. MODELING TRAVEL TIME UNCERTAINTY

In this chapter, a framework is developed to simultaneously model expected bus travel time between two stops and the uncertainty associated with this estimate. The results are compared with a more traditional method of modeling travel times that does not provide reasonable uncertainty values. The proposed method uses the accelerated failure time survival model (AFT survival model). The dependent variable in an AFT survival model is the time remaining until some event occurs; in this case, the time until the bus arrives at a particular stop. As previously mentioned, this modeling framework provides the full distribution of the dependent variable, which can be used to estimate the expected value (mean travel time) and variation (variance of travel times). This is compared with a linear regression model, which was found to be one of the commonly used modeling methods for bus travel times based on a review of the literature. The linear regression model provides only an estimate of the expected travel time. The variance of this expected travel time can be inferred from the linear regression model as the root mean square error (RMSE) of the model predictions. However, this variance is assumed to be the same for all estimates of mean travel time and thus might not be useful to estimate travel time uncertainty in real time.

This framework is applied to a single stop-pair available along the Blue Loop dataset to demonstrate its feasibility. Ideally, individual models would be needed for every stop-pair to estimate bus travel times between any two locations. However, errors in this specific dataset precluded this. The primary error involved how travel times were calculated along the Blue Loop. To detect the time a bus arrives at a stop, all 15 stops along the route are defined using geocoded “stop-zones” around the bus stop area. The time buses pull into a stop-zone and pull out of a stop-zone are noted by the AVL system and recorded. A “stop report” is created for each pull-in and pull-out maneuver at a particular stop. This report provides the pull-in time and the dwell time, which is equal to the difference between the pull-out and pull-in times. In many cases, stop-zones actually overlap, particularly when adjacent stops are in close proximity (e.g., stops 3-5, 14-15; see Figure 4). Thus, a bus can be reported as having pulled-in to one stop before it technically pulled-out of another. This yields unrealistic negative travel times between adjacent stop pairs along the BL that were not seen as reasonable to develop travel time models.

Since the goal was not to provide actual travel time models but just to assess the usefulness of this newly proposed modeling approach, a model was developed to estimate travel times between a single stop-pair: from stop 9 (Nittany Community Center) to stop 15 (Walker Building). This particular 1.2-mile (1.9-km) segment was selected because it provides a mix of traffic control devices along the route (stop signs and traffic signals), numerous pedestrian crosswalks, and travel along a busy signalized arterial (College Ave). This segment also avoids the stop 6 (Jordan East Parking), which serves as a headway checkpoint that may result in significantly long dwell times. Thus, only observations at stop 9 were used, and each bus’s travel time from stop 9 to stop 15 was considered as the dependent variable for analysis. A total of 15,421 observations were available in the cleaned modeling database for this purpose. Table 5 provides the summary statistics for the relevant variables that were considered for modeling purposes. Note that the Travel\_Time values are from Stop 9 to Stop 15 and are provided in seconds. Also, the weather-related variable “snowdepth” is coded as an indicator variable to represent the presence of

snow on the ground. When the actual snowdepth is larger than 30, the snowdepth variable is given a value 1; otherwise, it is 0.

**Table 5. Summary Statistics for Data Used in Travel Time Models Representing Observed Travel Times from Stop 9 to Stop 15 (N=15,421)**

Variable	After Cleaning			
	Mean	Std. Dev.	Min.	Max.
Travel_Time	590.21	131.32	367	1782
TTPreBus	589.63	131.39	367	1782
Headway_Deviation	139.67	207.45	-590	1197
Onboard	11.84	11.37	0	77
Temperature	41.65	18.25	-5.8	91.4
Variable	Fraction of observations with value of 1			
Scheduled_Headway 300		0.60		
Scheduled_Headway 360		0.27		
Scheduled_Headway 600		0.12		
Scheduled_Headway 1200		0.01		
Time_Period 1		0.04		
Time_Period 2		0.04		
Time_Period 3		0.01		
Time_Period 4		0.05		
Time_Period 5		0.01		
Time_Period 6		0.05		
Time_Period 7		0.01		
Time_Period 8		0.05		
Time_Period 9		0.01		
Time_Period 10		0.05		
Time_Period 11		0.01		
Time_Period 12		0.04		
Time_Period 13		0.01		
Time_Period 14		0.04		
Time_Period 15		0.18		
Time_Period 16		0.01		
Time_Period 17		0.03		
Time_Period 18		0.03		
Time_Period 19		0.02		
Time_Period 20		0.03		
Time_Period 21		0.04		
Time_Period 22		0.04		
Time_Period 23		0.02		
Time_Period 24		0.03		
Time_Period 25		0.03		
Time_Period 26		0.02		
Time_Period 27		0.12		
SnowDepth 0		0.78		
SnowDepth 1		0.22		

Table 5 shows that the expected travel time in this segment is 590 seconds, with a standard deviation of 131 seconds for individual values. This average historical value can provide a basic travel time estimate. However, it is clear that these travel times might change throughout the course of the day. Instead, historical information could be used to estimate the average travel time for each of the Time\_Periods considered and these then used as historical travel time estimates. Table A1 (in the Appendix) provides these historical travel time estimates for each Time\_Period. These historical values will be used as a baseline with which to compare the regression models developed as a part of this work.

## REGRESSION RESULTS

A linear regression model and several forms of the AFT survival models (each differing in the specific form of the disturbance term assumed) were each estimated to predict bus travel times along this segment. At the first stage, simple models with only primary effects (i.e., no interaction terms between variables) were estimated. Non-parametric survival models were applied first to improve understanding of the data. With non-censored data, a sharp increase in the hazard was observed when travel time is high, which might be caused by the existence of some extreme values. These extreme values for travel times likely occurred due to malfunctioning AVL units or during special events (e.g., Penn State University activities—such as sporting events—which cause significantly different travel patterns). To address the problem, an upper limit for travel time was set, and the top 1% percentile observations were censored such that a better-shaped hazard was obtained. The resulting AFT models had a much better fit, as the parametric models considered could accurately describe the hazard function. Summary statistics for these primary models are provided in Table 6.  $R^2$  reflects the overall fitness of a linear model; AIC (Akaike Information Criterion) is a measure of the relative model quality and provides a means to select between models; and RMSE (Rooted Mean Square Errors) gives an indication of difference between estimated values and actual values.

**Table 6. Summary Measures for Travel Time Regression Models without Interactions**

Form	Error Term Distribution	Independent variables	$R^2$	AIC	RMSE (seconds)
Linear	Normal	Time_Period	0.271	97715	111
AFT Survival	Weibull	Onboard	-	116407	111
	Log-normal	TTPreBus	-	112991	111
	Log-logistic	Headway_deviation	-	112858	111
	Generalized Gamma	Scheduled_Headway_300	-	112858	111
		SnowDepth Temperature	-	112625	113

In both the linear and AFT survival models, Time\_Period, peak scheduled headway (of 300 seconds), headway deviation, current occupancy, and travel time of the previous bus through the segment were found to be statistically significant, and their parameter estimates were all consistent with expectations. SnowDepth and Temperature were not statistically significant at 95% confidence level, but these variables were kept in the model because they provided a better fit (demonstrated by the lower AIC and RMSE values).



The RMSE was used as the primary means to compare each model's predictive ability. In terms of RMSE, the performances of all the models are close (even across different forms and distributions). This is likely due to the shape of the hazard curve not fitting any particular distribution better than the others. For each prediction, there is a mean error of about 110 seconds, which is approximately one-fifth of the mean travel time. For comparison, the RMSE value that would be obtained using the historical travel times is 126.80 seconds. Therefore, the regression models improve travel time estimation accuracy by an average of 12.5%. Note that  $R^2$  values are generally not provided for the survival models.

Estimates for the variable coefficients in all model forms have consistent signs, although the magnitudes of these estimates vary between modeling forms (as expected). The coefficients of the linear model are provided in Table A2 as an illustrative example because the coefficients in this model time have a physical meaning (the change in travel time based on a unit change in the associated variable). Notice that regular up-and-down fluctuations were observed from the estimates of the *Time\_Period* variable, which indicated that the peak time periods between classes generally have higher travel times. As expected, the *Onboard* coefficients had positive estimates because it is positively related with higher dwell times as well as busier time periods. *TTPreBus*, which represents the travel time of the previous bus in this segment, had a positive coefficient, which suggests that travel time will increase as the travel time of the previous bus increased. This is consistent with engineering intuition, as larger previous bus travel times indicate congestion and other effects that might slow down subsequent buses. *Headway\_deviation* had a positive estimate, which indicated that buses behind schedule have larger travel times. This is consistent with Newell's research findings about the inherent instability in bus systems. The coefficient estimate for the dummy variable *Scheduled\_Headway* (300 seconds) was positive and statistically significant because short headways are generally applied during busy time periods when both running times and dwell times increase. All other dummy variables for scheduled headway were not statistically significant, indicating a lack of significant differences for other scheduled headway values. This could also indicate that headway effects are picked up in the *Time\_Period* variable (which is likely). Also, not surprisingly, *SnowDepth* was positively related with travel time, which verifies that buses travel slower when the roads have snow. *Temperature* also had positive estimates, which indicates longer travel times on warmer days. This is likely due to larger delays at pedestrian crosswalks due to increased pedestrian activity on warmer days.

Models with secondary effects (i.e., that include interactions between variables) were also estimated to see if they provided additional estimation accuracy. As shown in Table 7, the inclusion of interactions between the time period and onboard passenger occupancy provided increased predictive ability over the models without variable interactions as measured by the lower AIC and RMSE values. These models provide more accurate predictions in terms of both the expected value and the uncertainty of travel time estimates. The results make sense: the impact of additional passengers on the bus is likely to be different during busier periods than less busy periods. However, the RSME values do not decrease significantly: on average, the additional interactions improve estimation accuracy by just 1 second (about 1%). In this situation, the AIC value can be used to assess if the interaction terms should be included (assessing the quality of estimates for the expected value rather than the uncertainty of the estimate). The drawback of the additional interaction

terms is that they require the estimation of many more model coefficients, which can lead to potential over-fitting of the model. However, the AIC value also accounts for the additional number of model coefficients that are estimated and compares this to the improved model fit. Lower AIC values indicate a better model. Comparing Table 6 and Table 7, we see that the models with interaction terms have lower AIC values in all cases, which means that, from a statistical perspective, the models with interaction terms are preferred. Although the improvement of prediction accuracy is marginal, the model form makes more sense and is used to compare the best linear and AFT survival models.

**Table 7. Summary Measures for Travel Time Regression Models Including Interaction Terms**

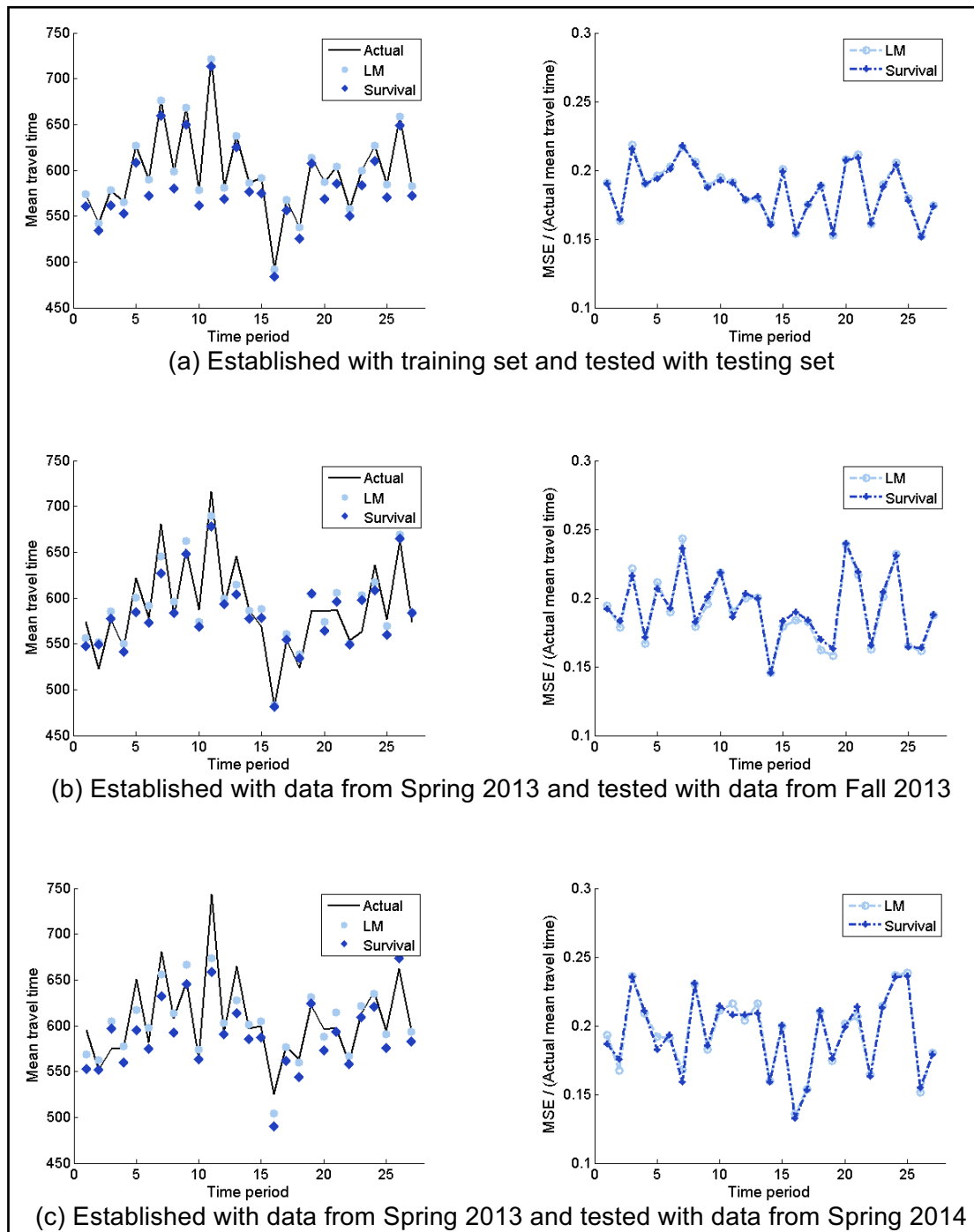
Form	Distribution	Independent variables	R <sup>2</sup>	AIC (unlogged for survival models)	RMSE
Linear	-	Time_Period Time_	0.28	97654	110
AFT Survival	Weibull	Period*Onboard	-	115088	111
		TTPreBus	-	111523	111
	Log-normal	Headway_deviation	-	111234	111
	Log-logistic	Scheduled_Headway_300	-	110957	112
	Generalized Gamma	SnowDepth	-		
		Temperature	-		

## MODEL TESTING AND VALIDATION

After the model establishment, the testing dataset was used to compare the predictive capabilities of the linear and survival models. This included comparisons of both the point estimate of travel time and variances associated with these predictions. The log-logistic survival model was used for comparison due to its low RMSE and AIC. The RMSEs of predictions using the testing data set were essentially identical with the results from training data, which indicated very good predictive validity. This also suggests that the model is not over-fit to the data because it shows similar predictions for the testing dataset. Figure 5a presents the mean travel times and RMSEs from the linear and log-logistic survival models for each of the 27 time periods. As shown, the mean values and RMSEs for predictions from both models are also close to each other and the RMSEs are stable. Regular fluctuations in mean travel times are observed across the day, which reflects different travel behaviors in the peak and off-peak time periods. The fluctuations are more obvious for time periods on MWF (time periods 1 to 15) because the class arrangements on TR are less concentrated. At Time\_Period 7 (noon on MWF), the relative error of predictions showed a high peak. This is likely due to the large fluctuations in demands during lunch time. For all time periods, the mean RMSEs are always less than one-quarter of the mean actual travel time.

The transferability of the models across different semesters was also examined to determine if a model with one set of data accurately reflects future conditions. The transferability of models was tested by estimating the models using data from an earlier semester and then testing the predictions for later semesters. In such a way, the predictive ability of models estimated using historical data can be validated. Figure 5b and Figure 5c show the

prediction results using models estimated with data from Spring 2013 (January-April 2013) and tested with the data from the following two semesters (August-November 2013 and January-April 2014). The linear model and log-logistic survival model performed similarly in the tests, although the log-logistic survival model showed better fitness in some time periods. The prediction results from Spring 2014 are slightly better than Fall 2013 in terms of MSEs, likely due to the different weather conditions and class arrangements between spring and fall semesters. Overall, the MSEs are stable from the tests, which implies that the models are transferable across semesters.



**Figure 5. Average Predicted Travel Time and Relative Errors as a Function of Time\_Period**

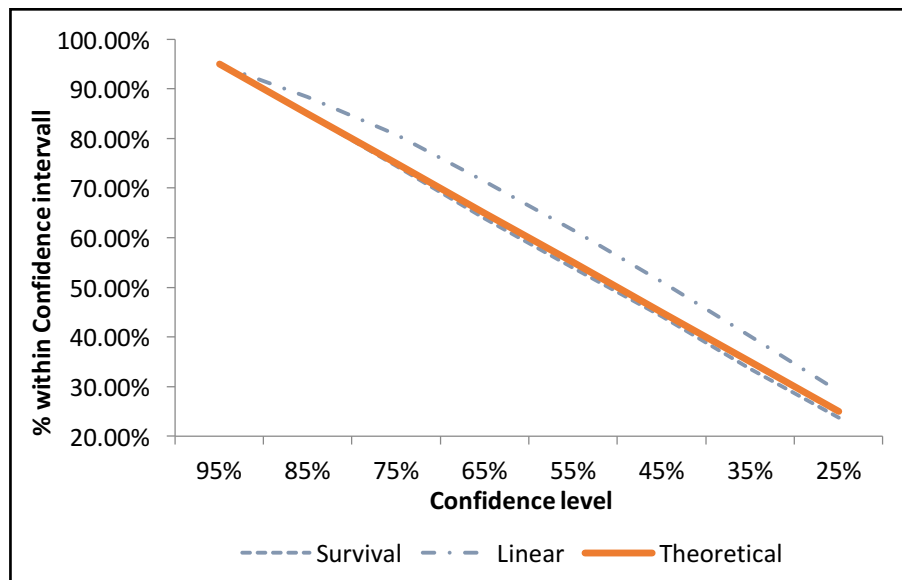
## UNCERTAINTY COMPARISONS

The primary benefit of the AFT survival models is that they provide an estimate of the distribution of travel times, which is contrary to the linear regression model, which provides just a point estimate alone. However, the distribution from the linear regression model can be inferred by recognizing that error terms are normally distributed with a standard deviation equal to the RMSE. Thus, the uncertainty of the travel time estimates can be calculated using both modeling approaches. In the linear regression model, the uncertainty is the same for all estimates. In the AFT survival models, the uncertainty is computed using the distribution obtained from the travel time distribution. However, uncertainty estimates are unique for each estimate because each estimate has a unique travel time distribution. Here, the uncertainties obtained using each modeling approach are compared to highlight the benefits of applying the AFT survival approach for modeling bus travel times.

Using the estimated distribution from the AFT survival models and the inferred distribution from the linear regression models, various confidence intervals were estimated for the travel time estimates. These confidence intervals were used to assess how well each model type could predict travel time uncertainty. Table 8 presents the results. The first column represents the particular confidence interval selected. The second and third columns represent the fraction of actual travel times that are observed within the confidence interval for the travel time estimate. If the models were perfect, one would expect the second and third columns to be equal to the first. Note the fraction of observations within the CI is much closer to what one would hope to expect for the survival model than for the linear model; this is shown graphically in Figure 6, which also contains a line representing the theoretical best-fit. While the linear model fractions are generally higher, this suggests that errors in the estimates are NOT normally distributed, and thus this assumption (which is made in the linear regression model) does not hold.

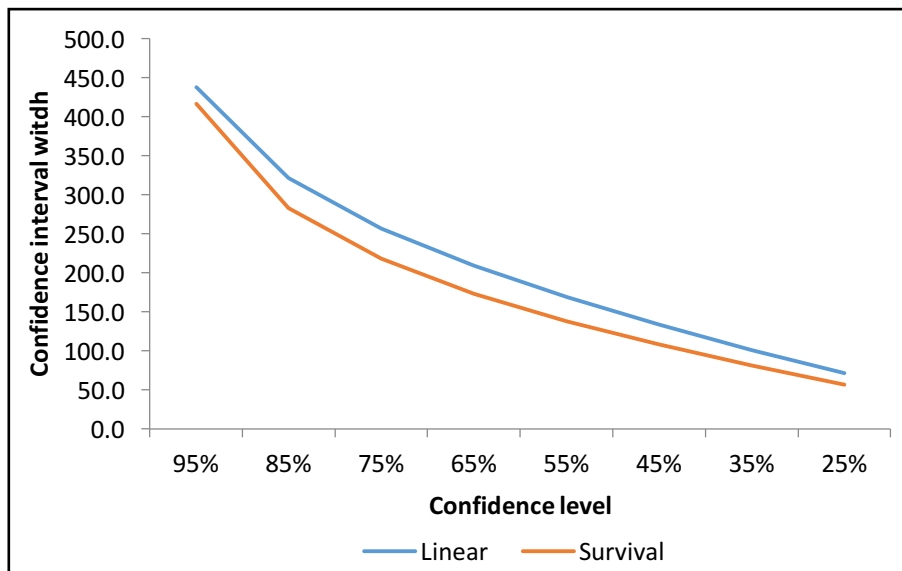
**Table 8. Confidence Intervals for Travel Time Estimates Using Linear and Survival Regression Models**

		% within CI		CI width (sec)	
		Linear	Survival	Linear	Survival
Confidence Level	0.95	94.8%	95.0%	437.8	416.4
	0.85	88.7%	84.5%	321.6	282.5
	0.75	80.9%	74.3%	257.0	218.0
	0.65	71.5%	63.8%	208.8	173.3
	0.55	62.4%	53.8%	168.7	138.1
	0.45	51.6%	44.0%	133.5	108.1
	0.35	40.3%	34.5%	101.4	81.4
	0.25	29.0%	24.5%	71.2	56.9



**Figure 6. Comparison of Confidence Interval with Fraction of Observations that are Observed within Confidence Interval (Travel Time)**

The last two columns of Table 8 provide the width of the various confidence intervals (in seconds). This information reveals that the uncertainty of the estimates is about 12% smaller on average when using the survival model than when using the linear regression model. This can be observed in Figure 7, which provides the size of the confidence interval for any given confidence level. Thus, the survival model estimates are tighter (generally have less uncertainty associated with them) than the linear model estimates.



**Figure 7. Size of Confidence Interval for Given Confidence Level (Travel Times)**

## FACTORS MOST IMPACTING UNCERTAINTY

As shown previously, the survival model can successfully model bus travel times for this particular stop-pair within the Blue Loop and also predict the level of uncertainty associated with these predictions. Therefore, one can use the survival model to unveil key characteristics that might impact travel time uncertainty. This would provide the transit agency with more information on the situations under which travel times are most certain and the situations that cause highly variable travel times. To do this, elasticities were used. The elasticity is defined as the percent change in a particular variable due to a particular change in another. Because the primary goal is to understand travel time uncertainty, the percentage change in the variance associated with the travel time estimate—a measure of how much uncertainty is associated with that estimate—versus a change in each of the explanatory variables is considered. For these calculations, all continuous variables are assumed to be equal to their mean value in the dataset. For the categorical variables, Time\_Period 1, a scheduled headway of 300 seconds and no snow on ground (snowdepth = 0) are selected as the baseline conditions. Elasticities for continuous variables (onboard passenger count, previous bus travel time, deviation from scheduled headway, precipitation, snow depth, and temperature) are provided as the percent change in travel time variance for a 1% increase in that variance. For the categorical variables, they are provided as the percent change in travel time variance from a change in the base condition. The elasticities are presented in Table 9.

**Table 9. Elasticities Describing Influence on Travel Time Variance**

Variable	Category	Elasticity in Variance
Onboard	-	0.067
TTPreBus	-	0.305
Headway_deviation	-	0.011
Temperature	-	0.018
SnowDepth	1	0.002
Time_Period	2	0.012
Time_Period	3	0.014
Time_Period	4	-0.057
Time_Period	5	-0.015
Time_Period	6	-0.026
Time_Period	7	0.013
Time_Period	8	-0.012
Time_Period	9	0.049
Time_Period	10	-0.042
Time_Period	11	0.098
Time_Period	12	-0.024
Time_Period	13	0.013
Time_Period	14	-0.032
Time_Period	15	0.035
Time_Period	16	-0.025
Time_Period	17	0.028
Time_Period	18	-0.068

Variable	Category	Elasticity in Variance
Time_Period	19	-0.003
Time_Period	20	-0.063
Time_Period	21	0.003
Time_Period	22	-0.065
Time_Period	23	0.030
Time_Period	24	-0.017
Time_Period	25	-0.051
Time_Period	26	0.056
Time_Period	27	0.034
Scheduled_Headway 300	0	-0.12

These elasticities reveal key information about travel times on this particular segment. Comparison of these elasticities with the parameter coefficients in reveals that, in general, travel time uncertainty increases as the mean travel time increases. This suggests that estimates of travel time are more uncertain for larger travel times than for smaller travel times. While not surprising, this confirms behavior that should be expected: the predictability of the travel time should decrease as the actual travel time decreases.

Of the continuous variables, the previous bus travel time appears to contribute most to travel time uncertainty. This makes sense, as longer bus travel times would reflect congested travel conditions, which are characterized by unpredictable travel speeds and thus travel times. Onboard passenger count also has one of the highest elasticities, which suggests that more passengers on the bus would increase travel time variability. This is also consistent with engineering intuition, as more passengers would result in more opportunities for the bus to stop at the intermediate stops, increasing travel times along the segment. Weather-related variables appear to have little influence on travel time uncertainty, which is surprising. However, higher amounts of precipitation and snow might cause all vehicles to travel more cautiously, helping to make bus travel times more consistent.

Table 9 shows that Time\_Period 11 (MWF 2:15 p.m.-2:30 p.m.) has the highest elasticity by far. This suggests that travel times are much more uncertain during this period than during the others. Time\_Periods 9, 15, and 26 also have larger positive elasticities, suggesting that travel times are more uncertain during these periods. The smallest elasticities were observed in Times\_Periods 18, 20, and 22, which suggest that travel times are more certain during these day/time combinations than in the rest. A non-300-second scheduled headway has a negative elasticity, which means travel times are more variable when a 300-second scheduled headway is used than others.

---

## VI. MODELING PASSENGER OCCUPANCIES

In this chapter, a framework is developed to predict the passenger occupancy of buses arriving at a downstream stop, given that real-time information would be available at its current or most-recent stop. Fortunately, the data collection issue that created unrealistic travel times did not impact the passenger occupancy values observed at each stop. Therefore, the entire dataset of 230,222 observations was used for this analysis. The list of variables and summary statistics are provided in Table 10. Note that the weather-related variables are coded as dummy variables. Snowdepth was coded as an indicator variable that represented if the snow depth was larger than 30 mm (1.2 inches), while snow fall and precipitation were coded as indicator variables that reflected if their magnitudes were greater than 15 mm (0.6 inches).

Models were created to estimate passenger occupancies for every downstream stop when the bus is currently at any of the stops along the route. To do this in a computationally efficient manner, several model frameworks were proposed. Using these frameworks provided insight into how passenger occupancies should be modeled along bus routes. Because no existing efforts have estimated real-time bus passenger occupancies, there is no existing baseline for comparison. However, three modeling techniques were used: linear regression models, count models, and quantile models. The first two focused on estimates of expected passenger occupancy, while the last provided information on the uncertainty associated with these estimates.

From Table 10, one can see that the average bus passenger occupancy is 14 passengers per bus. However, this is the average value across the entire route. The average passenger occupancy of buses upon leaving each of the bus stops during each Time\_Period is provided in Table A3. This information could be used to provide bus occupancy estimates along the route based only on historical data at each location during any day and time period. These historical estimates will be used as a baseline for comparing the proposed modeling approaches.



**Table 10. Summary Statistics for Data Used in Passenger Occupancy Models (N=230,222)**

Variable	Mean	Std. Dev.	Min.	Max.
Onboard	14.2	13.81	0	80
Future_onboard	14.4	13.92	0	80
Headway_Deviation	136.86	241.32	-600	1200
Temperature	48.91	18.89	6.1	90
Variable	Fraction of observations with value of 1			
Scheduled_Headway 300	0.59			
Scheduled_Headway 360	0.27			
Scheduled_Headway 600	0.12			
Scheduled_Headway 1200	0.02			
Time_Period 1	0.04			
Time_Period 2	0.04			
Time_Period 3	0.01			
Time_Period 4	0.05			
Time_Period 5	0.01			
Time_Period 6	0.05			
Time_Period 7	0.01			
Time_Period 8	0.04			
Time_Period 9	0.01			
Time_Period 10	0.04			
Time_Period 11	0.01			
Time_Period 12	0.04			
Time_Period 13	0.01			
Time_Period 14	0.04			
Time_Period 15	0.18			
Time_Period 16	0.01			
Time_Period 17	0.03			
Time_Period 18	0.03			
Time_Period 19	0.02			
Time_Period 20	0.03			
Time_Period 21	0.04			
Time_Period 22	0.04			
Time_Period 23	0.02			
Time_Period 24	0.02			
Time_Period 25	0.03			
Time_Period 26	0.02			
Time_Period 27	0.12			
SnowDepth 0	0.79			
SnowDepth 1	0.21			
Snow 0	0.22			
Snow 1	0.08			
Precipitation 0	0.78			
Precipitation 1	0.22			

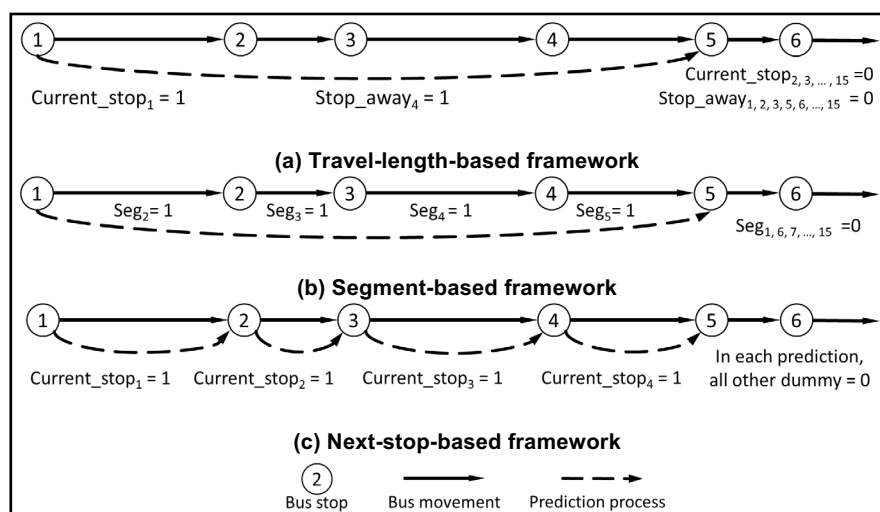
## MODELING FRAMEWORKS

Ideally, it would be desirable to predict the passenger occupancy of a bus as it arrives to any future stop when the bus is currently at any location along the route. One way to do this would be to create a separate model for each of the current-future stop-pairs. For the Blue Loop, this would require the estimation of unique regression models. Such a method might not be scalable to larger routes with more stops. Instead, three modeling frameworks were examined that could be used to model passenger occupancies between all stop pairs in a more scalable way. The next three subsections describe each of these frameworks. They also visually depict each of these frameworks for a hypothetical prediction made for the occupancy of a bus as it arrives to stop 5 after it has left stop 1.

### Travel-Length-Based Framework

For this approach, the onboard passenger count at a randomly selected future stop within the next cycle (i.e., the next 15 downstream stops for that bus) was used as dependent variable in regression. Care was taken to ensure that no operation interruptions occurred between the current observation and the future stop. Two sets of dummy variables were created to represent the current bus location and distance to the future prediction: 1) *Current\_Stop*, a set of dummy variables representing the current bus stop; and, 2) *Stops\_Away*, a set of dummy variables representing the number of stops between the current stop and the stop for which the occupancy prediction will be made. With both of these terms, any current-future stop-pair can be included in the model. Figure 8a illustrates this framework for a hypothetical prediction made for the occupancy at stop 5 when the bus is currently at stop 1. In this example, the dummy variable for the starting stop, *Current\_Stop\_1*, has a value of 1, and the dummy variable for travel length, *Stops\_away\_4*, is set 1. All the other variables for starting stop and travel length are 0.

This model was specified to capture the influence of how far downstream the prediction is being made. It is expected that the further away the prediction (i.e., the higher value of *Stops\_away* used), the less of an effect real-time information from the current bus stop will play on the prediction.



**Figure 8. Graphical Depiction of the Three Modeling Frameworks Considered to Estimate Passenger Occupancies**

## Segment-Based Framework

In this approach, the dependent variable was chosen in the same manner as the travel-length-based framework. Here, a set of 15 dummy variables,  $Seg_i$ , were created to represent the segments between two adjacent stops along the route. Each of these dummy variables indicates if the bus travels on that particular segment between the current and future locations. Figure 8b illustrates the framework for a hypothetical prediction made for the occupancy at stop 5 when the bus is currently at stop 1. Dummy variables for segments between stop 1 and stop 5 ( $Seg_1, Seg_2 \dots Seg_4$ ) all have a value of 1 while the remaining dummy variables ( $Seg_5 \dots Seg_{15}$ ) each have a value of 0.

This framework was easy to understand conceptually and to implement. This framework can be used to reveal the busiest segments along the route (i.e., the segments in which the most passengers enter the bus and the most passengers exit the bus). This framework is also more computationally efficient, as only 15 dummy variables are considered, instead of the 30 dummy variables included in the travel-length-based framework.

## Next-Stop Framework

The two previous frameworks directly model the occupancy at any future stop for any current stop that the bus might be at. In this framework, the future stop selected in the regression model is always the next downstream stop; e.g., if the bus is currently leaving stop 1, the model will estimate the passenger occupancy expected when the bus pulls out of stop 2. Consequently, only a single set of dummy variables are needed to indicate the current stop. This results in a very simple regression procedure, as the dependent variable is always the future bus occupancy after it reaches the next stop.

Predictions of bus occupancy can still be made for future stops that are located multiple stops downstream. In this case, an iterative process is used in which the prediction at the intermediate stop is used as an estimate for the current passenger occupancy at that stop. This is conceptually similar to how an auto regressive model would be applied. Figure 8c illustrates how this framework can be applied for a hypothetical prediction made for the occupancy at stop 5 when the bus is currently at stop 1. First, a prediction will be made for the bus occupancy at stop 2. This prediction would then be used to estimate a prediction for stop 3, and the procedure repeated until a prediction is obtained for stop 5. This is conceptually similar to how an auto-regressive model would be applied.

This framework focused on the relationship between adjacent stops. The model form was simpler, and the results were more predictable. As a result, it should provide better predictions for nearby downstream stops than do other frameworks. However, this framework is not likely to be as accurate for occupancy predictions at farther stops because systematic errors in the predictions will become compounded.

## REGRESSION RESULTS

For each of the three modeling frameworks proposed, a linear model and a zero-inflated negative binomial model were estimated. (Note that traditional NB binomial models were also estimated, but the ZINB outperformed the traditional NB in all cases.) Similar to the travel-time models, baseline models with primary effects (i.e., no interaction terms) were developed first. As shown in Table 11, for each prediction, there is a mean error of just higher than 10 passengers in the linear modeling frameworks. For comparison, the RSME value that would be obtained using the historical passenger occupancies at each stop from Table A3 is 13.2. Therefore, the linear regression models can improve estimation accuracy by about 20%. The linear models outperform the ZINB in terms of AIC and RMSE for all frameworks. Thus, the effects of the variables appear to be additive rather than multiplicative. This makes physical sense because passengers are being added and subtracted as the bus moves between two locations, with separate processes occurring that affect the number of passengers that board and alight at each stop.

**Table 11. Summary Measures for Passenger Occupancy Regression Models without Interactions**

Framework	Form	Predictor	Adj. R <sup>2</sup>	AIC	RMSE
Travel-length-based	LM	Onboard Time LastOccuDiff Current_Stop Stops_away Headway_Deviation Schedule_Headway Temperature Precipitation Snowdepth	0.406	1048386	10.7
	ZINB	Same, Zero model: Onboard Time Current_Stop Stops_away	-	1284807	12.0
Segment-based	LM	Onboard Time LastOccuDiff Seg Headway_Deviation Schedule_Headway Temperature Precipitation Snowdepth	0.433	1039800	10.5
	ZINB	Same, Zero model: Onboard Time Seg	-	1278609	11.7
Next-stop-based	LM	Onboard Time Current_Stop LastOccuDiff Headway_Deviation Schedule_Headway Snowdepth	0.875	790624	4.9
	ZINB	Same, Zero model: Onboard Time Current_Stop	-	1164303	8.1

Table A4 to Table A6 provide the model coefficients for each of the linear model estimates across the three different frameworks. In all model frameworks, the variables headway deviation, scheduled headway, current occupancy, difference in onboard number from the previous bus, precipitation, snow depth, and temperature were found to be statistically significant with signs for the coefficients that are consistent with expectations. *LastOccuDiff* had positive estimates in all models, revealing a positive correlation between the change in occupancies for the previous and current buses. The sign of the *Headway\_Deviation* coefficient was positive for all modeling frameworks, which reaffirms that passenger occupancies will be higher the further a bus falls behind its target headway. The estimates for dummy variable *Schedule\_Headway* consistently observed the following sequence from higher to lower coefficient values: 300s, 360s, 600s, and 1200s. When the scheduled headway is 300s, the average passenger occupancies are higher than in all the other situations, which indicates that the increase in travel demands offsets and exceeds the

impact of shorter time headways between vehicles. Conversely, when the scheduled headway is 600s, the average occupancies on the bus are less than the buses with a 1200s scheduled headway. This indicates that the 600s headway periods have less than double the demand that occurs during the 1200s headway periods. The weather coefficients reveal interesting patterns. *Precipitation* had a negative coefficient, which indicates less bus use during wetter periods. This is reasonable, as people generally make fewer trips while rain is falling. *Temperature* has a negative coefficient, which suggests that buses have lower ridership during warmer temperatures and higher ridership during colder temperatures. *SnowDepth* and *snow* had opposite coefficients. The former represents snow depth on the ground and suggests that bus ridership is lower when there is snow on the ground. The latter represents current snowfalls and suggests that ridership increases during periods with snowfall. This seems reasonable, as snow on the ground might discourage people from making trips, while people might use the bus during snowfall because buses provide a respite from the elements. However, it should also be noted that these findings are limited to a university route serving students and may not represent the general population that uses public transit services.

Models with secondary effects (i.e., those that included interactions between variables) were also estimated to see if they could significantly improve estimation accuracy. As shown in Table 12, the addition of the interaction terms improved model fit (measured by the RMSE, indicating less uncertainty in the estimates) in all cases. The improvement in the RSME with the addition of the interaction terms appear to be about one passenger, which represents about 10% of the total value. Additionally, the reduction in the AIC values suggests that these interactions are statistically valid because the improvement in accuracy is not outweighed by the additional model coefficients.

**Table 12. Summary Measures for Passenger Occupancy Models Including Interaction Terms**

Framework	Form	Predictor	Adj. R <sup>2</sup>	AIC	RMSE
Travel-length-based	LM	Onboard*Time_Period Onboard *Current_Stop*Stops_away LastOccuDiff*Time_Period Time_Period*Current_Stop Headway_Deviation Schedule_Headway Temperature Precipitation Snowdepth	0.475	1025986	10.1
	ZINB	Same, Zero model: Onboard Time_Period Current_Stop*Stops_away (Vuong test statistic = 91.51)	-	1277399	11.3
Segment-based	LM	Onboard*Time_Period Onboard*Seg LastOccuDiff*Time_Period Time_Period*Seg Headway_Deviation Schedule_Headway Temperature Precipitation Snowdepth	0.499	1017647	9.9
	ZINB	Same, Zero model: Onboard Time_Period* Seg (Vuong test statistic = 84.33)	-	1224611	10.4

Framework	Form	Predictor	Adj. R <sup>2</sup>	AIC	RMSE
Next-stop-based	LM	Onboard*Time_Period Onboard*Current_Stop LastOccuDiff*Time_Period Time_Period*Current_Stop Headway_Deviation Schedule_Headway Snowdepth	0.901	748233	4.4
	ZINB	Same, Zero model: Onboard Time_Period*Current_Stop (Vuong test statistic = 126.52)	-	1077642	6.4

The segment-based and travel-length-based models are directly comparable because they both model the passenger occupancy at a randomly chosen downstream stop. Based on the regression results, the best linear model for segment-based framework is slightly better than the travel-length-based framework in terms of adjusted R<sup>2</sup>, AIC, and RMSE values. The RMSEs for the best segment-based and travel-length-based models are about 10, which means that the average error in prediction of future passenger occupancies is about 10 passengers. In comparison, the RMSE for the best next-stop-based model is only 4.4. The result is not surprising because predictions would be more accurate for estimates of passenger occupancy at the next downstream stop compared with stops further downstream. A more direct comparison of these modeling frameworks is provided in the model testing section.

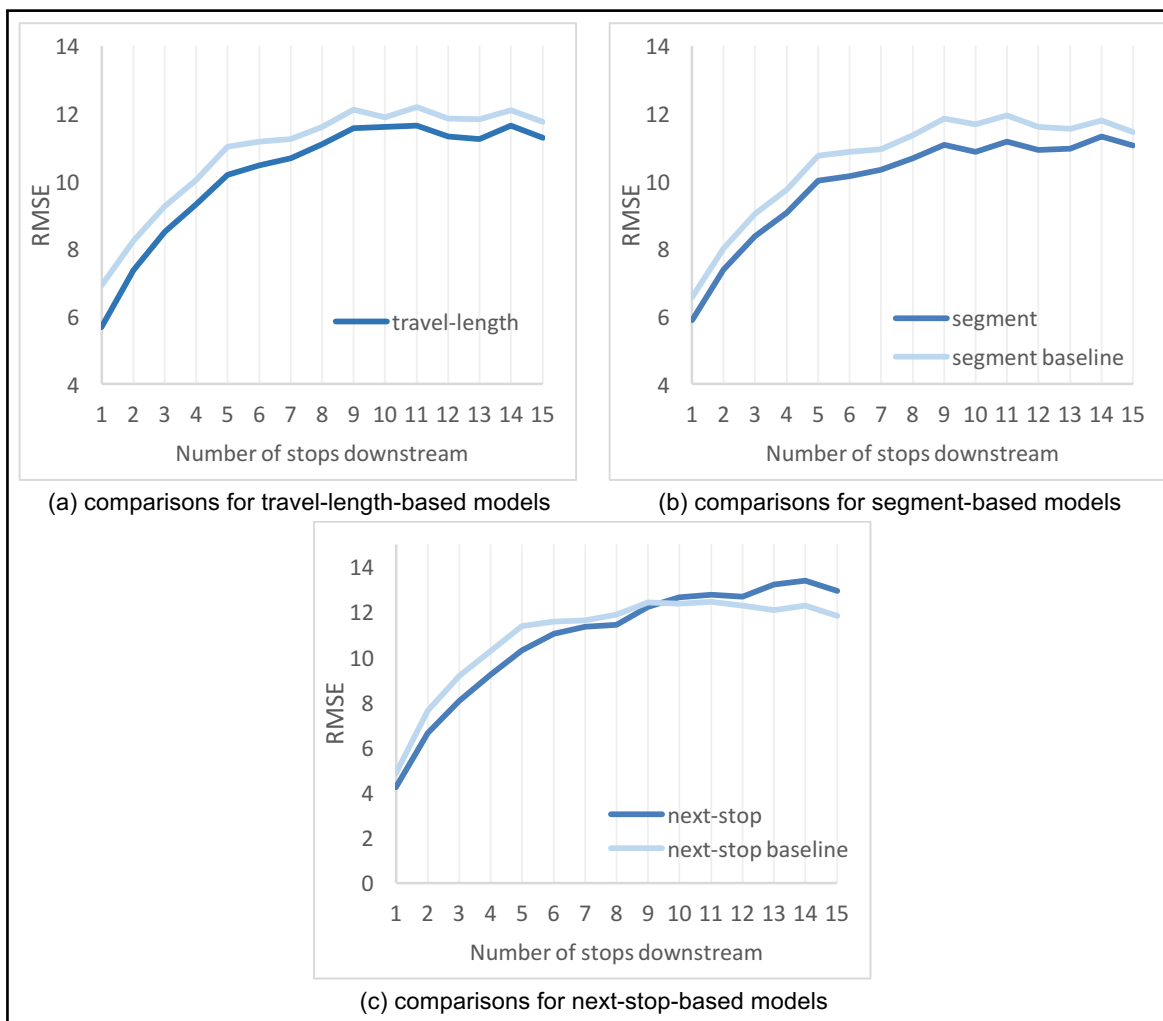
## MODEL TESTING

The prediction capabilities of the regression models were further examined in terms of the prediction accuracy using the testing dataset, the transferability, and the ability to identify full buses. Models from all three frameworks, along with their baseline counterparts, were tested in this section. For the segment-based and travel-length-based frameworks, the regression results were used directly, while for the next-stop-based framework, the iterative prediction approach was used. All predictions smaller than 0 or larger than 80 were manually set to 0 or 80, respectively, as these represent the limits of realistic values. For the next-stop-based framework, the ZINB model was not found to be useful, as multiplicative effects resulted in very large (approaching infinity) or small (0) estimates of passenger occupancy when applied recursively for downstream estimates due to compounding systematic errors in the prediction process.

## Prediction Accuracy

Figure 9 illustrates the accuracy of the three different modeling frameworks (measured in the RMSE) for the primary models (baselines without interactions) and the models with interactions. These models were established using the training dataset, and the results were calculated through applying the established models to the testing dataset. The RMSEs are presented based on the number of stops downstream that the occupancy prediction is being made to better understand the nature of the models. As the previous regression results indicated, travel-length-based models and segment-based models with interaction terms always outperformed their baseline versions, although the magnitudes of differences were not large. For next-stop-based models, the baseline model performed better for predictions more than nine stops downstream. This is likely because the model is built to predict the bus

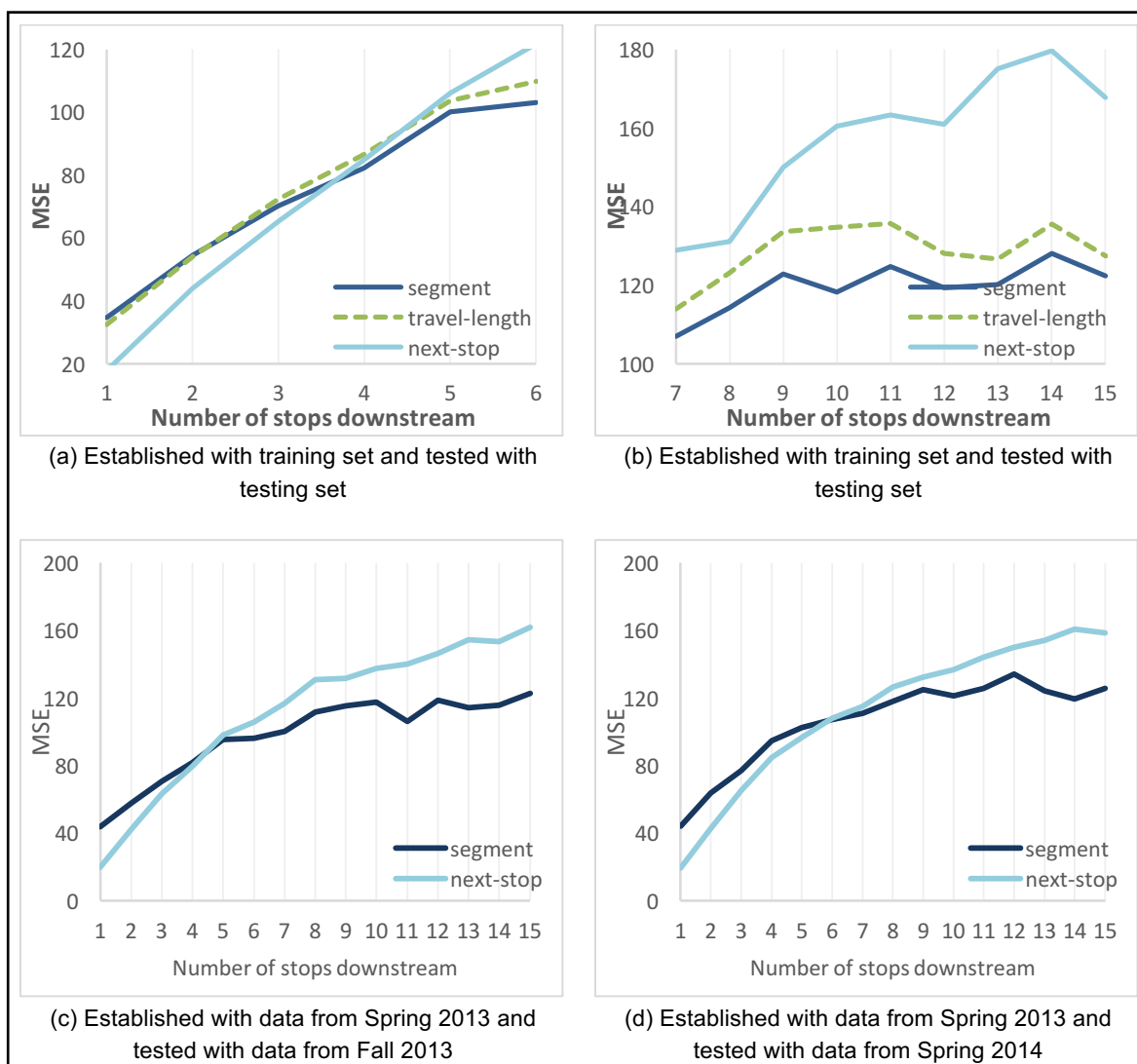
occupancy at the next (downstream) stop, and errors in the predictions build as the model is used to predict bus occupancies further and further downstream. Overall, the models with interaction terms performed better than the baseline models, so the remaining analyses will be only based on models that include interaction terms.



**Figure 9. Comparison of Model Accuracy Across Modeling Frameworks Both Without and With Interaction Terms**

Figure 10a and Figure 10b compare the RMSEs across the three different modeling frameworks at varying numbers of stops downstream. The total RMSEs obtained from the testing data are almost identical to the RMSEs in training data, indicating good predictive validity. The RMSE generally grows as the further downstream future stops are, which is logical, as less uncertainty exists when predicting passenger occupancy for closer stops. Interestingly, the segment-based model almost always performs better than the travel-length-based model. More importantly, the best framework to use changes based on how far away the prediction is being made. The next-stop-based framework performs the best for predictions less than four stops downstream. The next-stop-based and segment-based perform similarly for predictions of four or five stops downstream. The segment-based model performs best for predictions more than five stops downstream, while the next-stop-based model gets very high RMSE values when the future stop gets further. Thus,

this suggests that a combination of frameworks should be used when implementing these models in practice. Predictions for nearby downstream bus stops should be made using the next-stop framework, while predictions further downstream should be made using the segment-based modeling framework.



**Figure 10. Accuracy of Best Passenger Occupancy Models as a Function of How Far Away Prediction is Being Made**

### Model Transferability

The transferability of passenger occupancy models was also tested in a similar way as done with the travel time models. Data from Spring 2013 was used as a training set, and then the models were tested with data from Fall 2013 and Spring 2014. Figure 10c and Figure 10d show the RMSE results from the transferability tests. Based on these results, the same patterns were found with segment-based and next-stop-based frameworks: the model with best predictions switches at around five stops downstream. The predictions for Fall 2013 had slightly higher RMSEs, possibly due to the different patterns (weather and class arrangements) between the spring and fall semesters. Overall, the models provide



stable predictions, which indicate that patterns do not change significantly over time, and the models are transferable over time.

### Identification of Full Buses

A threshold of 60 passengers onboard was used to define a full bus, and this threshold was used to determine how accurately the model predictions were able to identify “full” buses. In general, it is subject to the judgment of the driver if a bus can hold more passengers. This threshold was selected because, in practice, buses start to become crowded with about 60 passengers onboard and drivers are then very likely to prohibit boarding activities. In these test, prediction accuracies from both the segment-based and the next-stop-based frameworks were compared, because both gave fairly similar estimation results. Within five stops downstream, ~80% of full buses can be identified. This accuracy drops to ~65% for all predictions between 1 and 15 stops downstream. Buses that are not full are almost always identified accurately (~98% of the time). The reason for this phenomenon is that: 1) “full” buses occur at more unpredictable high-demand time periods, when more fluctuations and extreme situations occur; and, 2) the occupancy range for “not-full” buses is from 0 to 60, so the tolerance of prediction errors is larger. Overall, both frameworks provide decent identification of full buses within 5 stops downstream. This can be very helpful during high-demand periods when three to four BL buses run concurrently and buses are generally only about four stops apart.

### UNCERTAINTY COMPARISONS

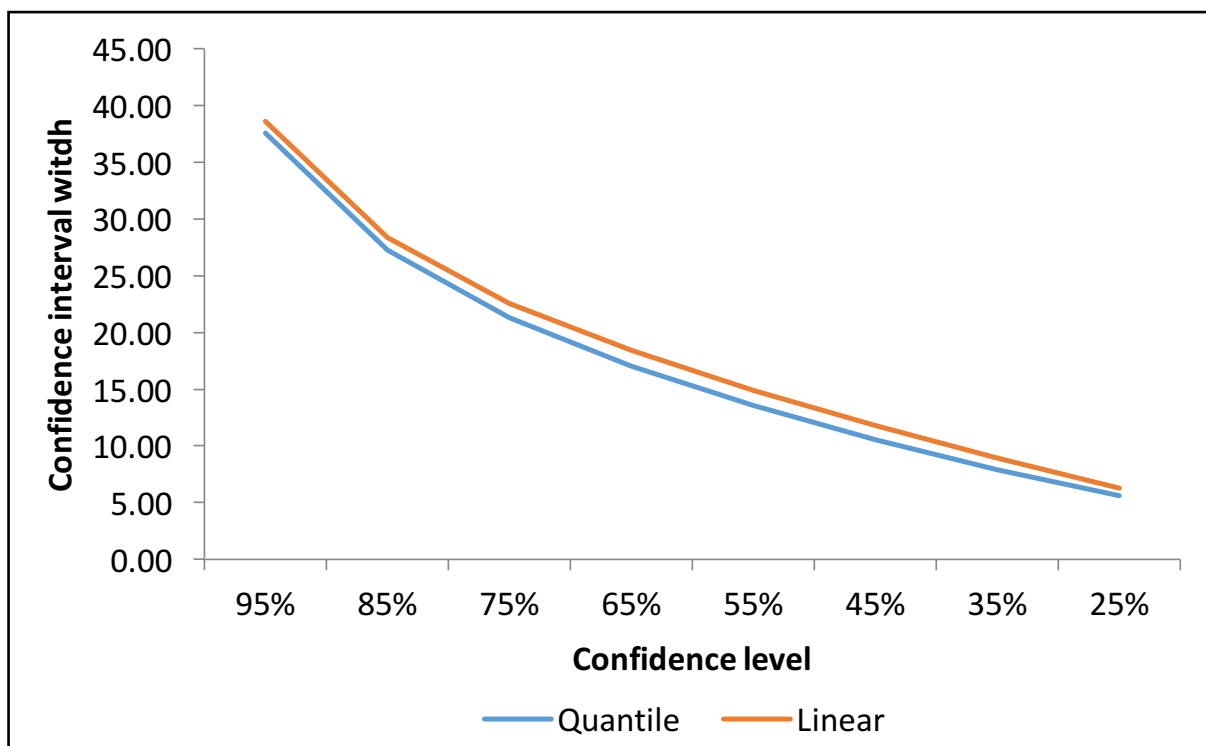
To understand the factors impacting the uncertainty of passenger occupancy estimates, quantile models were developed using the segment-based modeling framework. Quantile model were developed using the same functional form as the linear models. An additive model was assumed because the linear regression models generally outperformed the count regression models, which assumed multiplicative effects. The most important difference between the two model types is that estimates of a quantile model were obtained by minimizing weighted absolute errors, rather than squared errors. The weights for error terms were assigned according to the percentile value of interest as well as the sign for the error. When estimates are made for a low-percentile value, low-percentile observations (which usually have negative errors) receive higher penalties, and thus the estimates are pulled toward them. Generally speaking, the prediction of a certain percentile in a quantile model can be taken as that particular percentile of predicted dependent variable distribution.

Using the quantile model outputs (not shown here for brevity), the confidence interval widths and prediction accuracies were computed. Unique confidence intervals can be obtained for each individual observation. Similar to the travel time models, confidence intervals can also be created using linear models, but the uncertainty associated with each estimate is assumed to be the same across all observations. Nevertheless, the two methods were compared to determine which provided a better fit to the data. Estimates from several percentiles were calculated, and the confidence interval can be obtained from them (e.g. the 95% confidence interval for an observation can be obtained by calculating the 2.5-percentile and 97.5-percentile values). Table 13 provides the confidence intervals and prediction accuracies from the segment-based linear model and quantile model.

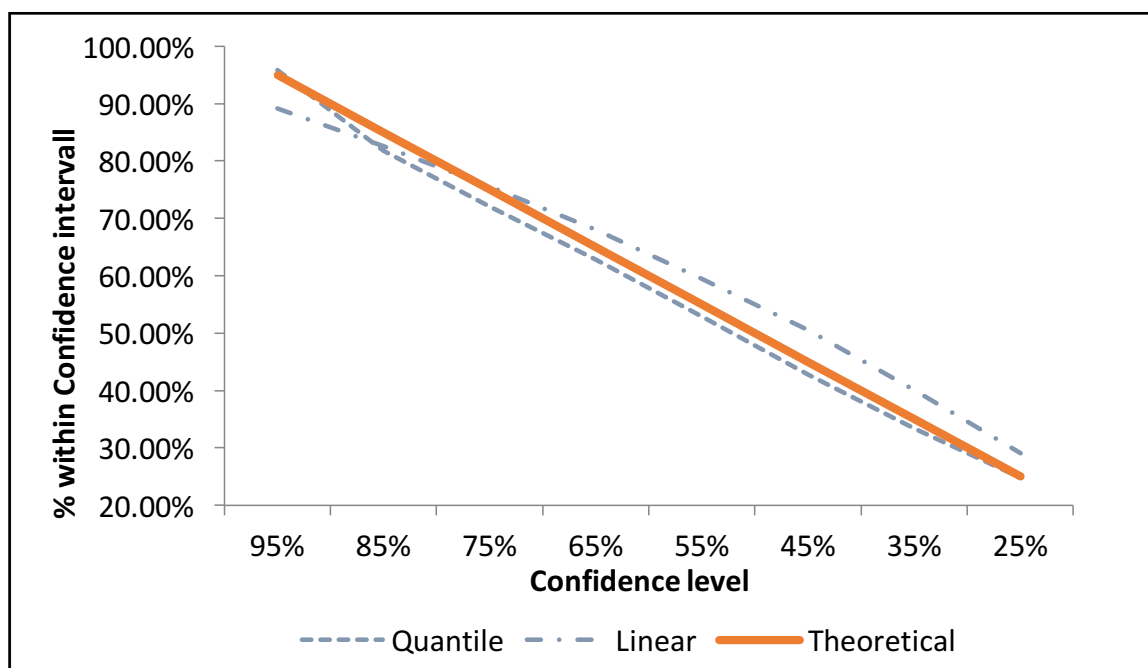
These results were obtained through applying models established with training data to testing dataset. As shown, the quantile models reduced confidence interval widths by between 3% to 12%; a graphical representation is provided in Figure 11. In terms of the prediction accuracies, the results from the quantile model were also closer to theoretical values (Figure 12). Again, the standard errors from the linear model were very likely to be inflated by some extreme values.

**Table 13. Confidence Intervals for Passenger Occupancy Estimates Using Linear and Quantile Regression Models**

		% within CI		CI width	
		Linear	Quantile	Linear	Quantile
Confidence Level	0.95	89.13%	95.87%	38.64	37.6
	0.85	82.59%	81.80%	28.38	27.28
	0.75	75.49%	72.10%	22.58	21.30
	0.65	68.05%	62.78%	18.42	17.03
	0.55	59.49%	52.86%	14.89	13.54
	0.45	50.45%	42.76%	11.78	10.52
	0.35	40.06%	33.41%	8.94	7.90
	0.25	29.09%	24.77%	6.28	5.62



**Figure 11. Size of Confidence Interval for Given Confidence Level (Passenger Occupancies)**



**Figure 12. Comparison of Confidence Interval with Fraction of Observations that are Observed within Confidence Interval (Passenger Occupancies)**

### Factors Most Impacting Uncertainty

It appears that the quantile models can accurately predict the uncertainty that exists in passenger occupancy estimates. Therefore, the quantile model results can be used to unveil key characteristics that most impact uncertainty in passenger occupancies. Based on the previous quantile regression results for the segment-based linear model, Table 14 gives the percentage change in the size of the 65% confidence interval caused by one unit increase in the magnitude of continuous variables; for categorical variables, the increase is caused by the change from basic conditions (Time\_Period = 1, scheduled\_headway = 1200).

As shown, onboard occupancy has a larger impact on interval width than occupancy change recorded for the previous bus. For each onboard passenger, the interval width increases by 2.5%, suggesting that current onboard passenger count significantly impacts the uncertainty of passenger occupancy estimates. Among all the time periods, Time\_Periods 2 and 17 (early morning on MWF and TR) have the most significant impact on the confidence interval width, which indicates higher variances in passenger occupancy predictions across those periods. The smallest interval widths increases were observed in Times\_Periods 5, 7 and 20, which suggest that bus occupancies are more certain during these day/time combinations than the rest. Among stops, the segment from stop 5 to stop 6 (Seg6) and stop 11 to stop 12 (Seg12) lead to highest increase in the interval width. This makes sense because stop 6 (Jordan East) is near the students parking lots, and stop 12 (Atherton Hall) serves the largest dormitory area on campus and has large fluctuations in bus occupancies. Also, observations with shorter schedule headways were found to be associated with wider confidence intervals. For weather variables, the presence of snow on ground and snowing lead to smaller intervals, while other precipitation leads to larger intervals.

**Table 14. Sensitivity of Confidence Interval Size to Various Independent Variables for Passenger Occupancy Model**

Variable	Percentage change in CI
Time_Period 2	5.7
Time_Period 3	-24.0
Time_Period 4	-24.9
Time_Period 5	-35.5
Time_Period 6	-28.2
Time_Period 7	-35.4
Time_Period 8	-29.2
Time_Period 9	-29.0
Time_Period 10	-19.2
Time_Period 11	-24.9
Time_Period 12	-23.3
Time_Period 13	-13.4
Time_Period 14	-19.5
Time_Period 15	-13.2
Time_Period 16	-27.5
Time_Period 17	1.8
Time_Period 18	-27.9
Time_Period 19	-7.9
Time_Period 20	-36.6
Time_Period 21	-21.5
Time_Period 22	-33.4
Time_Period 23	-4.3
Time_Period 24	-30.2
Time_Period 25	-13.8
Time_Period 26	-19.9
Time_Period 27	-12.1
onboard	2.5
LastOccuDiff	1.0
Seg1	7.1
Seg2	4.1
Seg3	-5.4
Seg4	-2.2
Seg5	-6.4
Seg6	11.8
Seg7	-1.3
Seg8	0.8
Seg9	5.3
Seg10	-0.2
Seg11	9.1
Seg12	19.4
Seg13	4.8
Seg14	3.7
Seg15	-2.2

---

Variable	Percentage change in CI
headway_deviation	0.02
scheduled_headway 300	39.5
scheduled_headway 360	25.6
scheduled_headway 600	14.1
precipitation	5.7
snowdepth	-24.0
snow	-24.9
temperature	-0.35

---

## VII. CONCLUDING REMARKS

This project used empirical data from a bus route in State College, PA to develop statistical models for bus travel times and passenger occupancies of individual buses. In both cases, models were created to provide an estimate of the expected value (i.e., mean value) as well as the uncertainty associated with the estimate. For travel times, this was accomplished using accelerated failure time survival models, which are able to predict the distribution of the time until an event occurs. Here, the event referred to the time until the bus arrives to a downstream stop. The AFT survival model was compared with linear regression models—which are fairly common to describe bus travel times—and were found to: 1) predict mean travel times more accurately; 2) accurately model the uncertainty associated with these predictions; and, 3) provide smaller uncertainty ranges—or confidence intervals—for the predictions. The survival models reveal that travel time uncertainty increases as the magnitude of the expected travel times increases. This is not surprising: the larger the travel time expected, the more room there is for uncertainty in the estimate. Nevertheless, it is satisfying that the model confirms this intuitive result. In terms of individual parameter contribution to uncertainty, the AFT survival model reveals that the travel time of the previous bus contributes most to travel time uncertainty. That is, the longer the travel time observed for the previous bus, the more uncertain the estimate of travel time for the current bus. Onboard passenger count also significantly contributes to travel time uncertainty; the more passengers currently on the bus, the less accurately travel times can be estimated. Weather related variables have little impact on travel time uncertainty. The modeling results also reveal that late afternoon peak periods have higher travel time uncertainty than other time periods, which is not surprising because these represent the most congested time periods.

For passenger occupancies, both linear regression models and negative binomial count regression models were considered to predict mean values. The regression results indicate that linear regression models are more appropriate for estimating bus passenger occupancies. This suggests that the impacts of independent variables on bus passenger occupancy are more additive than multiplicative. Three different modeling frameworks were considered to develop a single model to estimate passenger occupancies at all stops along the entire route. The next-stop model was found to be most accurate for passenger occupancy predictions as the bus travels one to five stops immediately downstream of its current location. For predictions of passenger occupancies at stops further away, the segment-based framework was found to be most accurate. Uncertainty estimates were predicted using quantile regression models, which can directly predict any desired confidence interval. The uncertainty analysis reveals that smaller bus headways are associated with more uncertainty in the passenger occupancy estimates. The presence of precipitation and lower temperatures increases passenger occupancy uncertainty, while snow reduces uncertainty. These values also reveal which time periods and segments along the route are most uncertain. A summary of these findings is provided in Table 15.

**Table 15. Summary of Major Findings**

Outcome (predicted)	Conclusions from modeling activities
Travel time	<ul style="list-style-type: none"> <li>• AFT survival models outperform linear regression models</li> <li>• Uncertainty in travel time increases with mean travel time</li> <li>• Travel time of previous bus and current onboard passenger occupancy associated with more uncertain travel times</li> <li>• Weather-related variables have little impact on travel time uncertainty</li> </ul>
Passenger occupancy	<ul style="list-style-type: none"> <li>• Linear regression models outperform count regression models</li> <li>• “Next-stop” modeling framework most accurate for predictions 1-5 stops away</li> <li>• “Segment-based” modeling framework most accurate for predictions &gt;5 stops away</li> <li>• Quantile regression model accurately describes uncertainty associated with estimates</li> <li>• Smaller bus headways found to have more uncertain passenger occupancies than larger bus headways</li> <li>• Precipitation and lower temperatures increase uncertainty, while snow reduced uncertainty in passenger occupancies</li> </ul>

One danger of developing statistical models for transit data is over-fitting of the model due to the large number of parameter coefficients that must be estimated. To address this concern, the transferability of the model across semesters was examined. The results revealed that models developed for one semester are appropriate for estimation in the next, which suggest that these models can be readily applied for prediction purposes. This is promising, as it suggests that the model is not only describing the data used to estimate the model; rather, the model is identifying general trends that are consistent across time and can be used for estimation purposes.

While the models created here are illustrative and cannot be applied directly to another route, they provide evidence that the proposed modeling approaches are feasible for modeling travel time and passenger occupancies on bus transit systems as well as a modeling framework that can be used to develop prediction models for other bus transit systems. These models are fairly easy to estimate using off-the-shelf statistical packages, and thus these modeling approaches can be easily applied to any transit system with the type of data used here. Furthermore, the estimates are computationally simple to apply, which suggests that these models can be readily applied to any real-time transit information system.

Of course, this study was limited by the data quality issues described in Chapter III. Some (actual) extreme values were likely eliminated from the dataset erroneously during the data cleaning stage. The use of a higher-quality dataset can overcome this issue for future efforts at modeling real-time bus occupancies. Future work in predicting passenger occupancy should also consider methods for estimating confidence intervals along with the point estimate for the mean. For the travel time models, only a single stop pair was considered. However, this modeling approach can and should be extended for all stop pairs in the network. To reduce the computational burden of doing this, the modeling frameworks developed for the passenger occupancy models should be examined for travel time estimations such that a single model can reflect travel times between any stop pair along the route.

## APPENDIX A: MODELING OUTPUTS AND TECHNICAL DETAILS

**Table A1. Historical Travel Time Estimates Based on Time of Day**

<b>Time_Period</b>	<b>Mean Travel Time (sec)</b>
1	581.18
2	542.06
3	591.86
4	566.47
5	628.86
6	586.75
7	673.69
8	597.15
9	668.95
10	578.49
11	718.85
12	590.63
13	651.41
14	590.81
15	591.83
16	491.11
17	570.18
18	546.54
19	611.98
20	586.11
21	602.6
22	559.1
23	600.68
24	626.5
25	588.37
26	666.45
27	585.36



**Table A2. Parameter Estimates for Linear Travel Time Model**

R-Square		RMSE	AIC
0.27		111.38	97717
Parameter	Estimate	St. dev	T value
Intercept	329.4240	7.6450	43.09
Time_Period 1	9.9528	7.0825	1.41
Time_Period 2	-11.3748	7.0267	-1.62
Time_Period 3	-32.1247	11.0738	-2.9
Time_Period 4	-36.8865	6.5430	-5.64
Time_Period 5	8.3536	10.3619	0.81
Time_Period 6	-17.9529	6.4422	-2.79
Time_Period 7	25.3866	11.3375	2.24
Time_Period 8	-14.8772	6.5877	-2.26
Time_Period 9	25.9732	10.2659	2.53
Time_Period 10	-27.3764	6.5321	-4.19
Time_Period 11	83.9700	10.7280	7.83
Time_Period 12	-29.3142	6.7663	-4.33
Time_Period 13	19.1800	10.5166	1.82
Time_Period 14	-22.8994	6.6366	-3.45
Time_Period 15	5.0368	4.5570	1.11
Time_Period 16	-25.5544	12.3104	-2.08
Time_Period 17	9.7274	7.4935	1.3
Time_Period 18	-48.3557	7.3076	-6.62
Time_Period 19	-1.5464	8.8863	-0.17
Time_Period 20	-34.8232	8.0494	-4.33
Time_Period 21	-9.9169	7.0516	-1.41
Time_Period 22	-51.4502	6.9534	-7.4
Time_Period 23	-3.7375	8.9796	-0.42
Time_Period 24	-8.5582	7.9614	-1.07
Time_Period 25	-15.6356	7.7398	-2.02
Time_Period 26	18.0299	9.6224	1.87
Time_Period 27	0.0000	.	.
Onboard Passenger Count	1.7170	0.1153	14.89
Travel Time of Previous Bus	0.3368	0.0095	35.3
Deviation from Scheduled Headway	0.0650	0.0067	9.64
Scheduled Headway = 300 Seconds	61.4502	3.2741	18.77
snowdepth	3.7864	3.3006	1.15
temperature	0.0883	0.0787	1.12

**Table A3. Historical Passenger Occupancy Estimates Based on Time of Day and Stop Location**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	7.9	3.9	2.8	2.6	6.0	18.2	18.4	17.5	20.5	14.0	12.2	17.4	13.6	17.9	14.6
2	8.5	5.2	4.3	4.3	5.7	17.8	17.9	17.7	19.4	13.5	12.4	15.7	13.2	15.3	13.9
3	22.7	17.0	7.6	7.3	7.0	16.9	18.4	19.2	21.7	16.3	19.8	29.3	25.1	27.6	24.7
4	11.3	9.1	6.0	6.1	6.0	11.8	11.5	11.6	12.8	10.1	11.1	14.9	13.7	16.4	14.8
5	24.8	22.3	13.1	12.8	10.1	11.7	11.5	10.4	13.8	11.8	16.1	26.2	27.4	33.2	29.9
6	12.0	10.0	7.8	7.7	6.7	9.5	9.2	9.4	10.4	10.1	10.3	13.7	12.9	14.4	13.9
7	28.2	25.7	18.6	15.6	10.2	10.3	10.8	10.8	12.0	13.1	17.9	26.8	26.0	29.2	31.5
8	14.9	12.9	9.4	8.8	7.3	9.1	8.7	8.7	10.2	9.9	11.0	15.3	16.3	17.2	16.4
9	26.9	23.7	13.6	13.6	8.1	9.3	9.3	9.2	10.7	12.4	18.3	24.8	29.1	30.2	28.8
10	17.0	16.4	11.8	12.2	8.6	8.2	7.9	8.3	9.4	9.8	11.1	16.7	17.9	19.3	18.7
11	29.6	27.4	18.8	18.6	12.7	9.3	7.9	7.5	9.6	11.9	18.1	27.7	30.2	32.9	31.9
12	17.0	19.0	14.0	14.3	9.8	7.5	7.6	7.6	8.4	8.7	11.5	15.5	17.1	17.6	16.1
13	32.6	33.0	24.0	20.2	12.3	7.1	7.0	7.4	8.1	10.9	18.7	25.4	27.8	26.7	26.6
14	21.8	25.3	19.9	20.3	12.9	8.2	8.0	8.1	8.2	9.5	12.2	17.0	19.6	19.4	19.0
15	21.9	24.6	18.8	18.2	12.8	8.6	8.6	9.3	10.3	11.7	14.0	18.3	18.6	18.4	18.4
16	2.7	2.2	1.9	1.7	3.4	10.5	9.9	10.8	10.5	7.0	4.7	6.3	4.8	5.0	4.3
17	8.6	5.3	4.2	4.1	6.3	19.1	18.4	18.5	19.9	14.0	12.1	16.3	12.9	14.6	12.7
18	10.7	7.2	4.8	4.9	5.6	14.7	14.9	15.7	17.2	11.6	11.1	16.6	14.3	17.7	15.4
19	13.4	8.8	6.5	6.3	7.2	14.8	14.1	14.2	17.1	14.5	14.6	21.3	18.1	22.1	20.1
20	16.6	14.5	8.5	9.9	8.0	10.2	11.3	11.4	13.2	12.1	13.5	17.9	18.2	19.6	18.7
21	16.7	16.3	11.0	10.6	8.4	10.3	9.9	10.3	11.9	11.8	14.2	18.4	18.7	19.3	18.5
22	18.2	16.5	10.9	11.5	8.6	9.3	9.7	9.4	10.2	10.3	12.3	17.3	17.4	19.1	18.4
23	20.2	19.3	14.0	12.7	9.5	10.5	9.5	10.1	12.5	12.9	15.0	19.8	21.1	22.4	22.4
24	22.6	23.6	17.2	17.2	12.2	9.9	9.7	9.5	10.9	12.2	15.8	20.4	21.7	23.2	22.3
25	21.2	23.1	15.1	15.8	10.6	8.2	7.5	7.8	8.7	9.8	13.1	17.5	19.7	19.0	18.7
26	32.7	35.7	27.2	27.8	16.9	10.7	10.2	10.0	11.2	13.6	17.2	24.3	27.5	28.4	29.6
27	21.9	24.6	19.2	18.9	12.5	7.9	8.2	8.8	9.5	11.0	13.5	18.0	18.9	18.3	18.3

Rows represent Time\_Period variable, while columns represent stop. Green represents smaller values, while red represents larger values.

**Table A4. Parameter Estimates for Linear Passenger Occupancy Model  
(Travel-Length Framework)**

R-Square		Root MSE	AIC
0.409		10.7	1047565
Parameter	Estimate	St. dev	T value
Intercept	3.1194	0.4018	7.76
Time_Period 1	-0.9849	0.1459	-6.75
Time_Period 2	0.9572	0.1479	6.47
Time_Period 3	-5.0445	0.2584	-19.52
Time_Period 4	-1.1759	0.1426	-8.24
Time_Period 5	-5.6532	0.2377	-23.78
Time_Period 6	-1.2269	0.1436	-8.54
Time_Period 7	-4.8590	0.2456	-19.78
Time_Period 8	-0.9010	0.1448	-6.22
Time_Period 9	-3.9814	0.2486	-16.02
Time_Period 10	-0.3150	0.1452	-2.17
Time_Period 11	-4.1555	0.2505	-16.59
Time_Period 12	-0.7661	0.1470	-5.21
Time_Period 13	-1.4610	0.2427	-6.02
Time_Period 14	-0.3127	0.1451	-2.15
Time_Period 15	-0.3306	0.0933	-3.54
Time_Period 16	-2.9043	0.2383	-12.19
Time_Period 17	0.7651	0.1664	4.6
Time_Period 18	-2.5340	0.1580	-16.04
Time_Period 19	1.0282	0.2021	5.09
Time_Period 20	-4.4494	0.1746	-25.48
Time_Period 21	-0.4890	0.1560	-3.13
Time_Period 22	-2.0661	0.1568	-13.17
Time_Period 23	2.4636	0.2032	12.13
Time_Period 24	-3.0845	0.1799	-17.15
Time_Period 25	1.0436	0.1734	6.02
Time_Period 26	-2.2596	0.2174	-10.39
Time_Period 27	0.0000	.	.
onboard	0.5970	0.0020	295.03
LastOccuDiff	0.3091	0.0020	155.55
Stops_Away 1	0.1200	0.1336	0.9
Stops_Away 2	0.0686	0.1375	0.5
Stops_Away 3	-0.0080	0.1400	-0.06
Stops_Away 4	-0.1355	0.1420	-0.95
Stops_Away 5	-0.1109	0.1440	-0.77
Stops_Away 6	-0.0548	0.1445	-0.38
Stops_Away 7	-0.0750	0.1461	-0.51
Stops_Away 8	-0.0289	0.1471	-0.2
Stops_Away 9	-0.0931	0.1491	-0.62
Stops_Away 10	0.0463	0.1490	0.31
Stops_Away 11	-0.0704	0.1509	-0.47
Stops_Away 12	0.0525	0.1515	0.35

---

Parameter	Estimate	St. dev	T value
Stops_Away 13	-0.1231	0.1513	-0.81
Stops_Away 14	-0.0699	0.1532	-0.46
Stops_Away 15	0.0000	.	.
Current_Stop 1	0.0038	0.1366	0.03
Current_Stop 2	-0.0076	0.1365	-0.06
Current_Stop 3	1.1773	0.1374	8.57
Current_Stop 4	1.3709	0.1397	9.81
Current_Stop 5	2.4058	0.1407	17.10
Current_Stop 6	2.0909	0.1417	14.76
Current_Stop 7	2.3240	0.1382	16.81
Current_Stop 8	2.2584	0.1381	16.36
Current_Stop 9	2.0202	0.1376	14.68
Current_Stop 10	2.3123	0.1373	16.84
Current_Stop 11	1.7886	0.1368	13.07
Current_Stop 12	0.3089	0.1362	2.27
Current_Stop 13	0.0908	0.1365	0.66
Current_Stop 14	-0.1343	0.1360	-0.99
Current_Stop 15	0.0000	.	.
headway_deviation	0.0005	0.0001	4.85
schedule_headway 300	4.0715	0.2136	19.07
schedule_headway 360	2.5790	0.2124	12.14
schedule_headway 600	1.2001	0.2185	5.49
schedule_headway 1200	0.0000	.	.
precipitation	-0.2573	0.0746	-3.45
snowdepth	-0.6308	0.0771	-8.18
snow	0.1861	0.1335	1.39
temperature	-0.0187	0.0079	-2.36

---

**Table A5. Parameter Estimates for Linear Passenger Occupancy Model  
(Segment-Based Framework)**

R-Square		Root MSE	AIC
0.436		10.4	1038767
Parameter	Estimate	St. dev	T value
Intercept	3.8256	0.3653	10.47
Time_Period 1	-0.8623	0.1424	-6.05
Time_Period 2	1.0124	0.1443	7.01
Time_Period 3	-5.1268	0.2522	-20.33
Time_Period 4	-1.1150	0.1392	-8.01
Time_Period 5	-5.6461	0.2321	-24.33
Time_Period 6	-1.1907	0.1402	-8.5
Time_Period 7	-4.9227	0.2397	-20.53
Time_Period 8	-0.8880	0.1413	-6.28
Time_Period 9	-4.0139	0.2426	-16.54
Time_Period 10	-0.2957	0.1417	-2.09
Time_Period 11	-4.1003	0.2445	-16.77
Time_Period 12	-0.7487	0.1435	-5.22
Time_Period 13	-1.3412	0.2370	-5.66
Time_Period 14	-0.3200	0.1416	-2.26
Time_Period 15	-0.3164	0.0911	-3.47
Time_Period 16	-2.8162	0.2326	-12.11
Time_Period 17	0.9220	0.1624	5.68
Time_Period 18	-2.6307	0.1542	-17.06
Time_Period 19	1.3422	0.1973	6.8
Time_Period 20	-4.6262	0.1705	-27.14
Time_Period 21	-0.2977	0.1523	-1.95
Time_Period 22	-2.1426	0.1531	-14
Time_Period 23	2.7159	0.1983	13.7
Time_Period 24	-3.3238	0.1756	-18.93
Time_Period 25	1.2416	0.1692	7.34
Time_Period 26	-2.2545	0.2122	-10.62
Time_Period 27	0.0000	.	.
onboard	0.6057	0.0019	312.33
LastOccuDiff	0.2536	0.0020	125.7
Seg1	-0.0897	0.0933	-0.96
Seg2	-0.1431	0.0927	-1.54
Seg3	-2.5267	0.0916	-27.58
Seg4	-0.4178	0.0905	-4.62
Seg5	-1.6803	0.0959	-17.52
Seg6	0.4047	0.0990	4.09
Seg7	-0.0637	0.0974	-0.65
Seg8	0.1390	0.0948	1.47
Seg9	0.6634	0.0948	7
Seg10	-0.2205	0.0943	-2.34
Seg11	0.9673	0.0934	10.35

---

Parameter	Estimate	St. dev	T value
Seg12	2.7273	0.0939	29.05
Seg13	0.1334	0.0930	1.43
Seg14	0.5292	0.0931	5.69
Seg15	-0.4318	0.0928	-4.65
headway_deviation	0.0005	0.0001	4.37
schedule_headway 300	4.1120	0.2082	19.75
schedule_headway 360	2.6113	0.2072	12.6
schedule_headway 600	1.2047	0.2132	5.65
schedule_headway 1200	0.0000	.	.
precipitation	-0.2509	0.0728	-3.45
snowdepth	-0.6351	0.0753	-8.44
snow	0.1946	0.1303	1.49
temperature	-0.0116	0.0077	-1.51

---

**Table A6. Parameter Estimates for Linear Passenger Occupancy Model  
(Next-Stop Framework)**

R-Square		Root MSE	AIC
0.864		5.0	806044
Parameter	Estimate	St. dev	T value
Intercept	0.8158	0.1114	7.32
Time_Period 1	-0.1867	0.0668	-2.8
Time_Period 2	0.1038	0.0699	1.49
Time_Period 3	-0.4397	0.1195	-3.68
Time_Period 4	-0.3812	0.0668	-5.7
Time_Period 5	-0.2778	0.1122	-2.48
Time_Period 6	-0.4754	0.0673	-7.06
Time_Period 7	-0.0219	0.1153	-0.19
Time_Period 8	-0.3862	0.0676	-5.71
Time_Period 9	0.1490	0.1163	1.28
Time_Period 10	-0.2747	0.0683	-4.02
Time_Period 11	0.6478	0.1162	5.58
Time_Period 12	-0.5161	0.0685	-7.53
Time_Period 13	0.9871	0.1127	8.75
Time_Period 14	-0.2788	0.0677	-4.12
Time_Period 15	-0.0198	0.0436	-0.45
Time_Period 16	-0.7801	0.1041	-7.49
Time_Period 17	0.3041	0.0776	3.92
Time_Period 18	-0.8294	0.0740	-11.21
Time_Period 19	0.7915	0.0941	8.42
Time_Period 20	-1.0137	0.0820	-12.36
Time_Period 21	0.3009	0.0734	4.1
Time_Period 22	-0.7317	0.0736	-9.94
Time_Period 23	0.8334	0.0956	8.72
Time_Period 24	-0.7689	0.0849	-9.05
Time_Period 25	0.2629	0.0818	3.21
Time_Period 26	0.0982	0.1021	0.96
Time_Period 27	0.0000	.	.
onboard	0.9199	0.0009	980.41
Current_Stop 1	0.0963	0.0638	1.51
Current_Stop 2	-4.3634	0.0638	-68.39
Current_Stop 3	-0.5333	0.0644	-8.28
Current_Stop 4	-3.4306	0.0654	-52.49
Current_Stop 5	0.3509	0.0661	5.31
Current_Stop 6	-0.4408	0.0659	-6.69
Current_Stop 7	-0.0966	0.0642	-1.5
Current_Stop 8	0.8357	0.0642	13.01
Current_Stop 9	-0.5426	0.0640	-8.47
Current_Stop 10	1.4492	0.0639	22.66
Current_Stop 11	4.6402	0.0637	72.82
Current_Stop 12	0.2993	0.0635	4.71
Current_Stop 13	1.1797	0.0638	18.49

---

Parameter	Estimate	St. dev	T value
Current_Stop 14	-0.5004	0.0636	-7.87
Current_Stop 15	0.0000	.	.
headway_deviation	0.0005	0.0001	9.13
schedule_headway 300	0.6433	0.1038	6.2
schedule_headway 360	0.4540	0.1034	4.39
schedule_headway 600	0.1155	0.1062	1.09
schedule_headway 1200	0.0000	.	.
snowdepth	-0.1221	0.0319	-3.82

---



---

## GLOSSARY OF ABBREVIATIONS

---

AFT	Accelerated failure time
ANN	Artificial neural network
APC	Automated passenger counters
AVL	Automatic vehicle location
BL	Blue Loop
CATA	Centre Area Transportation Authority
KF	Kalman-filter
MWF	Monday, Wednesday, and Friday
NB	Negative binomial
PSU	Pennsylvania State University
RMSE	Root mean square error
TR	Tuesday and Thursday
ZINB	Zero-inflated negative binomial

---

## BIBLIOGRAPHY

- Alfa, A.S., W.B. Menzies, J. Purcha, R. Mcpherson, 1988. A regression model for bus running times in suburban areas of Winnipeg. *Journal of Advanced Transportation* 21(3), 227-237.
- Anas, A., 1981. The estimation of multinomial logit models of joint location and travel model choice from aggregated data. *Journal of Regional Science* 21(2), 223-242.
- Bates, J., J. Polak, P. Jones, A. Cook, 2001. The valuation of reliability for personal travel. *Transportation Research Part E: Logistics and Transportation Review* 37(2), 191-229.
- Ben-Akiva, M., B. Boccara, 1995. Discrete choice models with latent choice sets. *International Journal of Research in Marketing* 12(1), 9-24.
- Brownstone, D., K.A. Small, 2005. Valuing time and reliability: assessing the evidence from road pricing demonstrations. *Transportation Research Part A: Policy and Practice* 39(4), 279-293.
- Carrel, A., A. Halvorsen, J. Walker, 2013. Passengers' perception of and behavioral adaptation to unreliability in public transportation. *Transportation Research Record: Journal of the Transportation Research Board* (2351), 153-162.
- Cathey, F., D. Dailey, 2003. A prescription for transit arrival/departure prediction using automatic vehicle location data. *Transportation Research Part C: Emerging Technologies* 11(3), 241-264.
- Chen, G., X. Yang, J. An, D. Zhang, 2011. Bus-arrival-time prediction models: Link-based and section-based. *Journal of Transportation Engineering* 138(1), 60-66.
- Chen, M., X. Liu, J. Xia, S.I. Chien, 2004. A dynamic bus-arrival time prediction model based on APC data. *Computer-Aided Civil and Infrastructure Engineering* 19(5), 364-376.
- Chen, M., J. Yaw, S.I. Chien, X. Liu, 2007. Using automatic passenger counter data in bus arrival time prediction. *Journal of Advanced Transportation* 41(3), 267-283.
- Chien, S.I.-J., Y. Ding, C. Wei, 2002. Dynamic bus arrival time prediction with artificial neural networks. *Journal of Transportation Engineering* 128(5), 429-438.
- Christofa, E., A. Skabardonis, 2011. Traffic signal optimization with application of transit signal priority to an isolated intersection. *Transportation Research Record: Journal of the Transportation Research Board* (2259), 192-201.

- Conrad, M., F. Dion, S. Yagar, 1998. Real-time traffic signal optimization with transit priority: Recent advances in the signal priority procedure for optimization in real-time model. *Transportation Research Record: Journal of the Transportation Research Board*(1634), 100-109.
- Daganzo, C.F., 2009. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B: Methodological* 43(10), 913-921.
- Daganzo, C.F., V.V. Gayah, E.J. Gonzales, 2012. The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions. *EURO Journal on Transportation and Logistics* 1(1-2), 47-65.
- Daganzo, C.F., J. Pilachowski, 2011. Reducing bunching with bus-to-bus cooperation. *Transportation Research Part B: Methodological* 45(1), 267-277.
- Delgado, F., J. Muñoz, R. Giesen, A. Cipriano, 2009. Real-time control of buses in a transit corridor based on vehicle holding and boarding limits. *Transportation Research Record: Journal of the Transportation Research Board* (2090), 59-67.
- Eichler, M., C.F. Daganzo, 2006. Bus lanes with intermittent priority: Strategy formulae and an evaluation. *Transportation Research Part B: Methodological* 40(9), 731-744.
- Frechette, L., A. Khan, 1998. Bayesian regression-based urban traffic models. *Transportation Research Record: Journal of the Transportation Research Board*(1644), 157-165.
- García-Ferrer, A., A. de Juan, P. Poncela, M. Bujosa, 2004. Monthly forecasts of integrated public transport systems: The case of the Madrid metropolitan area. *Journal of Transportation and Statistics* 7(1), 39-59.
- Glazer, A., E. Niskanen, 2000. Which consumers benefit from congestion tolls? *Journal of Transport Economics and Policy*, 43-53.
- Golob, T.F., E.T. Canty, R.L. Gustafson, J.E. Vitt, 1972. An analysis of consumer preferences for a public transportation system. *Transportation Research* 6(1), 81-102.
- Gonzales, E.J., C.F. Daganzo, 2012. Morning commute with competing modes and distributed demand: User equilibrium, system optimum, and pricing. *Transportation Research Part B: Methodological* 46(10), 1519-1534.
- Greene, W.H., 2011. *Econometric Analysis*, 7th ed. Prentice Hall, Upper Saddle River, NJ.
- Guler, S.I., M.J. Cassidy. 2012. Strategies for sharing bottleneck capacity among buses and cars. *Transportation research part B: Methodological* 46(10), 1334-1345.

- Guler, S.I., M. Menendez, 2014. Analytical formulation and empirical evaluation of pre-signals for bus priority. *Transportation Research Part B: Methodological* 64, 41-53.
- Haykin, S., 2004. *Kalman filtering and neural networks*. John Wiley & Sons.
- Hilbe, J.M., 2011. *Negative Binomial Regression*. Cambridge University Press.
- Hunt, P., D. Robertson, R. Bretherton, M.C. Royle, 1982. The SCOOT on-line traffic signal optimisation technique. *Traffic Engineering & Control* 23(4).
- Jariyasunant, J., B. Kerkez, R. Sengupta, S. Glaser, A. Bayen., 2011. Mobile transit trip planning with real-time data.
- Jeong, R., L. Rilett. 2005. Prediction model of bus arrival time for real-time applications. *Transportation Research Record: Journal of the Transportation Research Board* (1927), 195-204.
- Kerkman, K., K. Martens, H. Meurs, 2015. Factors Influencing Stop Level Transit Ridership in the Arnhem Nijmegen City Region, *Transportation Research Board 94th Annual Meeting*.
- Koenker, R., G. Bassett Jr., 1978. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33-50.
- Koenker, R., K. Hallock, 2001. Quantile regression: An introduction. *Journal of Economic Perspectives* 15(4), 43-56.
- Lee, S., M. Hickman, D. Tong, 2013. Development of a temporal and spatial linkage between transit demand and land-use patterns. *Journal of Transport and Land Use* 6(2), 33-46.
- Lord, D., F. Mannering, 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice* 44(5), 291-305.
- Lord, D., S.P. Washington, J.N. Ivan, 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention* 37(1), 35-46.
- Mazloumi, E., G. Currie, G. Rose, 2009. Using GPS data to gain insight into public transport travel time variability. *Journal of Transportation Engineering* 136(7), 623-631.
- National Oceanic and Atmospheric Administration, 2015. National Oceanic and Atmospheric Administration.
- Neter, J., M.H. Kutner, C.J. Nachtsheim, W. Wasserman, 1996. *Applied linear statistical models*. Irwin Chicago.

- Newell, G.F., 1974. Control of pairing of vehicles on a public transportation route, two vehicles, one control point. *Transportation Science* 8(3), 248-264.
- Newell, G.F., R.B. Potts, 1964. Maintaining a bus schedule, *Australian Road Research Board (ARRB) Conference, 2nd, 1964, Melbourne*.
- Nowlin, L., K. Fitzpatrick, 1997. Performance of queue jumper lanes, *Traffic Congestion and Traffic Safety in the 21st Century: Challenges, Innovations, and Opportunities*.
- Patnaik, J., S. Chien, A. Bladikas, 2004. Estimation of bus arrival times using APC data. *Journal of Public Transportation* 7(1), 1-20.
- Paulley, N., R. Balcombe, R. Mackett, H. Titheridge, J. Preston, M. Wardman, J. Shires, P. White, 2006. The demand for public transport: The effects of fares, quality of service, income and car ownership. *Transport Policy* 13(4), 295-306.
- Pennsylvania State Climatologist, 2015. The Pennsylvania State Climatologist.
- Poch, M., F. Mannering, 1996. Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering* 122(2), 105-113.
- Prashker, J.N., 1979. Direct analysis of the perceived importance of attributes of reliability of travel modes in urban travel. *Transportation* 8(4), 329-346.
- Rosenbaum, P.R., 2002. *Observational Studies*. Springer.
- Ryan, S., L.F. Frank, 2009. Pedestrian environments and transit ridership. *Journal of Public Transportation* 12(1), 3.
- Shalaby, A., A. Farhan, 2004. Prediction model of bus arrival and departure times using AVL and APC data. *Journal of Public Transportation* 7(1), 3.
- Shankar, V., F. Mannering, W. Barfield, 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis & Prevention* 27(3), 371-389.
- Sherman, R., 1972. Subsidies to relieve urban traffic congestion. *Journal of Transport Economics and Policy*, 22-31.
- Skabardonis, A., 2000. Control strategies for transit priority. *Transportation Research Record: Journal of the Transportation Research Board* (1727), 20-26.
- Smith, H.R., B. Hemily, M. Ivanovic, 2005. Transit signal priority (TSP): A planning and implementation handbook.
- Sun, A., M. Hickman, 2005. The real-time stop-skipping problem. *Journal of Intelligent Transportation Systems* 9(2), 91-109.

- Sun, D., H. Luo, L. Fu, W. Liu, X. Liao, M. Zhao, 2007. Predicting bus arrival time on the basis of global positioning system data. *Transportation Research Record: Journal of the Transportation Research Board* (2034), 62-72.
- Vanajakshi, L., Subramanian, S.C., Sivanandan, R., 2009. Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses. *IET Intelligent Transport Systems* 3(1), 1-9.
- Viegas, J., B. Lu, 2001. Widening the scope for bus priority with intermittent bus lanes. *Transportation Planning and Technology* 24(2), 87-110.
- Viegas, J., B. Lu, 2004. The intermittent bus lane signals setting within an area. *Transportation Research Part C: Emerging Technologies* 12(6), 453-469.
- Viegas, J.M., R. Roque, B. Lu, J. Vieira, 2007. Intermittent bus lane system: Demonstration in Lisbon, Portugal, *Transportation Research Board 86th Annual Meeting*.
- Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307-333.
- Washington, S.P., M.G. Karlaftis, F.L. Mannering, 2010. *Statistical and Econometric Methods for Transportation Data Analysis*. CRC press.
- Wu, J., N. Hounsell, 1998. Bus priority using pre-signals. *Transportation Research Part A: Policy and Practice* 32(8), 563-583.
- Xuan, Y., J. Argote, C.F. Daganzo, 2011. Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis. *Transportation Research Part B: Methodological* 45(10), 1831-1845.
- Xuan, Y., Gayah, V., Cassidy, M., Daganzo, C., 2012. Presignal Used to Increase Bus-and Car-Carrying Capacity at Intersections: Theory and Experiment. *Transportation Research Record: Journal of the Transportation Research Board* (2315), 191-196.
- Xuan, Y., Gayah, V., Daganzo, C.F., Cassidy, M.J., 2010. Multimodal traffic at isolated signalized intersections: new management strategies and a framework for analysis, *89th Annual Meeting of the Transportation Research Board, Washington, DC*.

---

## ABOUT THE AUTHORS

### VIKASH V. GAYAH

Vikash V. Gayah, PhD, is an assistant professor in the Department of Civil and Environmental Engineering at the Pennsylvania State University and served as the principal investigator for this project. He received his BS and MS degrees from the University of Central Florida (2005 and 2006, respectively), and his PhD degree from the University of California, Berkeley (2012). His research interests include urban transportation operations and network modeling, transit system operations, traffic safety, and statistical and econometric modeling of transportation data. He has over 10 years of research, teaching, and industry experience. Dr. Gayah currently serves as a member of the Traffic Flow Theory and Characteristics committee (AHB 45) of the Transportation Research Board and is an editorial advisory board member of Transportation Research Part B. His recent recognitions include Outstanding Student of the Year for the University of California Transportation Center (2011-2012), a Dwight D. Eisenhower Graduate Fellowship (2011-2012) and the New Faculty Award presented by Cambridge Systematics and the Council of University of Transportation Centers (2015-2016).

### ZHENGYAO YU

Zhengyao Yu is a PhD candidate in his first year of study in the Department of Civil and Environmental Engineering at Penn State. He received his BS degree (2014) from Tongji University (Shanghai, China) and his MS degree (2015) from Penn State, both in civil engineering. His research interests include traffic network modeling, traffic flow theory, and transit system operations. He is a recipient of an American Public Transportation Foundation Scholarship (2015) and a College of Engineering Recruitment Fund Graduate Fellowship at Penn State (2014).

### JONATHAN S. WOOD

Jonathan S. Wood is a PhD candidate in his final year of study in the Department of Civil and Environmental Engineering at Penn State. He received his BS (2011) and MS (2012) degrees in civil and environmental engineering from the University of Utah. His research interests include advanced statistical and econometric analysis of transportation data, traffic safety, geometric design, and public transit systems. He currently serves as a young member of the Geometric Design Committee (AFB 10) of the Transportation Research Board. Mr. Wood is a recipient of a College of Engineering Distinguished Teaching Fellowship at Penn State (2015-2016), a Dwight D. Eisenhower Graduate Fellowship (2014-2015), and was the Outstanding Student of the Year for the Mountain Plains Consortium (2013) and the Mid-Atlantic Universities Transportation Center (2016).

## **PEER REVIEW**

San José State University, of the California State University system, and the MTI Board of Trustees have agreed upon a peer review process required for all research published by MNTRC. The purpose of the review process is to ensure that the results presented are based upon a professionally acceptable research protocol.

Research projects begin with the approval of a scope of work by the sponsoring entities, with in-process reviews by the MTI Research Director and the Research Associated Policy Oversight Committee (RAPOC). Review of the draft research product is conducted by the Research Committee of the Board of Trustees and may include invited critiques from other professionals in the subject field. The review is based on the professional propriety of the research methodology.



# MTI FOUNDER

**Hon. Norman Y. Mineta**

## MTI/MNTRC BOARD OF TRUSTEES

**Founder, Honorable Norman Mineta (Ex-Officio)**  
Secretary (ret.), US Department of Transportation  
Vice Chair  
Hill & Knowlton, Inc.

**Honorary Chair, Honorable Bill Shuster (Ex-Officio)**  
Chair  
House Transportation and Infrastructure Committee  
United States House of Representatives

**Honorary Co-Chair, Honorable Peter DeFazio (Ex-Officio)**  
Vice Chair  
House Transportation and Infrastructure Committee  
United States House of Representatives

**Chair, Nuria Fernandez (TE 2017)**  
General Manager and CEO  
Valley Transportation Authority

**Vice Chair, Grace Crunican (TE 2016)**  
General Manager  
Bay Area Rapid Transit District

**Executive Director, Karen Philbrick, Ph.D.**  
Mineta Transportation Institute  
San José State University

**Joseph Boardman (Ex-Officio)**  
Chief Executive Officer  
Amtrak

**Anne Canby (TE 2017)**  
Director  
OneRail Coalition

**Donna DeMartino (TE 2018)**  
General Manager and CEO  
San Joaquin Regional Transit District

**William Dorey (TE 2017)**  
Board of Directors  
Granite Construction, Inc.

**Malcolm Dougherty (Ex-Officio)**  
Director  
California Department of Transportation

**Mortimer Downey\* (TE 2018)**  
President  
Mort Downey Consulting, LLC

**Rose Guilbault (TE 2017)**  
Board Member  
Peninsula Corridor Joint Powers Board (Caltrain)

**Ed Hamberger (Ex-Officio)**  
President/CEO  
Association of American Railroads

**Steve Heminger\* (TE 2018)**  
Executive Director  
Metropolitan Transportation Commission

**Diane Woodend Jones (TE 2016)**  
Principal and Chair of Board  
Lea+Elliot, Inc.

**Will Kempton (TE 2016)**  
Executive Director  
Transportation California

**Art Leahy (TE 2018)**  
CEO  
Metrolink

**Jean-Pierre Loubinoux (Ex-Officio)**  
Director General  
International Union of Railways (UIC)

**Michael Melaniphy (Ex-Officio)**  
President and CEO  
American Public Transportation Association (APTA)

**Abbas Mohaddes (TE 2018)**  
CEO  
The Mohaddes Group

**Jeff Morales (TE 2016)**  
CEO  
California High-Speed Rail Authority

**David Steele, Ph.D. (Ex-Officio)**  
Dean, College of Business  
San José State University

**Beverley Swaim-Staley (TE 2016)**  
President  
Union Station Redevelopment Corporation

**Michael Townes\* (TE 2017)**  
Senior Vice President  
Transit Sector, HNTB

**Bud Wright (Ex-Officio)**  
Executive Director  
American Association of State Highway and Transportation Officials (AASHTO)

**Edward Wytkind (Ex-Officio)**  
President  
Transportation Trades Dept., AFL-CIO

(TE) = Term Expiration or Ex-Officio  
\* = Past Chair, Board of Trustee

## Directors

**Karen Philbrick, Ph.D.**  
Executive Director

**Peter Haas, Ph.D.**  
Education Director

**Brian Michael Jenkins**  
National Transportation Safety and Security Center

**Hon. Rod Diridon, Sr.**  
Emeritus Executive Director

**Donna Maurillo**  
Communications Director

**Asha Weinstein Agrawal, Ph.D.**  
National Transportation Finance Center





**SAN JOSÉ STATE**  
**UNIVERSITY**

Funded by U.S. Department of  
Transportation

