# Optimization and Technology-Based Strategies to Improve Public Transit Performance Accounting for Demand Distribution 

Charalampos Sipetas<br>University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2
Part of the Transportation Engineering Commons

## Recommended Citation

Sipetas, Charalampos, "Optimization and Technology-Based Strategies to Improve Public Transit Performance Accounting for Demand Distribution" (2021). Doctoral Dissertations. 2140.
https://scholarworks.umass.edu/dissertations_2/2140

[^0]
# OPTIMIZATION AND TECHNOLOGY-BASED STRATEGIES TO IMPROVE PUBLIC TRANSIT PERFORMANCE ACCOUNTING FOR DEMAND DISTRIBUTION 

A Dissertation Presented<br>by<br>CHARALAMPOS SIPETAS

Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2021

Civil and Environmental Engineering Department
(C)Copyright by Charalampos Sipetas 2021

All Rights Reserved

A Dissertation Presented<br>by<br>CHARALAMPOS SIPETAS

Approved as to style and content by:

Eric Gonzales, Chair

Eleni Christofa, Member

Jimi Oke, Member

[^1]John E. Tobiason, Department Head Civil \& Environmental Engineering

## ACKNOWLEDGMENTS

It was back in 2016 when I decided to move to the US in order to pursue my PhD studies here. I was excited about the opportunities and also concerned about the challenges that such a decision could entail. Today, I feel so thankful for making this decision which allowed me to test my strengths and weaknesses on both a personal and professional level. A PhD is not a smooth path to walk, but with the right people by your side nothing is impossible.

First of all, I would like to thank my advisor Prof. Eric J. Gonzales, who was the person who gave me this opportunity. His guidance, support, and respect towards me are priceless and should serve as inspiration for all other teachers, mentors and/or advisors in academia. His involvement in original and innovative research allowed me to investigate fascinating research areas, while his deep knowledge in the field provided me with all the required skills to excel. His office's door was always open for me to address my questions and concerns and he could always find a solution to any of my problems. For all these (and so many other) reasons I will be forever grateful.

Regarding my committee members, Prof. Eleni Christofa was the person who introduced me to UMass and I will always be thankful for that. She has been a remarkable instructor and a great mentor whenever I needed her advice. I would like to thank her deeply for her support and her great contribution in the fulfillment of my studies. I would also like to thank Prof. Jimi Oke for being a member of my committee and offering me his valuable guidance. I so much enjoyed taking his class and I appreciated his enthusiasm during the lectures. I have a big "thank you!" for Prof. Nagurney for her kindness and generosity towards me. It was a great experience for me to attend one of her courses at Isenberg School of Management and also serve as an Officer at the Student Chapter that she advises with great success.

At this point I need to thank all the faculty members of the Civil and Environmental Engineering Department, as well as all the faculty members from other Departments with which I had the opportunity to interact. More specifically, I would like to thank Prof. Song Gao, Prof. Michael Knodler, Prof. Daiheng Ni, Prof. Chengbo Ai, Prof. Hari Balasubramanian, Prof. Senay Solak and Prof. Ahmed Ghoniem. They have all contributed greatly in my effort to acquire as much knowledge as possible and build skills that will allow me to achieve my future goals. I should also thank here all the UMass staff for the amazing job they do in helping students with any issue that might occur in different aspects of their student life.

The biggest part of my student life at UMass took place in Marston and I am so thankful to all the graduate students who were there during my studies. They were all good friends and colleagues and I am glad that our PhD seminar allowed us all to interact regularly and exchange ideas to improve our research. More specifically, I would like to thank Qing, Xinlian, Sayeeda, Tao, Hossein, Xiang, Rana, Farnoush, Foroogh, Emily, Nick, Alyssa and Francis. Special thanks go to Katerina for her constant support, as well as to Andronikos and Lancelot for sharing Marston 139A with me. Outside the Transportation program, I would like to thank deeply George, Fani, Panos, Androcles, Iro, Rodrigo and Deniz. Finally, I would like to thank my parents and my siblings for always allowing me to be myself and follow my own path. Their contribution in achieving my goals is priceless.

## ABSTRACT

# OPTIMIZATION AND TECHNOLOGY-BASED STRATEGIES TO IMPROVE PUBLIC TRANSIT PERFORMANCE ACCOUNTING FOR DEMAND DISTRIBUTION 

February 2021
Charalampos Sipetas

Diploma and M.S., University of Patras, Greece Ph.D., University of Massachusetts Amherst<br>Directed by: Dr. Eric Gonzales

Public transit is important to societies worldwide. The operation of public transit systems is generally associated with great benefits for the users, but there are also cases in which these systems demonstrate inefficient performance. Quantifying transit performance is an important area of research over the last decades. This dissertation presents models to improve transit system performance through optimization techniques and new technologies, recognizing the effects of non-uniform distribution of demand over space and time. The contributions span fixed route transit services and on-demand transit, as well as models for flexible transit operations that lie in between.

Regarding fixed route systems, a methodology is proposed to estimate the number of passengers being left-behind subway train vehicles due to overcrowding. Methods to identify appropriate time periods and locations for studying this phenomenon are presented. The effects of overcrowding on passenger waiting times are also investigated. The challenging case of transit networks where passengers tap-in only upon entrance is analyzed, adding a new methodology to a very short list of similar studies and enhancing previous work in this field.

For demand responsive systems, this dissertation focuses on optimizing the operation of paratransit services through coordination with alternative providers in order to decrease high operating costs of such a service. The analysis includes a heuristic-based method. The proposed model is more detailed than existing aggregated methods and is able to perform well in high demand levels, unlike existing exact approaches. This part of the dissertation also assists in making transportation network companies a complementary part of public transit, rather than a competitor.

Finally, flexible transit systems are studied to identify the operational and demand related characteristics of a service area that could serve as indicators of such systems' efficient performance. The focus here is on route deviation flexible services. Continuous approximation is used to model this flexible system. A new optimized hybrid transit system with elements of both fixed route and flexible services is proposed. Finally, it is highlighted that the current COVID-19 pandemic has proven the need for public transit systems that could be adjusted to accommodate changes in transit demand.

## TABLE OF CONTENTS

Acknowledgments ..... iv
Abstract ..... vi
List of Figures ..... xvii
List of Tables ..... xviii
1 INTRODUCTION ..... 1
1.1 Motivation ..... 1
1.2 Research questions ..... 4
1.3 Dissertation organization ..... 5
2 LITERATURE REVIEW ..... 6
2.1 Distribution of transit demand ..... 6
2.2 Transit performance ..... 8
2.3 Crowding on fixed route systems ..... 10
2.3.1 Crowding measures and user perception ..... 11
2.3.2 Utilization of Intelligent Transportation Systems ..... 12
2.3.3 Left behind passengers ..... 13
2.4 Operating costs of on-demand systems ..... 14
2.4.1 Early studies on DRT ..... 14
2.4.2 Recent studies on DRT ..... 15
2.4.3 Dial-a-Ride Problem ..... 15
2.4.4 The use of taxis and TNCs for people with disabilities ..... 16
2.5 Efficient implementation of flexible systems ..... 17
2.5.1 Case studies ..... 17
2.5.2 User preferences ..... 18
2.5.3 Modeling approaches ..... 19
2.6 Contributions ..... 20
2.6.1 Fixed route systems ..... 20
2.6.2 On-demand systems ..... 21
2.6.3 Flexible systems ..... 22
3 CROWDING ON FIXED ROUTE SYSTEMS ..... 24
3.1 Introduction ..... 24
3.2 Digital image processing for object detection ..... 26
3.3 Methodology ..... 28
3.3.1 Identification of study locations and times ..... 29
3.3.2 Automated dwell time estimation ..... 30
3.3.3 Automated passenger counts from video ..... 31
3.3.4 Model estimation and validation for left behind passengers ..... 31
3.3.5 Model implementation ..... 32
3.4 Study site ..... 32
3.4.1 Raw data ..... 32
3.4.1.1 Automatic Fare Collection (AFC) ..... 33
3.4.1.2 Automatic Passenger Counter (APC) ..... 34
3.4.1.3 Train Tracking Records (TTR) ..... 34
3.4.2 Models and inferred data ..... 35
3.4.2.1 Origin-Destination-Transfer model (ODX) ..... 35
3.4.2.2 Rail flow ..... 38
3.4.3 Surveillance video feeds ..... 39
3.5 Identification of study sites ..... 39
3.5.1 Crowding analysis ..... 40
3.5.2 Station geometry and camera views ..... 44
3.6 Manual data collection ..... 46
3.6.1 Train door opening and closing times ..... 47
3.6.2 Number of passengers left behind ..... 47
3.6.3 Number of passengers waiting on platform ..... 48
3.7 Automated detection of passengers on platforms in video feeds ..... 51
3.7.1 Calibration of parameters ..... 51
3.7.2 Raw video counts ..... 52
3.7.3 Accuracy of detected left behind passengers ..... 52
3.8 Modeling left-behind passengers ..... 55
3.8.1 Model estimation ..... 57
3.8.2 Model validation ..... 60
3.8.2.1 Number of passengers left behind ..... 60
3.8.2.2 Occurrence of a train leaving behind passengers ..... 62
3.8.2.3 Estimating distribution of experienced waiting times ..... 65
3.9 Summary ..... 69
4 OPERATING COSTS OF ON-DEMAND SYSTEMS ..... 71
4.1 Introduction ..... 71
4.2 Aggregate operations models ..... 73
4.2.1 Vehicle Miles Traveled (VMT) model ..... 74
4.2.2 Vehicle Hours Traveled (VHT) model ..... 75
4.2.3 Fleet size (M) ..... 76
4.2.4 Total operating and marginal cost ..... 76
4.3 Proposed algorithm ..... 77
4.3.1 Algorithm to construct representative routes ..... 78
4.3.2 Estimation of marginal cost of each trip ..... 80
4.3.2.1 Marginal cost of each paratransit trip ..... 80
4.3.2.2 Cost of taxi and TNC trip ..... 81
4.3.3 Procedure to allocate trips to paratransit or taxi/TNC ..... 83
4.4 Study site ..... 85
4.4.1 Description ..... 86
4.4.2 ADA paratransit customer data ..... 88
4.4.3 The RIDE trip records ..... 89
4.4.4 Relevant explanatory variables ..... 91
4.4.5 Observed operational outputs ..... 93
4.5 Results ..... 95
4.5.1 Aggregate model of The RIDE paratransit operations ..... 95
4.5.2 Optimized allocation of trips to paratransit and TNCs ..... 99
4.5.2.1 Marginal cost of paratransit trips ..... 100
4.5.2.2 Shifted trips characteristics ..... 102
4.6 Discussion ..... 105
4.6.1 Special equipment ..... 105
4.6.2 Taxi/TNC pricing ..... 107
4.6.3 Environmental impacts ..... 108
4.7 Summary ..... 110
5 EFFICIENT IMPLEMENTATION OF FLEXIBLE SYSTEMS ..... 113
5.1 Introduction ..... 114
5.2 System description ..... 115
5.3 Modeling route deviation ..... 118
5.4 Modeling costs ..... 120
5.4.1 Agency costs ..... 120
5.4.2 User costs ..... 121
5.4.2.1 Walking ..... 121
5.4.2.2 Waiting ..... 122
5.4.2.3 Riding ..... 122
5.4.3 Total weighted generalized costs ..... 123
5.5 Optimal spacing and flexible region ..... 124
5.6 Numerical analysis ..... 126
5.6.1 Optimal decision variables ..... 128
5.6.2 Optimized system cost components ..... 130
5.7 Comparison between fixed, hybrid and route deviation system costs ..... 130
5.7.1 Optimal percent flexibility ..... 133
5.8 Optimization of station spacing based on fixed route and route devia- tion systems ..... 135
5.9 Sensitivity analysis ..... 137
5.9.1 Effect of headway ..... 137
5.9.2 Effect of flexible service demand ..... 139
5.9.3 Effect of cost weights ..... 140
5.10 Simulation ..... 142
5.10.1 Optimal flexible region and station spacing ..... 142
5.10.2 Simulated demand ..... 143
5.10.3 Simulated scheduling ..... 143
5.10.4 Simulated costs ..... 145
5.10.4.1 Walking costs ..... 146
5.10.4.2 Waiting costs ..... 147
5.10.4.3 Riding costs ..... 148
5.10.4.4 Fleet costs ..... 148
5.10.4.5 VHT costs ..... 149
5.10.4.6 VMT costs ..... 149
5.10.5 Comparison between simulation and analytical modeling ..... 150
5.11 Summary ..... 151
6 CONCLUSIONS, TRANSFERABILITY AND FUTURE EXTEN- SIONS ..... 154
6.1 Conclusions ..... 155
6.1.1 Crowding on fixed route systems ..... 155
6.1.2 Operating costs of on-demand systems ..... 156
6.1.3 Efficient implementation of flexible systems ..... 157
6.2 Transferability ..... 158
6.3 Future extensions ..... 160
A APPENDIX ..... 163
References ..... 165

## List of Figures

3.1 Flowchart of proposed methodology ..... 29
3.2 Count of passengers a) boarding by station, and b) alighting by station, for Northbound Orange Line, 5:15-5:30pm ..... 41
3.3 Inferred passenger crowding using a) inferred passenger occupancy (Winter 2017), and b) train trajectories with inferred passenger loads (PM Peak, November 15, 2017), for Northbound Orange Line trains . ..... 45
3.4 Selected camera views from North Station, Orange Line, Northbound direction ..... 46
3.5 Regression model to estimate dwell time from Train-Tracking Records (TTR) ..... 48
3.6 Cumulative number of passengers entering the platform and boarding vehicles, North Station, November 15, 2017 ..... 50
3.7 Timeseries of passengers on platform from manual counts, North Sta- tion, November 15, 2017 ..... 50
3.8 Effect of threshold values on object detection in a sample video frame ..... 52
3.9 a) Raw (unsmoothed) and b) smoothed passenger counts from video, November 15, 2017 ..... 53
3.10 Automated passenger counts from surveillance video, North Station,
November 15, 2017 ..... 54
3.11 Linearity of explanatory variable of a) dwell time and b) video counts with respect to logit, Model 3 ..... 58
3.12 Distribution of passenger wait times considering comparison between a) expected (without left-behind passengers) and observed (with actual left-behind passengers), and b) observed and estimated (with estimated left-behind passengers) distributions of waiting times, North Station January 31, 2018 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67
4.1 Paratransit trip configuration ..... 80
4.2 Flow chart of proposed methodology ..... 85
4.3 MBTA paratransit service's monthly pick-up requests and garage lo- cations ..... 87
4.4 Percentage of total time by number of passengers onboard for MBTA paratransit service ..... 92
4.5 Network versus straight-line distance for a) Provider 1, b) Provider 2,
c) Provider 3 service area ..... 95
4.6 Marginal cost of the paratransit service with respect to demand shifted to TNC and the TNC cost for a) Provider 1, b) Provider 2, c) Provider 3 service area ..... 98
4.7 Calibrating greedy algorithm's outputs for a) VHT and b) VMT in the North Region during January 2017 ..... 99
4.8 Comparison of costs by incrementally shifting trips to TNC under dif- ferent patterns for a) Provider 1, b) Provider 2, c) Provider 3 service area ..... 101
4.9 Distribution of estimated marginal costs for all trips, North, January 23, 2017 ..... 102
4.10 Distribution of remaining ADA paratransit trips by time of day, Provider 1, January 23, 2017 ..... 103
4.11 Percent of trips shifted to TNCs per town considering a) 100 first removed trips, b) 200 first removed trips, c) 300 first removed trips, d) 400 first removed trips, and e) 500 first removed trips for Provider 1, Jan. 23, 2017
4.12 Change in cost by incrementally shifting trips with greatest net marginal cost to TNC, North, January 23, 2017

### 4.13 Change in trip allocation by increasing the value of TNC cost coefficients, North, January 23, 2017

4.14 Total VMT of serving paratransit trips in coordination with TNCs, North, January 23, 2017

### 4.15 Total VMT of serving paratransit trips in coordination with TNCs for various pick-up distance factors, North, January 23, 2017

5.1 Examples of system configuration for a) conventional fixed route, b) flexible with route deviation, c) hybrid fixed with route deviation ..... 116
5.2 Service area configuration for a) $W=1$, b) $W=2$, and c) $W=3$ ..... 129
5.3 Daily user costs of a) walking, b) waiting, and c) riding ..... 131
5.4 Daily agency costs of a) fleet size, b) VHT, and c) VMT ..... 1325.5 Percent user benefits from implementing hybrid transit instead of fixedroute for a) $Q=2.5$, b) $Q=5$, c) $Q=7.5$, and route deviation for d)$Q=2.5$, e) $Q=5$, and f) $Q=7.5$134
5.6 Optimal percent flexibility of a service area with length $\mathrm{L}=10$ (miles) and headway equal to a) 0.5 hours $/ v e h$, b) 1 hours $/ v e h$, and c) 1.5 hours/veh ..... 136
5.7 Optimized decision variable of a) station spacing, and b) flexible region for a corridor with $W=2, L=10$, and $Q=5$ ..... 137
5.8 Percent cost difference for a) $Q=2.5$, b) $Q=5$, and c) $Q=7.5$ ..... 138
5.9 Optimal decision variable of a) $S^{*}(x)$ for various headways, $H$, b) $A^{*}(x)$ for various headways, $H$, c) $S^{*}(x)$ for various percentages, $a$, d) $A^{*}(x)$ for various percentages, $a$, e) $S^{*}(x)$ for various weights, $w_{U C}$, f) $A^{*}(x)$ for various weights, $w_{U C}$. . . . . . . . . . . . . . . . . . . . . . . . . 141
5.10 Comparison between analytical model and simulation costs of a) users and b) agency . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 151

## List of Tables

> 3.1 Accuracy of video counts of left-behind passengers, North Station, $3: 30-6: 30 \mathrm{pm}$, November 15,2017 . . . . . . . . . . . . . . . . . . . . 55
3.2 Logistic regression model parameters, North Station, November 201759
3.3 Validation of probability and count of left behind passengers, January
31, 2018 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63
3.4 Validating count of left behind passengers using actual number on plat-
form, North Station, January 31, 2018 . . . . . . . . . . . . . . . . . 64
3.5 Validation of estimated occurrence of left-behinds, January 31, 2018 . 65
3.6 Comparison of distributions of passengers wait time (seconds) . . . 68
4.1 Average values of relevant explanatory variables . . . . . . . . . . . . 94
4.2 Average VMT and VHT by time of day, 2017 . . . . . . . . . . . . . 96
4.3 Modeled values of $r_{V M T}$ and $r_{V H T}$ by time of day, 2017 . . . . . . . . 97
5.1 Input values . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 127
5.2 Percent agency benefit from implementing the optimized hybrid transit
instead of Fixed Route (FR) and Route Deviation (RD) . . . . . . . 135
5.3 User and agency costs per day for different headways and percent of
demand served curb-to-curb . . . . . . . . . . . . . . . . . . . . . . . 139

## 1 INTRODUCTION

Public transit (also known as public transportation, public transport or mass transit) is a key component of societies around the world. The range of flexibility for a system's routes and schedules defines different types of public transit services. First, fixed route systems are those with the lowest flexibility, meaning that they operate on predefined routes and with specific schedules (e.g., heavy rail), without the ability to deviate. Second, flexible transit systems are the ones that allow some level of flexibility, as for example the deviation from the fixed route and/or schedule in order to serve passengers door-to-door or curb-to-curb. Third, on-demand (or demand responsive) services have the greatest level of flexibility, with all passengers being served at locations and times of their choice (e.g., paratransit services). In this dissertation, flexible and on-demand services are often mentioned with the general term "non-fixed route" systems. Designing and operating fixed and non-fixed route public transit systems depends greatly on the demand, among other factors. The non-uniform distribution of demand over space and time results in challenges that must be addressed to assure the efficient operation of these systems for both agencies and users.

This Chapter is organized as follows. Section 1.1 explains the motivation for conducting the research included in this dissertation. Section 1.2 presents the research questions that are addressed here. Finally, the organization of this dissertation is described in Section 1.3.

### 1.1 Motivation

The impact of public transit on a city's planning and development, as well as on an individual's daily activities, made it a topic of great interest over the years. The
benefits to users and non-users expand over many fields. Transit systems assure user accessibility to employment, health care and entertainment. In cases of emergency, their role has proven critical in assisting evacuation processes. Neighborhoods with well-functioning transit stations are associated with prosperous businesses and high levels of safety and security. The importance of public transit can also be highlighted by the increase in demand that results from current trends, such as urbanization, environmental concerns, aging populations, etc.

Cases in which public transit systems demonstrate inefficiencies emerge regularly and trigger questions regarding their operation and performance improvements. Regarding operation, delays and long waiting times are among the top reasons for transit user dissatisfaction. Discomfort during the trip due to aged vehicles or overcrowding is common as well. In economic terms, the efficient performance of a transit system leads to increased operating costs and thus subsidies. Demand responsive systems, and more specifically paratransit services, are often associated with very high operating costs per trip. Issues are also related with transit station accessibility. Travelers for instance, try to avoid long walking distances to access a station for various reasons. Comfort related purposes and phenomena of criminality around some transit stops are among these reasons.

Strategies to improve transit systems are consistently of interest for operating authorities. Such strategies include increasing service frequency, purchasing new vehicles and replacing the old ones, developing information systems regarding transit operations (e.g., train arrival times), and modernizing payment methods. Apart from improving equipment and infrastructure, transit authorities focus on alternative efforts such as making trips more affordable through partnering with other entities. The costs of such enhancements, though, are often high and raise the need for evidence that the expense of public transit is indeed justified in terms of satisfying the
users' needs and the operators' goals.
The importance of having healthy public transit systems is demonstrated in part by the number and diversity of people who use their services. According to the American Public Transportation Association, 9.9 billion public transit trips were completed in 2018 in the US alone. This is despite the fact that $45 \%$ of Americans do not have access to public transit, according to the same source. A survey conducted by the Pew Research Center (2015) revealed that public transit is preferred by demographic groups aged 18-29. Various studies support the need of public transit opportunities for elderly people as well (Davey, 2007). This highlights the importance of operating more flexible route transit systems for this age group.

Regarding fixed route systems, a study by Buehler and Pucher (2012) highlights a big contrast between rail and bus users in the USA. Rail passengers tend to have the highest incomes compared to other modal user groups and even higher than the national average income. Buses on the other hand are more often used by ethnic minorities and people with low income. Bullard et al. (2004) attributes such a contrast to the spatial distribution of households, where high income population is gathered in the suburbs and low income in the inner city. The purpose of the trips could also assist in portraying the users of public transit systems. According to the Federal Highway Administration (2002) work appears to be the primary destination of the travelers with a percentage of approximately $50 \%$, followed by social, church or personal business and shopping (approximately 13\%).

It is evident that public transit is used by many travelers within the US and covers a wide variety of demographic groups who travel daily to satisfy basic needs. Public transit systems are made to serve those needs, and their design and operation are strongly related with the demand distribution over space and time, in many different ways. For fixed route services, peak hours and high demand stations cause
overcrowding phenomena. As a result, waiting times increase and passengers may turn to other modes of transportation, either inside or outside the public transit umbrella. In the case of non-fixed route services, the non-uniform distribution of demand over time results in increasing the fleet size on a daily basis which in turn leads to high operating costs. On the contrary, spatial aggregation of trip requests may lead to faster service and with fewer vehicles, which is beneficial for both the demand and supply side of a transit system.

### 1.2 Research questions

Public transit affects the lives of many people on a daily basis and improving its operation could both increase user satisfaction and reduce operating costs. The general goal of this dissertation is to understand:

- What are the effects of demand distribution on transit service performance?
- What are the challenges and opportunities that different distributions of demand pose for public transit systems?

The purpose of this dissertation is to identify methods to quantify these effects, focusing on fixed and non-fixed route systems. In terms of fixed route services, subway overcrowding phenomena are studied in an effort to answer:

- What are the effects of crowding on passengers' experiences and the transit system's reliability?

Regarding non-fixed route systems, on-demand paratransit systems are studied to identify:

- How can trips be efficiently allocated to alternative services to reduce high operating costs?

In the area of non-fixed route systems, flexible transit services were also investigated in order to answer:

- What are the operational and demand-related characteristics of a service area that indicate efficient performance of such a system?

Both the demand and supply sides are studied to address these research questions and real data obtained from transit authorities are considered, when available. Existing tools and databases are utilized and no further costs are required for the developed models' proper implementation. For the cases where real data were not available, simulation techniques were used to validate all assumptions made during the model development.

### 1.3 Dissertation organization

This dissertation is organized as follows. Chapter 1 includes the motivation for this research, as well as the investigated research questions. Chapter 2 presents the literature review on the respective research fields and the contribution of this work on fixed route and non - fixed transit systems. In Chapter 3, the effects of non-uniform demand on fixed route systems are presented through the analysis of crowding phenomena of subway transit vehicles. Demand related challenges and opportunities for on-demand transit systems are included in Chapter 4. The operational and demand - related characteristics of a service area that is a good candidate for flexible systems implementation are investigated in Chapter 5. Finally, Chapter 6 presents conclusions, transferability of methods and future extensions.

## 2 LITERATURE REVIEW

Demand is distributed unevenly over space and time, which affects transit systems' operation. Section 2.1 presents literature on investigating the spatio-temporal patterns of transit users mobility in an attempt to comprehend their impact on transit performance. Performance measures developed for any type of transit system are included in Section 2.2. Focusing on fixed route systems, Section 2.3 investigates existing performance measures related to subway vehicles' crowding phenomena, which are a very common cause of transit users inconvenience. Regarding demand responsive transit systems, their high operating costs lead to investigate related studies, such as modelling and scheduling approaches, which are presented in Section 2.4. Section 2.5 describes literature on flexible services, emphasizing on different types of such services and highlighting the need for specific guidelines and optimization techniques for their implementation. Finally, Section 2.6 presents the contribution of this dissertation to the existing literature of fixed route systems, on-demand and flexible services.

### 2.1 Distribution of transit demand

Human mobility patterns have been investigated in many studies over the years in areas such as urban planning, transportation, geography and crisis management (Gong et al., 2012; Zhang et al., 2018). Multiple passengers move simultaneously in both spatial and temporal space and studying their overall behavior leads to identifying their network mobility patterns (Faroqi et al., 2019). Even though the first studies on understanding travel behavior focused on temporal aspects (Agard et al., 2006), more recent studies consider both spatial and temporal characteristics (Yu and He, 2017). Neutens et al. (2012) highlight the importance of understanding that the
transit user activities are spatially and temporally linked. In fact, measuring passenger flows in a large spatio-temporal scale is becoming more and more feasible due to emerging demand data sources (Luo et al., 2019).

According to Hasan et al. (2013), available mobility data sources include phone calls, credit card transactions, bank notes dispersal, and location detection through social network applications, among others. Research on transit users' spatio-temporal mobility is often performed through the use of smart cards in existing literature. In terms of equity, a study by Farber et al. (2016) confirms that more marginalized groups tend to travel at times of the day that are not considered as peak and are consequently poorly served. From a spatial perspective, Scott and Horner (2008) conclude that due to uneven distribution of population groups around transit stops, some of them might be favored over others. Ma et al. (2017) provide useful insights for policymakers to achieve a more balanced job-housing relationship in a given area by visualizing spatial distribution of homes and workplaces for both commuters and non-commuters. Manley et al. (2018) identify spatial and temporal clusters of travel events in order to study regularities and irregularities in travel patterns.

A significant body of research focuses on the relationship between spatio-temporal characteristics of transit demand and the surrounding land uses. Lee et al. (2013) implement aggregated stop methods to identify the relationship between land use types and demand patterns within a specified area around the transit stop. In Gong et al. (2012), the authors explain how the homogeneity and high density of land uses around a station can lead to morning and afternoon demand peaks. Similarly, Shi et al. (2018) focus on the station level and study the association between the hourly ridership and the characteristics of the built environment and topology. Yu et al. (2019) worked on cases in which the station peak hours are not completely consistent with those of the city to which they belong. Hu et al. (2016) present a general
framework of modelling transit and land use relationship that could shed light on both urban planning procedures and transit systems improvements.

Tang et al. (2018) affirm that the non-uniform spatio-temporal distribution of demand requires the implementation of proper strategies in order to maintain a transit system's operating efficiency. A preceding step that is of equal importance is relating this non-uniformity directly with transit performance and quantifying its effects. Evans and Wener (2007), for instance, investigate crowding inconveniences for transit users as a result of peak hours. Regarding non-fixed systems, Faroqi et al. (2019) emphasize on the importance of knowing the spatial and temporal distribution of demand in creating clusters that could be beneficial for group-based transit services (e.g., DRT). Thus, it is apparent that these effects can be both negative and positive.

### 2.2 Transit performance

The interest towards public systems' performance is constantly increasing due to concerns of policymakers, stakeholders and citizens on both quality and cost of public services (Stanley, 2004). The evident need for quantitatively evaluating the performance of public transit has led to the development of many performance measures (Fielding et al., 1977; Karlaftis, 2004; Lem et al., 1994; Talley, 1986). In addition to evaluating public transit in satisfying users' daily need for commuting, studies have also measured its performance in terms of sustainability (Miller et al., 2016), accessibility (Mamun et al., 2013) and emergency preparedness (Nakanishi et al., 2003), among others.

The first edition of the Transit Capacity and Quality of Service Manual (TCQSM) was published in 1999 (Kittelson \& Associates, Inc., 1999) and included six performance measures, namely service frequency, hours of service, service coverage, passenger loading, reliability, and transit vs. automobile travel time. The TCRP Guidebook
for Developing a Transit Performance-Measurement System (Transportation Research Board, 2003) includes four regularity indicators, namely headway adherence, service regularity, observed to scheduled headway ratio, and headway regularity index. In fact, according to Ruan and Lin (2009) the most commonly used metric in existing literature is related to service regularity and it is the average passenger wait time proposed by Osuna and Newell (1972).

Efficiency, effectiveness and impact were the dominating measure areas for many decades according to Phillips (2004). In this study, efficiency refers to the production of a given output using the least possible resources (e.g., labor, vehicle, maintenance). Effectiveness refers to the comparison between the intended and the actual output (e.g., utilization of service, operating safety, passenger convenience, service reliability). Regarding impact, although it partly reflects the efficiency and effectiveness, it also includes the effects of public transit on society, economy, and environment (e.g., energy consumption, user accessibility and pollution reduction). One of the first studies that supports these three areas in measuring transit performance is Gilbert and Dajani (1975). In their study, among others, the authors argue whether the transit related goals of the government should be considered when developing performance measures or not.

Eboli and Mazzulla (2012), separate transit performance measures in two broad categories, the subjective and the objective. The first refers to indicators based on passenger perception whereas the latter is expressed through numerical values of quantitative measures. Rietveld (2005) highlights the importance of studying the transit quality indicators from both the demand and supply side, since there is evidence that there are systematic differences. This approach is also followed by Trépanier et al. (2009), where the supply-based indicators include vehicle-kilometers and vehicle-hours per route per day and the demand-based indicators are a single passenger's travel
distance and time on a single run. All these indicators are then integrated into two broader ones, which are average vehicle occupancy and vehicle capacity ratio. In fact, many other studies tend to combine categories as well in order to develop integrated tools of performance measurement (Tyrinopoulos and Antoniou, 2008).

There are many methods developed in existing literature to estimate the performance of a public transit system. In terms of efficiency, Yao et al. (2019) separate these methods in two basic groups, the parametric and the non-parametric ones. The first group is primarily represented by the stochastic frontier approach, whereas the second by the data envelopment analysis. Karlaftis and Tsamboulas (2012) refer to the stochastic frontier approach, the data envelopment analysis as well as to neural networks as three basic approaches in estimating efficiency and effectiveness of transit performance. Gattoufi et al. (2004) study data envelopment analysis and concludes that its implementation is associated with significant advantages compared to other methods.

### 2.3 Crowding on fixed route systems

Crowding is a major challenge for public transit systems all over the world. It is associated with increases in waiting and travel times and decreases in operating speeds, reliability, and passenger comfort, among others (Tirachini et al., 2013). Studies show that crowding in public transit increases anxiety, stress, and feelings of invasion of privacy for passengers (Lundberg, 1976). Many recent studies are focused on understanding how overcrowding levels may affect a traveler's behaviour. Some of these use stated and/or revealed preferences data (Batarce et al., 2015; Tirachini et al., 2016), whereas others utilize available smartcard data (Kim et al., 2015). Cats et al. (2016) develop a stochastic model based on simulations and use Stockholm subway as the application network.

TCQSM (Kittelson and Associates Inc et al., 2013) determines some guidelines for measuring the quality of service to track passenger related metrics. It states that crowding affects several aspects of transit availability and all the elements of comfort and convenience related to the quality of service. The indicators of availability as presented on the TCQSM are frequency, service span, and access. The indicators of comfort and convenience as defined by TCQSM are passenger load, reliability, and travel time. Evidence has shown that the users perceive waiting and travel times to be longer in crowded conditions than in uncrowded conditions due to the additional crowding discomfort (Fan et al., 2016).

### 2.3.1 Crowding measures and user perception

Li and Hensher (2013) reviewed objective and subjective (or psychological) measures of crowding. In their study it is highlighted that a common key factor used to evaluate transit vehicle crowding in the USA is the load factor (passengers per seat), defined as the number of passengers divided by the number of seats. Another commonly used objective measure for crowding is the number of standing passengers per square meter $\left(\mathrm{m}^{2}\right)$ of a vehicle. However, unacceptable crowding levels may vary across countries and transit services. For example, transit agencies in the USA consider five standees per square meter as the limit of accepted crowding in bus services, whereas the same limit equals four in Australia and Europe (Diec et al., 2010; Furth et al., 2006). The use of density as a crowding measure, however, lacks in consideration of individual perceptions of crowding (Cox et al., 2006; Turner et al., 2004). Perceived crowding is investigated by many authors over the years (Van der Reis, 1983; Sundstrom, 1978). Batarce et al. (2016) found that the value of time of a user experiencing an overcrowded situation (equal to six standing passengers per square meters in this study) is 2.5 times larger than the respective value if there were empty
seats available.
Most of the literature on crowding has focused on passenger discomfort. It has been shown that waiting and in-vehicle travel time savings are inversely proportional to the number of people in the platform or vehicles (Douglas and Karpouzis, 2005). This is the basis for estimating the crowding externality or crowding cost (Kraus, 1991). Many studies have investigated the value of crowding from the perspective of the user, in terms of value of time and willingness to pay an extra fee to avoid crowding (Haywood and Koning, 2015; Haywood et al., 2017; Hörcher et al., 2017; Li and Hensher, 2011). Furthermore, various studies have aimed to determine the effect that crowding has on passengers' travel decisions and path choice (Raveau et al., 2014). For instance, research in Seoul, South Korea, suggests that crowding affects the path choice in networks that are large and connected enough to offer multiple path choices to users between origin-destination pairs (Kim et al., 2015).

### 2.3.2 Utilization of Intelligent Transportation Systems

In Camacho et al. (2012) the authors examine the potential of utilizing Intelligent Transportation Systems (ITS) in improving transit system reliability and user satisfaction. An example of a suggested solution to the current transit problems is the use of a system that would inform passengers for seat availability in a transit vehicle. Such an approach is presented by Zhang and Chen (2014), who develop a real-time broadcast system for crowding based on the Internet of things. Utilizing existing ITS, Nuzzolo et al. (2016) present a mesoscopic transit assignment model that could be used to predict the number of passengers on-board a transit vehicle in real time. Noursalehi et al. (2019) propose a decision support platform for realtime crowding prediction and information generation to assist passengers' decision on whether to board a train or not. The effects of real-time crowding information in
public transport networks is investigated by Drabicki et al. (2017).

### 2.3.3 Left behind passengers

When transit vehicles are overcrowded, commuters can't board the vehicle they wish. These commuters are "left-behind passengers" or "passengers denied boarding", and determining their numbers is crucial (Ma et al., 2019). Surprisingly, the existing literature on this topic is not as wide as expected. A table included in Ma et al. (2019) summarizes all of the existing papers in the area, presenting detailed comparison between developed methods. One of the existing approaches in estimating the number of left-behind passengers uses statistical techniques at a station level. Following statistical inference, Zhu et al. (2017) make use of Automatic Fare Collection (AFC) and Automatic Vehicle Location (AVL) datasets from metro systems that include smartcard tap not only in the entrance but also in the exit. Another proposed approach to estimate the number of left-behind passengers refers to the network level and proposes a network assignment method (Stasko et al., 2016).

To date, most studies that infer travel patterns of transit system users are developed using farecard data from transit systems that require passengers to tap-in and tap-out for zone-based fare collection (Ma et al., 2013). Such systems give exact information about arrivals, departures, and travel times of passengers in the system. Pelletier et al. (2011) presents the various uses of smartcard technology. Most existing studies to detect left-behind passengers have been conducted with data from the London Underground, which offers both tap-in and tap-out information (Zhu et al., 2018). Only one of the existing studies by Miller et al. (2018) proposes a method that could be utilized in transit systems where passengers tap their cards at the entrance, but not at the exit, making it hard or even impossible in some cases to identify the passenger flow within the transit network with certainty (e.g., Boston,

New York, Chicago). The method utilizes archived data and the results are promising for successful implementation in overcrowded conditions.

### 2.4 Operating costs of on-demand systems

Paratransit services are one of the many systems under the umbrella of Demand Responsive Transportation (DRT) services, which are considered a fully flexible-route transit system. Transit agencies in the United States are required to provide door-to-door paratransit service for customers with disabilities under the Americans with Disabilities Act (ADA) of 1990, which had a great impact on the operation of such systems (Quadrifoglio et al., 2008). Lewis et al. (1998) study the impact of this regulatory framework on the system's operating costs, focusing mostly on the restriction of zero denial rate from the supply side. The National Transit Summaries and Trends (NTST) of 2017 reports that the cost per passenger trip on a demand responsive system is higher than any other pubic transit mode.

### 2.4.1 Early studies on DRT

All the above highlight the need for studying DRT and understanding the fundamental elements of its operation. In fact, research on DRT systems exists in literature from the early 1970's already and it has been vivid over the last decades. The integration of DRT with traditional transit is studied by Aex (1975). Lerman and Wilson (1974) focused on predicting a DRT system's performance. The inefficient operation of conventional taxi services led to the development of more advanced DRT, such as Dial-A-Ride Transit (DART), with pre-arrangement and ride-sharing opportunities (Stein, 1978; Wilson et al., 1976). Vitt et al. (1970) used survey data and analytic techniques to identify the importance of user attributes in DRT systems. The implementation of DRT systems in various case studies is widely investigated (Carlson,

1976; Flusberg, 1976; Guenther and Authority, 1976). The importance of demand responsive systems for the private sector is investigated by Heathington et al. (1974). The results prove the economic viability of these systems and their crucial contribution to the overall public transportation system.

### 2.4.2 Recent studies on DRT

Over the next decades, research on DRT expanded and various approaches are implemented. Many studies investigate the impacts of zoning and time window strategies on DRT performance (Diana et al., 2006; Quadrifoglio et al., 2008; Rahimi and Gonzales, 2015; Shen and Quadrifoglio, 2012). Simulation-based approaches are widely applied by many authors, as for example for simulating DRT requests (Deflorio, 2011), paratransit services (Fu, 2002c), and scheduling strategies (Torkjazi and Huynh, 2019), among many other applications. A review of simulating DRT is presented by Ronald et al. (2015). Another category of existing approaches accounts for stochasticity in DRT systems (Chevrier et al., 2006; Daganzo, 1978; Fu, 1999, 2002b; Lerman and Wilson, 1974). Such methods, however, are both time and cost consuming due to the level of detail and precision that is required for their proper development. Approximate analytical models offer an alternative approach to analyze the DRT operating characteristics (Daganzo, 1978, 1984; Figliozzi, 2008, 2009).

### 2.4.3 Dial-a-Ride Problem

Scheduling a DRT system can be achieved through the implementation of Dial-ARide Problems (DARP), which consist of designing vehicle routes and schedules on a static - all requests known in advance (Desrosiers et al., 1986) or dynamic way requests occur real-time (Attanasio et al., 2004). The route scheduling of paratransit services in the past has been achieved mostly through the implementation of DARP
approaches. According to Cordeau and Laporte (2007) these models differ from others (e.g., Vehicle Routing Problem with Time Windows - VRPTW) since they account for the human perspective as well. One of the first approaches included a heuristic for multiple vehicle static DARP (Jaw et al., 1986). The heuristic selects users starting from the earliest feasible pick-up request and gradually inserts all requests into vehicle routes. In the area of multi-vehicle DARP, existing studies have considered the coordination of a regular demand responsive service with taxis, in an effort to serve all requested trips. A real-life problem concerning service of people with disabilities with taxis using a penalty cost is tackled by Toth and Vigo (1996). A detailed comparison between DARP approaches is included in Cordeau and Laporte (2003, 2007). Studies on DARP are vivid over the last decade, as well (Bongiovanni et al., 2019; Ritzinger et al., 2016; Tellez et al., 2018; Torkjazi and Huynh, 2019).

### 2.4.4 The use of taxis and TNCs for people with disabilities

The use of taxis in DRT services for people with disabilities is investigated in many studies (Burkhardt, 2010; Chia, 2008; Ellis, 2016; Tuttle and Eaton, 2012). Examples of successful collaboration among taxi and paratransit services include transit services in California, Illinois and Washington, D.C., among others (Burkhardt et al., 2008). Existing studies of paratransit operations provide modeling capabilities to quantify the effect of changes in demand, as they may result from diverting some trips to taxis or TNCs. In Rahimi and Gonzales (2015) and Amirgholy and Gonzales (2016) a quantitative basis for decision making on trip allocation is provided. Also, in Turmo et al. (2018) a study of the Pioneer Valley Transit Authority (PVTA) ADA paratransit service provided an initial analysis of the potential cost savings from coordinating with taxis or TNCs. According to Tirachini (2019), the relationship between public transport and ride hailing systems, such as TNCs, is one of the most interesting
research areas regarding the increasing use of ride hailing in general. This is mainly because TNCs can both substitute and complement public transport.

### 2.5 Efficient implementation of flexible systems

According to Mulley and Nelson (2009), the main goals of flexible systems refer to improving convenience of public transport and maintaining a comparable price to existing public transit systems. A survey by Koffman (2004) reveals that most flexible transit services are planned and designed without established guidelines. Errico et al. (2013) classified existing studies on flexible transit into two categories. The first group includes studies that describe practical experiences, whereas the second refers to methodological contributions to assist planning processes. They also concluded that there are a few cases of implementing optimization techniques for actual flexible systems, with which Potts et al. (2010) and Scott (2010) also agree. Many approaches are based on analytical modeling, considering rectilinear distances because a rectilinear movement of the vehicle is a good approximation of reality according to Dessouky et al. (2005).

### 2.5.1 Case studies

Existing literature includes flexible transit related surveys that aim at portraying the current conditions under which flexible transit services operate. According to Koffman (2004), at the time of the study development flexible transit services were implemented in more than 50 transit agencies throughout North America. Weiner (2008) complements Koffman (2004) by focusing on integrated flexible transit services that either were designed according to ADA (1990) or have proved beneficial for riders with disabilities. The report also presents and discusses cases in USA where integrated services discontinued, such as Sarasota County Area Transit, Calgary Transit, and

Access-A-Ride in New York. In Potts et al. (2010) the authors aim at providing a practical guide regarding the implementation of flexible transit services through the identification of 26 agencies as best practices for further research, including Mason County Transportation Authority and Jacksonville Transportation Authority, only to mention a few.

According to Fu (2002a), the main benefit of flexible services that involve vehicle route deviation is that they serve trips that would not be otherwise served or that would be served by a more expensive alternative, such as driving or additional fixed transit routes. A service area of width $W=2.4$ miles and length $L=10$ miles is simulated as part of their investigation for issues in designing flexible route systems, considering stochasticity in passenger demand.

### 2.5.2 User preferences

Most of the studies on public transit user preferences focus on the competition between fixed route and demand responsive services (Commins and Nolan, 2011; Hensher et al., 2013). Few studies have focused on flexible route transit more generally (Zheng et al., 2018a). In Broome et al. (2012) the authors completed a study showing the public's positive perception of flexible transit systems. In Chavis and Gayah (2017) a stated preference survey is performed to develop a mode choice model that can be used to describe how transit users select among competitive transit options. Their study covered the entire public transit spectrum, from traditional fixed route to flexible and pure on-demand services (including e-hailing such as Uber and Lyft). Although there are passengers that always choose the same mode, the results also indicated that there are statistically significant predictors of the flexible service type selected, such as monetary cost, expected in-vehicle, waiting, and walking time.

In Broome et al. (2012) the performance of a flexible route is evaluated in Hervey

Bay, Queensland, Australia. Analysis of ticket sales data showed that the replacement of the conventional fixed route by a flexible service led to approximately doubled use of the service by elderly people. The authors conclude that flexible route bus transport is a promising technology for meeting the transport needs of elderly people. Among the many types of flexible transit services, deviated fixed route services are the most widely used (Qiu et al., 2015).

### 2.5.3 Modeling approaches

In Zheng et al. (2018a) a methodology is proposed to support the decision-making process when choosing between a route deviation policy and a point deviation policy. In Nourbakhsh and Ouyang (2012) the agency and user cost components of a flexible transit system are analyzed considering idealized square cities. In Kim et al. (2019) a planning model is presented for optimizing a flexible system serving many-to-one and one-to-many demand patterns, identifying relations among optimal zone sizes, headways, and relevant exogenous factors. A study included in Pei et al. (2019) summarizes valuable findings from the existing literature on modeling approaches for flexible transit systems.

Continuous approximation methods are widely implemented in existing literature for transportation systems in general (Quadrifoglio and Li, 2009). An early study on this topic was conducted by Newell (1973). The optimized coordination between rail and bus transit services through analysis of the user and agency benefits is presented in Wirasinghe et al. (1977). A recent study by Chen et al. (2018) investigates the utilization of local route and short-turn services to complement a regular fixed route transit service by implementing a continuous approximation to model the proposed hybrid system. A detailed review of continuous approximation techniques in existing literature for transportation systems is presented in Ansari et al. (2018).

### 2.6 Contributions

### 2.6.1 Fixed route systems

The contributions of this dissertation on fixed route systems pertain to estimating the number of left behind passengers, which are a result of crowding phenomena in subway systems. The methodology proposed here addresses a more challenging case than the ones presented in Section 2.3. More specifically, the developed method accounts for transit systems where the passengers only tap-in when they enter the system and thus their movements within the transit network and their destinations can only be inferred. Existing literature includes only one study that considers this challenging case. The study developed by Miller et al. (2018) proposes a method based on statistical regression. Both AVL and AFC data are utilized and the passenger movements within the network are based on calibrations including either survey data or the origin-destination-transfer model developed by Sánchez-Martínez (2017).

The method presented in this dissertation outperforms Miller et al. (2018) for estimating low numbers of left-behind passengers. This method complements Miller et al. (2018), which is more appropriate for high levels of crowding. The two approaches could, thus, be also combined for the same transit network, depending on the fluctuations in the level of crowding within a day and across stations. The innovative study presented here is the first to implement image processing techniques for estimating the number of left behind passengers due to overcrowding. It also offers unique and valuable insights on the effect of left behind passengers on the transit reliability measure of passenger waiting times. Even in cases of low demand, as for example the current pandemic, this study provides significant guidance for modeling experienced waiting times and estimating important values that are not directly available for the transit agency (i.e., train dwell times). With fewer passengers using
the system, the methods for monitoring passenger counts on station platforms can be used to track the level of crowding, even when demand is low enough that no passengers are left behind. Finally, the overall approach adopted here highlights the potential of fusing archived transit data and object detection techniques in quantifying the effects that non-uniform spatial (i.e., crowded stations) and temporal (i.e., peak hours) distribution of demand has on the performance of a subway system.

### 2.6.2 On-demand systems

The contribution of this dissertation in the field of on-demand services refers primarily in presenting a new methodology for identifying paratransit trips that would be better served by alternative providers (e.g., taxis or TNCs) in order to reduce the high operating costs. This part of the dissertation illustrates the opportunities provided by spatially clustered demand (e.g., shorter travel times and thus lower costs of service by the transit agency) and the challenges of temporal peaks of requests (e.g., peak morning hours when the need for alternative providers is increased) in developing more efficient operating strategies for demand responsive systems. The new method proposed here is the first one to estimate the marginal cost of a paratransit trip, as a result of its relationship with the other requested trips. The great challenge of making TNCs serve as complements and not substitutes for public transport is addressed here through the proposed method for a strategic coordination of the two systems. The high significance of this study is highlighted by the major need for such quantifiable studies in this area according to recent literature (Tirachini, 2019).

Existing aggregated models in this field (Rahimi et al., 2018; Turmo et al., 2018) can answer part of the question that is set here. More specifically, they achieve to estimate the number of paratransit trips that should be assigned to an alternative provider, but they do not determine which specific trips. There are also earlier studies
in this area of research, that focus on a trip's level. In Toth and Vigo (1996), such a method is developed and tested for a real life case with a maximum of 312 requests, whereas Wong and Bell (2006) propose a model that is further tested using 150 artificial customers. The method proposed here is based on a heuristic algorithm and can be utilized for service areas with much greater demand. More specifically, the developed algorithm has been tested on a dataset with more than 3,000 trip requests. In fact, according to Toth and Vigo (1997), exact approaches for the solution of reallife transport of people with disabilities are not practicable and the authors propose the use of heuristics.

### 2.6.3 Flexible systems

This dissertation contributes greatly in the existing literature of flexible services through the development of a hybrid transit system, with elements of both fixed route and flexible route deviation services. This system is optimized to minimize the total generalized costs, considering the size of the flexible region within the service area where vehicles are able to deviate to pick up passengers on request and the spacing between fixed stops as the decision variables. According to existing studies there are a few cases of implementing optimization techniques for actual flexible systems (Errico et al., 2013; Potts et al., 2010; Scott, 2010), and this study aims at providing transit agencies with valuable quantifiable tools to support the design and operation of such optimized systems.

Most of the existing studies in this field focus on the analysis of a specific type of flexible service or the comparison between different systems. This study is the first to introduce a hybrid service model that allows the degree of flexibility to be optimized continuously for every location along a transit route. The model combines elements of two different transit systems, considering optimization techniques based on continu-
ous approximation approach. The result is a unique model that calibrates the degree of flexibility to the characteristics of the region and demand served. On one extreme, the lowest flexibility converges to a conventional fixed-route service. On the other extreme, a fully flexible service would allow deviation to serve passengers anywhere in corridor. A study that implements similar methods to optimize the coordination between rail and bus transit systems was developed from the early 1970s (Wirasinghe et al., 1977). That study, however, focuses on optimizing the coordination of different types of fixed route transit services and also considers different service area configuration than the ones considered here. The same holds for more recent studies, such as Chen et al. (2018). The method presented here considers the case where the level of flexibility for a service area is determined by the proposed models, leading often to a hybrid operation where the same fleet of vehicles serves passengers at fixed stops and at curb-to-curb locations within an optimized flexible region.

Finally, this part of the dissertation offers major insights on what are the operational features and the spatio-temporal characteristics of demand that determine whether a service area would be better served as a conventional fixed route system, as a fully flexible service, or as an intermediate system. Such insights are essential, since the existing literature includes various examples of unsuccessful implementation of flexible services in practice. For example, Weiner (2008) reports cases where integrated services were discontinued, such as Sarasota County Area Transit (SCAT), Calgary Transit, and Access-A-Ride in NY, among others. The implementation of the methodology proposed here could reveal the level of flexibility that is required in serving a given area. If this level is very low, the flexible service and the respective losses for both the agencies and the users could be avoided.

## 3 CROWDING ON FIXED ROUTE SYSTEMS

The focus of this Chapter is to identify methods to quantify the effects of nonuniform spatio-temporal distribution of demand on fixed route systems as a result of crowding phenomena. The proposed methodology addresses the challenging case of systems where passengers only tap-in when they enter the system and their movements within the transit network can only be inferred. To perform this study both archived data and video image processing techniques are required. The results indicate that this methodology is particularly valuable for detecting the number of left behind passengers, and its performance is greater in low crowding levels.

This Chapter is organized as follows. Section 3.1 presents an introduction to the research topic. Section 3.2 includes a literature review on existing image processing studies for object detection. The methodology developed to achieve the goals of this study is explained in Section 3.3. The case study considered here is described in Section 3.4. The identification of study sites is presented in Section 3.5. In Section 3.6, the process and results of manual data collection are explained. The content of Section 3.7 refers to the automated detection of passengers waiting on the platform. Section 3.8 includes the logistic regression models that were developed to estimate the number of left behind passengers fusing archived and real-time data. The summary of this research is presented in Section 3.9.

### 3.1 Introduction

Peak hours and highly utilized stations often lead to vehicle crowding on subway systems, which results in passengers not being able to board the first arriving train. These passengers are called left behind passengers and their number can affect significantly the transit system's measures of reliability, even though it is not often
taken into consideration. The proposed methodology is a technique for estimating the number of left behind passengers at a station and involves multiple steps, referring either to the system or the station level. The case study considered here refers to the Boston's subway operated by the Massachusetts Bay Transportation Authority (MBTA), where the system's reliability is measured through the percentage of passengers experiencing waiting times longer than a headway. The lower this percentage, the greater the system's reliability.

Existing rail data sources include AFC, AVL, the inferred model of origin-destina-tion-transfer (ODX) and the Rail Flow tool (based on ODX), which are utilized to identify stations and time periods with the highest probability of detecting left behind passengers. ODX could be the main data source for a study like this, but it is based on the fundamental assumption that everyone is able to board the first departing train. Although its utilization remains valuable for addressing crowding related issues, the quantification of left behind passengers requires further analysis.

The only existing study to the date that can be implemented in transit systems with only tap-in upon entrance to estimate left behind passengers is a recent study by Miller et al. (2018). The authors use AFC and AVL data in order to define a measure of cumulative transit vehicle capacity shortage. This measure is proven to be correlated with the number of left behind passengers. Manual counts from video feeds are used to calibrate the model. The results indicate that the model performs better in very crowded conditions. The results presented in this Chapter prove that the study proposed here can complement the study by Miller et al. (2018) depending on the level of crowding at a given station.

The methods and results presented here offer insights to the potential utilization of existing rail data sources and emerging technologies, such as object detection tools, in measuring the number of left behind passengers. An existing body of research
investigates algorithms for tracking pedestrian movements within transit stations and a number of studies have developed image processing tools to track pedestrians in video footage (Li et al., 2014; Mukherjee et al., 2011; Ozer and Wolf, 2014; Yan-yan et al., 2014). The object detection tool adopted here is called You Only Look Once (YOLO) and is described in the following Section.

### 3.2 Digital image processing for object detection

There are a number of technologies that can be used to observe, count, and track pedestrians and pedestrian movements in an area. Digital image processing for object detection is an appealing approach for transit systems because surveillance videos are already being recorded in transit stations for safety and security purposes. The video feed records passenger positions and movements in the same way that a person would observe them, as opposed to infrared or wireless signal detectors that merely detect the movement of a person passed a point or their proximity to a detector. The detection of objects in surveillance videos is an invaluable tool for passenger counting and has numerous applications. For example, object detection can be used for passenger counting or tracking, recognizing crowding, and hazardous object recognition. In a relevant application, Velastin et al. (2006) uses image processing techniques to detect potentially dangerous situations in railway systems. Computer vision is the duplicate of human vision aiming to electronically perceive, understand and store information extracted from one or more images (Sonka et al., 2014).

There are various techniques to use computers to process an image for object detection by extracting useful information. Recent methods use feature-based techniques rather than segmentation of a moving foreground from a static background, which was used in the past. Then, the detected features are extracted and classified, typically using either boosted classifiers or Support Vector Machine (SVM) methods
(Cheng et al., 2015; Viola, 1993). SVM is one of the most popular methods used in object detection algorithms and especially passenger counting, because it offers a method to estimate a hyperplane that splits feature vectors extracted from pedestrians and other samples (Cheng et al., 2015), differentiating pedestrians from other unwanted features. Boosting uses a sequence of algorithms to weight weak classifiers and combine them to form a strong hypothesis when training the algorithm to attain accurate detection (Zhou, 2012). Current methods for object detection take a classifier for an object and evaluate it at several locations and scales in a test image, which is time-consuming and creates numerous computational instabilities at large scales (Deng et al., 2010).

The most recent methods, such as Region Based Convolutional Neural Network (R-CNN), use another method to decrease the region over which the classifier runs and includes the SVM. First, category-independent regions are proposed to generate potential bounding boxes. Second, the classifier runs and extracts a fixed-length feature vector for each of the proposed regions. Finally, the bounding boxes are refined by the elimination of duplicate detections and rescoring the boxes based on other objects on the scene using SVMs (Girshick et al., 2014). The bounding box is a rectangular box located around the objects in order to represent their detection (Coniglio et al., 2017; Lézoray and Grady, 2012). The resulting object detection datasets are images with tags used to classify different categories (Deng et al., 2009; Everingham et al., 2010).

The open-source software tool called YOLO uses a different method than the above-mentioned techniques for object detection. It generates a single regression problem to estimate bounding box coordinates and class probabilities simultaneously by using a single convolutional network that predicts multiple bounding boxes and class probabilities for these boxes (Redmon et al., 2016). Another advantage of YOLO
is that, unlike other techniques such as SVMs, it sees the entire image globally instead of sections of the image. This feature enables YOLO to implicitly transform contextual information to the code about classes and their appearance and at the same time makes YOLO more accurate, making fewer than half the number of errors compared to Fast R-CNN (Redmon et al., 2016). YOLO uses parameters for object detection that are acquired from a training dataset. YOLO can learn and detect generalizable representations of objects, outperforming other detection methods, including R-CNN. The ability to train YOLO on images has the potential to directly optimize the detection performance and increase the bounding box probabilities (Redmon et al., 2016).

The calibration of parameters for object detection using an algorithm like YOLO requires training datasets with a large number of tagged images. Although a custom training set that is specific to the context of application (e.g., MBTA transit stations) would be desirable for achieving the most accurate object detection outcomes, it is very costly to create a large tagged training set from scratch. The Common Objects in Context (COCO) dataset is a large-scale object detection, segmentation, and captioning dataset that is freely available to provide default parameter values for YOLO. The COCO dataset is not specific to passengers or transit stations, but it is a general dataset that includes 328,000 images, 2.5 million tagged objects and 91 object types, including "person" (Lin et al., 2014). Nevertheless, the tool is effective for identifying individual people in camera feeds, and the use of general training data allows the same tool to be applied in other contexts without requiring additional training data.

### 3.3 Methodology

The proposed methodology aims to estimate the number of left behind passengers at a transit station when trains are too crowded to board. Figure 3.1 presents a flowchart of the data and methods used in this study in order to provide a roadmap
for the analysis described in this study. The methods rely heavily on two data sources that are automatically collected and recorded (shown in blue): train tracking records that indicate train locations over time, and surveillance video feeds. Additional archived data on inferred travel patterns from farecard records is used only to identify the most crowded parts of the system (shown in purple), and manual counts are used to estimate and validate models (shown in red). For model implementation, the proposed models require only the automatically collected input data.


Figure 3.1: Flowchart of proposed methodology

### 3.3.1 Identification of study locations and times

The first step of the analysis presented in this study is to identify the stations and times of day when crowding is most likely to cause passengers to be left behind
on the platform. This analysis is used only for determining where to collect data to demonstrate the implementation of the proposed model. This step could be skipped for cases in which the locations for implementation are already known.

The identification of study sites involves a crowding analysis that makes use of two data sources: train tracking records, which denote the locations of trains over time; and inferred passenger flows. As discussed later in this study, the study site is MBTA where passenger flows are inferred through ODX model using passenger farecard data. Other similar models can be equally applied, if available by the transit authority that implements this methodology. Peaks in train occupancy and numbers of boarding passengers show where and when passengers are most likely to be left behind, as described in Section 3.5.1. Then, Section 3.5.2 describes an analysis of surveillance camera views to determine which stations have unobstructed platform views and station geometry that allows the automated video analysis techniques to be used to count passengers.

### 3.3.2 Automated dwell time estimation

Train tracking data, which includes the time each train enters a track circuit, is automatically recorded in transit networks, including the MBTA Research Database. By comparing this data against manual observations of the times that train doors open and close in the station, a linear regression model is estimated to predict dwell time from the train tracking records, as described in Section 3.6.1. This model is used to obtain automated dwell time estimates as inputs to the model of left behind passengers.

### 3.3.3 Automated passenger counts from video

Automated counts of the number of passengers on each station platform are obtained using YOLO, an automated image detection algorithm. The parameters of the algorithm are associated with the freely-available COCO training dataset, as described in Section 3.2. The threshold for object identification is calibrated, as described in Section 3.7.1, by applying the algorithm to the surveillance video feed and comparing with manual counts of the passengers remaining on the platform after the doors have closed (Section 3.6.2) and the passengers entering and exiting the platform (Section 3.6.3). With the parameter values and calibrated threshold, YOLO produces estimates of the number of passengers on the platform as a time series. The number of passengers that remain on the platform after the doors close is a raw automated passenger count, as shown in Section 3.7.2. These raw counts are not very accurate as a direct measure (Section 3.7.3), but they provide a useful input for modeling the number of left behind passengers.

### 3.3.4 Model estimation and validation for left behind passengers

A logistic regression is used to predict the probability that a passenger is left behind on the station platform based on automated dwell time estimates and/or automated passenger counts from video. The model parameters are estimated using the manually observed counts of passengers left behind on the station platforms as the observed outcome. The diagnostics, parameters, and fit statistics of the models developed in this dissertation are presented in Section 3.8.1. The explanatory variables in the study presented here are automated dwell time estimates and automated passengers counts. The quality of the proposed models is evaluated through validation against manually collected counts on a different day than the one used for model
estimation. The accuracy of the model predictions is then calculated relative to manually observed passenger counts on the same day as the one used for prediction, as shown in Section 3.8.2.

### 3.3.5 Model implementation

Implementation of the model to make ongoing estimates of the numbers of passengers left behind each departing train requires only train tracking data and surveillance video feeds as model inputs. The manual observations of door opening/closing times and the number of passengers on the platforms are used only for estimating model parameters. The models then produce predictions of the number of passengers left behind each departing train based only on data that is automatically collected. Therefore, the numbers of left behind passengers and the associated impact on the distribution of wait times experienced by passengers could be tracked as a performance measure over time. If data feeds were processed as they are recorded, it would also be possible to implement the models to make real-time predictions of the left behind passengers.

### 3.4 Study site

### 3.4.1 Raw data

The case study considered here is the MBTA subway system, where there are three main sources of raw data related to passenger and vehicle movements, as described in the following sections.

### 3.4.1.1 Automatic Fare Collection (AFC)

Automatic fare collection data is collected from the fare collection system at station fare gates and on-board buses and light rail vehicles. The AFC records are associated with events in which Charlie Cards (MBTA's farecard) are used to load value, pay a fare, or validate a pass. The data is partitioned by month and year, and includes records of Charlie Card transactions from individual fare cards as well as passes. Relevant AFC data that could potentially be useful for assessing crowding are:

- Unique identifier of the device that records the AFC event
- The station location of the device (e.g., fare gate, firebox, or ticket vending machine) that recorded the event
- The timestamp of the event
- Card/ticket serial number from the AFC system
- Type of transaction (e.g., top-up, validation, or fare deduction)

From this raw data, counts of passengers entering transit stations can be tracked over time based on the transactions' times and locations. The dataset includes good coverage of passengers entering fare gate-controlled stations on the red, orange, and blue lines. However, passengers are able to board inbound green line trains without necessarily validating a ticket, so some passengers are able to enter the system and make transfers without being counted.

The MBTA's rapid transit fare system charges a single fare for entry to the system, and passengers do not tap out when they leave the system. As a result, AFC records
only account for station and vehicle entry, and there are no direct observations of exits.

### 3.4.1.2 Automatic Passenger Counter (APC)

Automatic passenger counters (APC) are devices that count the number of passengers boarding and alighting each vehicle. APC devices are not in widespread deployment on MBTA rail vehicles, so this is not a data source that can be reliably used for assessing crowding in the system.

### 3.4.1.3 Train Tracking Records (TTR)

The train-tracking system records the position of heavy rail vehicles as they move from track circuit to track circuit through the system. The analogous data for tracking bus positions on the network are reported through the AVL systems. Since much of the heavy rail operations are in tunnels, track circuits are used to identify train locations. There is typically one track circuit associated with each station, and a few circuits between consecutive stations. The relevant TTR data for this study are:

- The timestamp of the train-tracking record
- Numeric code for heavy rail line
- Letter code for heavy rail line
- Numeric code identifying a trainset
- Latitude associated with track circuit
- Longitude associated with track circuit
- Unique identification number for track circuit
- The direction of train traffic on the track circuit
- The location type of the track circuit
- Name of station associated with the track circuit

The AVL data provides detailed data about vehicle movements in the system that can be compared against passenger data from the AFC data. From the AVL data, it is possible to piece together the progression of an individual vehicle along a line. It is also possible to look at the headways of departures from a specific station.

### 3.4.2 Models and inferred data

The raw data collected and logged by the MBTA contains extensive (although not complete) information about passenger entrances to rail stations and boarding buses at bus stops. It also contains comprehensive records of vehicle movements. By itself, this data is sufficient to count passenger entries and track performance of transit vehicles for schedule or headway adherence. In order to assess crowding, additional processing of the data is necessary to link records and infer travel patterns.

### 3.4.2.1 Origin-Destination-Transfer model (ODX)

A model to link trip records and infer origin-destination and transfer patterns in the system has been developed to populate a database of ODX records. Inference models based on farecard data have been improved over the years. The most recent advances make use of dynamic programming to minimize generalized disutility for travelers, accounting for path-specific waiting time, in-vehicle time, and transfers (Sánchez-Martínez, 2017). The model identifies records from AFC that can be linked
to infer transfers or return trip patterns. For example, a passenger using a Charlie Card to enter a rail station and later board a bus near a different rail station can be assumed to have used the rail system and then transferred to the bus. Another passenger who enters one rail station in the morning and enters a different rail station in the afternoon may be completing a round-trip commute, so the destination of the morning and afternoon trips can be inferred by linking the two trips. Through this method, the model infers values for $97 \%$ of trip origins, $75 \%$ of trip destinations, and $92 \%$ of transfers.

The ODX model is structured in three levels:

1. Ride - One ride; boarding and alighting one vehicle
2. Stage - One fare card tap; this could be a single ride, boarding a bus and riding to a destination stop to alight. This could also be a station entry that is followed by a ride on a train and then a gateless transfer to another train
3. Journey - One trip from origin to destination; this may consist of one or more rides and stages. For example, a multi-stage journey could include a first stage consisting of a ride on a bus and then a second stage consisting of entry to a rail station. The stages are each recorded by a separate tap (on the bus and at the fare gate), but a transfer from one mode or route to another may be required to complete a trip.

The ODX records are based on the raw data from AFC and AVL, but the dataset contains information related to journeys by inferring the destination and transfer locations and times associated with each origin. The relevant data from the ODX records are:

- Serial number of card, or arbitrary assigned number for cash transactions
- Location of the stop or station where fare transaction was recorded
- Timestamp of fare transaction
- The sequence of the journey for a specific card
- Sequence of the stage within the journey
- Total number of stages within the journey
- Recorded or inferred journey origin location
- Inferred journey destination location
- Timestamp when the stage starts, based on vehicle's departure time from origin stop
- Timestamp when the stage ends, based on vehicle's arrival at destination stop
- Timestamp when the stage ends, based on vehicle's arrival at destination stop
- The route of the vehicle trip or the route of the station where the fare card was tapped
- The direction of the vehicle trip
- Code indicating if the origin was inferred, or the reason it was not inferred
- Code indicating if the destination was inferred, or the reason it was not inferred
- Code indicating if a transfer was inferred, or the reason it was not inferred
- The given or inferred origin of a ride, usually a bus stop or station platform
- The time at which the vehicle departed from the ride's origin
- The inferred destination of a ride, usually a bus stop or station platform
- The time at which the vehicle arrived at this ride's destination

The ODX data provides a comprehensive and useful view of travel patterns in the MBTA system. Although it appears at a glance to provide the same information as records from a tap-in and tap-out AFC would provide, it is important to be mindful of the assumptions on which inferences are based. Notably, for this study, inferred stages are based on the assumption that passengers are always able to board the next arriving vehicle. Therefore, destination times provide an optimistic estimate, assuming that crowding did not prevent a passenger from boarding the next arriving vehicle.

### 3.4.2.2 Rail flow

The Rail Flow tool provides processed and aggregated data based on the ODX records. This data includes estimates of passenger boardings and alightings at stations for 15 minute increments. In this way, the ODX model provides valuable data for estimating the level of crowding in the system. The tool shows the variability of passenger flows between stations and provides an indication of locations and times that are likely to be experiencing the greatest crowding. However, Rail Flow does not provide an indication of left-behind passengers, because the ODX data is built on the assumption that passengers are not left behind.

Perhaps a subsampling of stage data could be extracted to consider only multistage journeys in which the start time of the second stage can be used to work backward to estimate when the previous stage likely ended. Comparing the estimate of stage end time to the passage of vehicles may provide a rough estimate of whether
or not a passenger was left behind. This would not provide a comprehensive measure of the left-behinds problem.

### 3.4.3 Surveillance video feeds

Stations throughout the MBTA are equipped with surveillance cameras for security purposes. The placement of cameras has been designed to provide coverage for security purposes, and the view angles are not necessarily optimized for counting passengers on platforms. Variations in station architecture (e.g., side platforms vs. island platforms, columned stations with low station ceilings vs. open vaulted ceilings) create many different contexts for video observation. A challenge is that columns and curvature in the station limit how much of the platform, where passengers may be walking or waiting, is visible in a single frame. The extensive placement of cameras, especially in recently renovated stations, provides multiple vantage points to observe platform crowding and vehicle boarding.

### 3.5 Identification of study sites

To test the implementation of object detection with video in transit stations, a first step is to identify locations and times to collect video feeds as well as direct manual observations of left-behind passengers. For this study, stations were selected based on a crowding analysis and evaluation of station geometry and camera view characteristics. The goal was to identify stations with the greatest likelihood of passengers being left behind during a typical morning or afternoon rush and where object detection techniques would be most successful. The analysis focused on the Orange Line, which is 11-miles long with 20 stations. Oak Grove and Forest Hills are the northern and southern end stations, respectively. There are two main reasons for choosing this specific line. First and most important, it has no branch lines, so all travelers
can reach their destination by boarding the next available train, which simplifies the identification of left-behind passengers. Second, it passes through several transfer stations in the center of Boston, which highlights its significance for passengers' daily commuting.

### 3.5.1 Crowding analysis

A crowding analysis is a necessary step to identify the times and stations where crowding is observed and left behinds have the highest probability of occurring. The data used in this part of the analysis have been extracted from the Rail Flow database in the MBTA Research and Analytics Platform. The Rail Flow dataset includes aggregated boarding and alighting counts by time of day with 15 -minute temporal resolution averaged across all days in a calendar quarter. An example is given in Figure 3.2 for $5: 15-5: 30 \mathrm{pm}$ in Winter 2017. These data are derived from the ODX model, which makes use of AFC and AVL systems to infer the flow of passengers within the subway (Sánchez-Martínez, 2017) and is described in Section 3.4.2.1.

For the crowding analysis in this study, cumulative counts of passengers boarding and alighting at each station have been created along the direction of train travel using the aggregated railflow data. For a 15 -minute time period, $B(n, t)$ is the cumulative count of all passengers that board trains in the direction of interest at stations preceding and including station $n$ during time interval $t$. Similarly, $A(n, t)$, is the cumulative count of passengers that are assumed to have exited trains traveling in the direction of interest at stations preceding and including station $n$ during time interval $t$. It should always be true that $A(n, t) \leq B(n, t)$, because passengers can only alight a train after boarding it.

The difference between the cumulative boardings, $B(n, t)$, and alightings, $A(n, t)$, is the estimated passenger flow, $Q(n, t)$, between station $n$ and $n+1$ during each


Figure 3.2: Count of passengers a) boarding by station, and b) alighting by station, for Northbound Orange Line, 5:15-5:30pm

15-minute time period.

$$
\begin{equation*}
Q(n, t)=B(n, t)-A(n, t) \tag{1}
\end{equation*}
$$

This calculation is approximate, because cumulative counts are calculated for a single 15-minute time period, and real trains take more than 15 minutes to traverse the length of a line.

To calculate the number of passengers per train, the passenger flow per time
period must be converted to passenger occupancy, $O(n, t)$ (passengers/train), which is calculated by multiplying the passenger flow by the scheduled headway of trains, $h(t)$ (minutes), at time $t$.

$$
\begin{equation*}
O(n, t)=Q(n, t) \times \frac{h(t)}{15} \tag{2}
\end{equation*}
$$

The headway is divided by 15 minutes to account for the fact that the passenger flow is per 15 -minute time period. This measure is an approximation of the number of passengers onboard each train that is based on the assumptions that headways are uniform and passengers are always able to board the next arriving train. In reality, variations in headways may lead to increased crowding after longer headways, increasing the likelihood that some passengers will be left behind.

The 2017 MBTA Service Delivery Policy (SDP) (MBTA, 2017) provides guidelines for reliability and vehicle loads. In the 2010 MBTA SDP (MBTA, 2010), the maximum vehicle load was explicitly defined as $225 \%$ of seating capacity in the peak hours (start of service to $9: 00 \mathrm{am} ; 1: 30 \mathrm{pm}-6: 30 \mathrm{pm}$ ) and $140 \%$ of the seating capacity in other hours. The 2017 SDP notes that accurately monitoring the passenger occupancy of heavy rail transit is not yet feasible on the MBTA system. Nevertheless, the guidelines from Table B2 in the 2017 SDP are used to identify general crowding levels, recognizing that each Orange Line train is six cars long and has a total of 348 seats.

A visualization of average train occupancy for the Winter 2017 Rail Flow data is shown in the color plot in Figure 3.3a. The color for each station and 15 -minute time interval corresponds to the value of $O(n, t)$. Since the trains have 348 seats, red parts of the plot indicate large numbers of standing passengers, with dark red indicating crowding near vehicle capacity. This figure shows that in the northbound direction, the most severe crowding occurs between Downtown Crossing and North

Station shortly before $6: 00 \mathrm{pm}$. Note that the crowding appears to decrease before rebounding again at $6: 30 \mathrm{pm}$. This is due to the change in scheduled headway at 6:30 pm from 6 minutes to 10 minutes, which increases occupancy, as calculated in equation (2).

A more detailed visualization combines transit vehicle location records and inferred origin-destination trip flows from a specific date. As mentioned already, the ODX trip flows are constructed with simplifying assumptions about passenger movements; for example, all passengers entering a station are assumed to board the first arriving train. Despite such assumptions, however, the model is valuable for many applications. The trajectories in Figure 3.3b are associated with the recorded arrival and departure times of train at each station. The colors are associated with the estimated train occupancy based on the inferred boardings and alightings, assuming that no passengers are left behind. The trajectory plot shows that the headway between trains can vary substantially, especially for the stations north of Downtown Crossing. Longer headways are followed by more crowded trains, because more passengers have arrived to board since the previous train. The occurrence of left-behind passengers would make actual train occupancies slightly lower for the trains following long headways. Those left-behind passengers would then be waiting to board the next train, thereby increasing the occupancy on one or more subsequent trains.

Tracking the average number of passengers onboard trains provides an indicator for the likelihood of passengers being left behind, because full trains leave little room for additional passengers to board. During the most crowded times of the day, it is also useful to look at the numbers of passengers boarding and alighting trains at each station. Passengers are most likely to be left behind at stations where trains arrive with high occupancy, few passengers alight, and many more passengers wait to board. By this measure, North Station in the afternoon peak appears to be an
ideal candidate for observing left behind passengers. Using the same method for the southbound direction, Sullivan Square station was identified as an ideal candidate location for data collection in the morning peak. Other candidate stations include Back Bay, Chinatown and Wellington stations.

### 3.5.2 Station geometry and camera views

In addition to identifying stations with the greatest likelihood of passengers getting left behind crowded trains, the stations that are selected for detailed analysis should also have characteristics that are amenable to successful testing of video surveillance counting methods. There are a variety of station layouts and architectures that contribute complicating factors to the analysis of left behind passengers, and the goal of this study is to identify the potential for the adopted detection method under the best possible conditions. Ideal conditions for the proposed analysis are:

- Dedicated Platform for Line and Direction of Interest - In this case, all passengers on a platform are waiting for the same train, so any passenger that does not board can be counted as being left behind. In the case of an island platform, observed passengers may be waiting for trains arriving on either track. In the MBTA system, more than half of the station platforms for heavy rail rapid transit in the city center (the most crowded part of the system) meet this criterion. ${ }^{1}$
- High Quality Camera Views - Surveillance cameras vary in age, quality, and placement throughout the MBTA system. Newer cameras have higher definition video feeds. The quality of the view is also affected by lighting conditions,

[^2]
(a)

(b)

Figure 3.3: Inferred passenger crowding using a) inferred passenger occupancy (Winter 2017), and b) train trajectories with inferred passenger loads (PM Peak, November 15, 2017), for Northbound Orange Line trains
especially at above-ground station where sunlight and shadows can affect the clarity of the images.

- Platform Coverage of Camera Views - The surveillance systems are designed to provide views of the entire platform area for security purposes. In some stations, the locations of columns obfuscate the views, requiring more cameras to provide this coverage.

Surveillance camera views were considered from five stations on the Orange Line (Back Bay, Chinatown, North Station, Sullivan Square, and Wellington) that were identified through crowding analysis as candidate stations. Ultimately, North Station was selected as the study site for the northbound direction afternoon peak period because the station exhibits consistent crowding and the geometry provided good camera views. Samples of the camera views from this station are shown in Figure 3.4.


Figure 3.4: Selected camera views from North Station, Orange Line, Northbound direction

### 3.6 Manual data collection

Manual observations on the platform needed to be collected to establish a ground truth against which to compare alternative methods for measuring and estimating the number of passengers left behind crowded trains. Detailed data collection at North Station was conducted during afternoon peak hours (3:30-6:30 pm) on midweek days during non-holiday weeks (Wednesday, November 15, 2017, and Wednesday, January

31, 2018). Three observers worked simultaneously on the station platform to record observations.

### 3.6.1 Train door opening and closing times

Although Train-Tracking Records (TTR) report the times that each train enters the track circuit associated with a station, there is no automated record of the precise times that doors open and close. Since passengers can only board and alight trains while the doors are open, recording these times manually is important for identifying when passengers board trains, when they are left behind, the precise dwell time in the station, and the precise headway between trains. Each of the three observers recorded the times of doors opening and closing. The average of these observations is considered the true value.

A simple linear regression model shows that observed dwell times (time from doors opening to doors closing) can be accurately estimated from automatic records of TTR arrival and departure times associated with each station. Figure 3.5 shows the data and regression results combining manual counts for November 15, 2017 and January 31, 2018. There is no systematic difference between records from different days, and the $R^{2}$ is greater than 0.9 , indicating a good fit.

### 3.6.2 Number of passengers left behind

Each observer counted the number of passengers left behind on the station platforms after the train doors closed. In order to avoid double-counting, each observer was responsible for observing passengers in a two-car segment of the six-car train (front, middle, and back). Some judgement was necessary in determining which passengers to count, because some passengers linger on the platform after alighting the train and some choose to wait for a later train even when there is clearly space avail-


Figure 3.5: Regression model to estimate dwell time from Train-Tracking Records (TTR)
able to board. The goal of the left-behind passenger count is to measure the number of passengers that are left behind due to crowding within $\pm 2$ passengers of the true number.

### 3.6.3 Number of passengers waiting on platform

In addition to counting the number of passengers left behind by crowded trains, it is important for model calibration to get an accurate count of the number of passengers waiting to board each arriving train. Given the large number of commuters using the heavy rail system during commuting hours, it is not possible to accurately count this total number of passengers in person.

Surveillance video feeds of escalators, stairs, and elevators used to access the platform of interest were used to manually count the number of passengers entering and exiting the platform offline. Specifically, an open-source software tool was used to track passenger movements by logging keystrokes to the video timestamp during playback (Campbell, 2012). Counts were conducted by watching the surveillance video playback of each entry and exit point from the platform and logging the entry
and exit of each individual passenger. The resulting data log records the time (to the nearest second) that each passenger entered and exited the platform. Since the platforms of interest serve only one train line in one direction, all entering passengers are assumed to wait to board the next train, and all exiting passengers are assumed to have alighted the previous train. Combining these counts with the direct observations of the number of passengers left behind each time the doors close provides an accurate estimate of the number of passengers that were successfully able to board each train. Figure 3.6 illustrates the cumulative numbers of passengers entering the platform (blue curve) and boarding the trains (orange curve). The steps in the orange curve correspond to the times that the train doors close. If passengers are assumed to arrive onto the platform and board trains in first-in-first-out (FIFO) order, the red arrow represents the waiting time that is experienced by the respective passenger, which is estimated as the difference between the arrival and the boarding time.

A timeseries of the actual number of passengers waiting on the platform is constructed by counting the cumulative arrivals of passengers to the platform over time and assuming that all passengers board departing trains except those that are observed to be left behind. This ground truth for data collected on November 15, 2017, is shown in blue in Figure 3.7. The sawtooth pattern shows the growing number of passengers on the platform as time elapses from the previous train. The drops correspond to the times when doors close. At these times, the platform count usually drops to zero. When passengers are left behind, the timeseries drops to the number of left behind passengers. One such case is illustrated with the red arrow in Figure 3.7.


Figure 3.6: Cumulative number of passengers entering the platform and boarding vehicles, North Station, November 15, 2017


Figure 3.7: Timeseries of passengers on platform from manual counts, North Station, November 15, 2017

### 3.7 Automated detection of passengers on platforms in video feeds

This part of the study was performed by the research team of Dr. Eric J. Gonzales and is not one of the author's individual accomplishments. However, the outputs obtained here are used as inputs for the remaining of the analysis for the estimation of left behind passengers, thus, a brief presentation of the implemented methodology and the main results are described in this section. For more details about the automated object detection methods adopted here, the readers are referred to Gonzales et al. (2018) and Sipetas et al. (2020).

### 3.7.1 Calibration of parameters

The YOLO algorithm uses pattern recognition to identify objects in an image. A threshold for certainty can be calibrated to adjust the number of identified objects in a specific frame. If the threshold is set too high, the algorithm will fail to recognize some objects that do not adequately match the training dataset. If the threshold is set too low, the algorithm will falsely identify objects that are not really present. In order to identify the optimal threshold, frames from 14 camera views were analyzed. Each frame was analyzed separately for threshold values ranging from $6 \%$ to $25 \%$ to determine the optimal threshold value in relation to a manual count of passengers visible in the frame. The optimal threshold across all camera views was found equal to $7 \%$. Figure 3.8 shows the identified objects at each threshold level for the same frame from camera installed in North Station.


Figure 3.8: Effect of threshold values on object detection in a sample video frame

### 3.7.2 Raw video counts

The output from YOLO is a text file that lists the objects detected for each frame and the bounding box for the object within the image. A time series count of passengers on the platform is simply the number of "person" objects identified in the corresponding frames from each sample video feed. Figure 3.9a shows the raw passenger counts on the platform at North Station for the time period from 5:00 $6: 30 \mathrm{pm}$ on November 15, 2017. Although there are noisy fluctuations, there is a clear pattern of increasing passenger counts until door opening times (green). To facilitate analysis of the automatic passenger counts from the surveillance videos, it is useful to work with a smoothed time series of passenger counts, as shown in Figure 3.9b.

### 3.7.3 Accuracy of detected left behind passengers

The smoothed video counts from the three surveillance camera feeds used to monitor the northbound Orange Line platform at North Station are shown as the green curve in Figure 3.10. The automated passenger counting algorithm clearly undercounts the total number of passengers on the platform. The reason for this large discrepancy is that the algorithm can only identify people in the foreground of the

(b)

Figure 3.9: a) Raw (unsmoothed) and b) smoothed passenger counts from video, November 15, 2017
images, where each person is large. Therefore, the available camera views do not actually provide complete coverage of the platform for automated counting purposes. Furthermore, when conditions get very crowded, it becomes more difficult to identify separate bodies within the large mass of people.

The problem of undercounting aside, it is clear that the automated counts gener-
ate a pattern that is representative of the total number of passengers on the platform. Using regression, the smoothed timeseries can be linearly transformed into a scaled timeseries (the orange curve in Figure 3.10), which minimizes the squared error compared with the manually counted timeseries. Using this scaling method, the data from November 15, 2017, were used to compare estimated counts of left-behind passengers in the peak periods with the directly observed values. This provides a measure of the accuracy of automated video counts. The total number of left-behind passengers estimated by this method is presented in Table 3.1, where the Root Mean Squared Error (RMSE) is calculated by comparing the number of passengers left-behind each time the train doors close.

The scaling process, which makes the blue and orange curves in Figure 3.10 match as closely as possible, results in substantially overcounted left behinds, because the scaling factor tends to over-inflate the counts when there are few passengers on the platform. As a direct measurement method, automated video counting is not sat-


Figure 3.10: Automated passenger counts from surveillance video, North Station, November 15, 2017

Table 3.1: Accuracy of video counts of left-behind passengers, North Station, 3:306:30pm, November 15, 2017

|  | Left-Behind Passengers | RMSE |
| :---: | :---: | :---: |
| Manual Observation | 198 |  |
| Unscaled Video Count | 73 | 16.7 |
| Scaled Video Count | 336 | 11.9 |

isfactory, at least as implemented with YOLO. However, Figure 3.10 shows a clear relationship between the video counts and passengers being left behind on station platforms, so there is potential to use the video feed as an explanatory variable in a model to estimate the likelihood of passengers being unable to board a train.

### 3.8 Modeling left-behind passengers

In order to improve the accuracy of estimates of the number of passengers left behind on subway platforms, a logistic regression model is formulated to estimate the probability that each passenger is left behind based on explanatory variables that can be collected automatically. A logistic regression is used to estimate the number of passengers left behind by way of estimating the probability that each waiting passenger is left behind, because the logistic function has properties that are more amenable to this application. Since passengers are only left behind when platforms and trains are very crowded, a linear regression has tendency to provide many negative estimates of left behind passengers, which are physically impossible. The binary logit model, by contrast is intended for estimating the probability that one of two possible outcomes is realized (e.g., a passenger is either left behind or not left behind). The estimated probability from a logit model is always between 0 and 1 , so the resulting estimate of the number of left-behind passengers is always non-negative and cannot exceed the total number of waiting passengers.

For estimation of the logistic regression, each passenger is represented as a separate observation, and all passengers waiting for the same departing train are associated with the same set of explanatory variables. Over the course of a 3-hour rush period, there are typically about 30 trains serving North Station, serving 1,500 to 3,000 passengers per period, and leaving behind well over 100 passengers. Logistic regression models are generally expected to give stable estimates when the data set for fitting includes at least 10 observations for each outcome, so there is sufficient data to estimate parameters for a model that is structured this way.

The logistic function defines the probability that a passenger is left behind by

$$
\begin{equation*}
P(x)=\frac{1}{1+e^{-\left(\beta_{0}+\boldsymbol{\beta} \boldsymbol{x}\right)}} \tag{3}
\end{equation*}
$$

where $\boldsymbol{x}$ is a vector of explanatory variables, $\boldsymbol{\beta}$ is a vector of estimated coefficients for the explanatory variables, and $\beta_{0}$ is an estimated alternative-specific constant. The estimation of the model can be thought of as identifying the values of $\beta_{0}$ and $\boldsymbol{\beta}$ that best fit the observed outcomes

$$
y=\left\{\begin{array}{l}
1, \beta_{0}+\boldsymbol{\beta} \boldsymbol{x}+\epsilon>0  \tag{4}\\
0, \text { else }
\end{array}\right.
$$

where $y=1$ corresponds to a passenger being left behind, and $y=0$ corresponds to a passenger successfully boarding.

The underlying assumption in this formulation is that the likelihood of being left behind can be expressed in terms of a linear combination of explanatory variables and a random error term, $\epsilon$, which is logistically distributed. The explanatory variables that are considered in this study are as follows:

1. Dwell time (time from door opening to door closing) or difference of TTR arrival
and departure times
2. Video count of passengers on platform following doors closing

These explanatory variables can all be monitored automatically, without manual observations. Video counts of passengers on the platform following doors closing are obtained from the object detection process described above. Although dwell time is an appropriate explanatory variable because doors stay open longer when trains are crowded, the dwell time is not directly reported in archived databases. As demonstrated in Figure 3.5, observed dwell times can be accurately estimated from automatic records of TTR arrival and departure times. This leads to using TTR reported values of difference between train arrival and departure instead of dwell times for the model development. Since these are essentially the same explanatory variable, we call this difference "dwell time" for the remainder of the study.

### 3.8.1 Model estimation

Initially, three models were estimated, making use of only TTR data (Model 1), only video counts (Model 2), and then fused TTR and video counts (Model 3). The data from November 15, 2017, were used to develop these models. The number of passengers waiting on the platform (as described in section 3.6.3) are used to determine the number of observations for estimating the parameters of the logit model. In total, 2167 passengers boarded arriving trains at North Station during the rush period and 198 of them were left behind. This leads to a sample size of 2365 passengers for the logistic models.

Models 1 and 2 are simple logistic regressions, each with only one independent variable. Neither model has influential values (i.e., values that, if removed, would improve the fit of the model). Model 3 uses both TTR data and video counts, so it
is important to diagnose the model's fit, especially with respect to the assumptions of the logistics regression. First, multicollinearity of explanatory variables should be low. The correlation between dwell time and video count is -0.643 and the variance inflation factor is 1.7 , both indicating that the magnitude of multicollinearity is not too high. Second, no influential values were identified. Third, the logistic regression is based on the assumption that there is a linear relationship between each explanatory variable and the logit of the response, $\log (p /(1-p))$, where $p$ represents the probabilities of the response. Figure 3.11 shows that dwell time is approximately linear with the logit response, while there is somewhat more variability with respect to the video counts. Neither plot suggests that there is a systematic mis-specification of the model.

(a)

(b)

Figure 3.11: Linearity of explanatory variable of a) dwell time and b) video counts with respect to logit, Model 3

A summary of the estimated model coefficients and fit statistics is presented in Table 3.2. The log likelihood is a measure of how well the estimated probability of

Table 3.2: Logistic regression model parameters, North Station, November 2017

| Parameter | Model 1 |  | Model 2 |  | Model 3 |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
|  | Value | p-stat | Value | p-stat | Value | p-stat |
| Constant for Left-Behind | -10 | 0.00 | -4.18 | 0.00 | -7.57 | 0.00 |
| Dwell Time (sec) | 0.0903 | 0.00 |  |  | 0.0487 | 0.00 |
| Video Count |  |  | 0.370 | 0.00 | 0.232 | 0.00 |
| Null Log Likelihood, $L L_{0}$ | -1639.29 |  | -1639.29 |  | -1639.29 |  |
| Model Log Likelihood, $L L$ | -551.94 |  | -533.28 |  | -514.43 |  |
| $\rho^{2}$ | 0.663 |  | 0.675 |  | 0.686 |  |
| AIC | 1107.9 |  | 1070.6 |  | 1034.9 |  |

a passenger being left behind matches the observations. The null log likelihood is associated with no model at all (every passenger is assigned a $50 \%$ chance of being left behind), and values closer to zero indicate a better fit. The $\rho^{2}$ value is a related measure of model fit, with values closer to 1 indicating a better model.

For all three models, the estimated coefficients have the expected signs and magnitudes. The positive coefficients for dwell time and video counts indicate a positive relationship with the probability of having left-behind passengers, which is intuitive. In order to compare models, the likelihood ratio statistic is used to determine whether the improvement of one model is statistically significant compared to another. The likelihood ratio test statistic is calculated by comparing the log likelihood of the restricted model (with fewer explanatory variables) to the unrestricted model (with more explanatory variables):

$$
\begin{equation*}
D=2\left(L L_{\text {unrestricted }}-L L_{\text {restricted }}\right) \tag{5}
\end{equation*}
$$

Comparing Model 1 (restricted) to Model 3 (unrestricted), one additional variable in Model 3, indicates one degree of freedom, which requires $D>3.84$ to reject the null hypothesis at the 0.05 significance level. Comparison between Models 1 and 3 gives $D=75.02$, indicating that Model 3 provides a significant improvement over Model

1 by adding video counts. Comparison between Models 1 and 2 gives $D=37.7$, which is also a significant improvement. The Akaike Information Criterion (AIC) is an additional model fit statistic that weighs the log likelihood against the complexity of the model. Although Model 3 has more parameters, the AIC is greater than for Model 1 or Model 2, indicating that the improved log likelihood justifies the inclusion of both TTR and video count data.

### 3.8.2 Model validation

### 3.8.2.1 Number of passengers left behind

The logistic regression provides an estimate of the probability that passengers are left behind each time the train doors close. In order to translate this probability into a passenger count, the estimated number of passengers waiting on the platform from the scaled video count is used as an estimate of the number of passengers waiting to board. Table 3.3 shows the validation results when the models were applied to data collected on January 31, 2018, for North Station. The scaling factor used for the number of passengers waiting on the platform is estimated from November 15, 2017 data. Considering the estimated number of left behind passengers for each train separately, it is observed that these models achieve higher accuracy when there are a few passengers left behind. Overall, Model 1 exhibits error of only $3.3 \%$ since it estimates that 116 passengers are left behind in total when 120 passengers were observed to be left behind. Model 3 gives a lower estimate of 100 passengers being left behind, which leads to an error of approximately $17 \%$.

As shown in Table 3.1 and Table 3.3, direct video counts (unscaled and scaled) do not provide accurate estimates of the total numbers of passengers left behind without some additional modeling. The unscaled video counts underestimate the total, while
the scaled video counts overestimate the total. The logistic regression provides much better results. Although there are some discrepancies for specific train departures, the estimated numbers of passengers left behind are not significantly biased and the total number of passengers left behind during the three-hour rush period is similar to the manually counted total.

The logistic regressions estimate the probability of a passenger being left behind using only the explanatory variables listed in Table 3.2. However, the estimated number of left behind passengers is calculated by multiplying the probability by the scaled video count of passengers on the platform at the time the doors opened, as estimated from the TTR data. Therefore, the estimated number of passengers left behind with Model 1 and Model 3 rely only on TTR data that is currently being logged and supplemented by automated counts of passengers in existing surveillance video feeds. The models therefore utilize explanatory variables that are monitored automatically, and they can be deployed for continuous tracking of left behind passengers without needing additional manual counts.

The logistic models could actually perform even better if there were a way to obtain a more accurate count of the number of passengers waiting for a train. During the morning peak period, the count of farecards entering outlying stations can provide a good estimate for the number of passengers waiting to board each inbound train. This is more challenging at a transfer station, like North Station, in which many passengers are transferring from other lines. In some cases, strategically placed passenger counters could provide useful data. Nevertheless, Table 3.4 presents the performance of the developed logistic regression models if their estimated probabilities are multiplied by the actual number of passengers on the platform instead of the estimated number as in Table 3.3. This reveals the value of more accurate data, because Model 3 decreases its error compared to Table 3.3. Model 3 in Table 3.4
estimates 122 passengers being left behind in the afternoon rush on the observed date when the previous estimate was 100, which is a reduction of error from $17 \%$ to $2 \%$ for this model compared to the 120 observed left behind passengers.

### 3.8.2.2 Occurrence of a train leaving behind passengers

Another way to evaluate the performance of the developed models is to consider whether or not trains that leave behind passengers can be distinguished from trains that allow all passengers to board. Through the course of data collection and analysis, the number of passengers being left behind because of overcrowding can only be reliably observed within approximately $\pm 2$ passengers. The reason for this is that sometimes people choose not to board a train for reasons other than crowding, and one or two passengers left on the platform did not appear to be consistent with problematic crowding conditions.

If a train is defined to be leaving behind passengers when more than 2 passengers are left behind, the results presented in Table 3.3 can be reinterpreted to evaluate each method by four measures:

1. Number of Trains Leaving Behind Passengers: The number of trains in a time period that leave behind passengers due to overcrowding.
2. Correct Identification Rate: The percent of trains that are correctly classified as leaving behind passengers or not leaving behind passengers, as compared to the manual count. This value should be as close to 1 as possible.
3. Detection Rate: The percent of departing trains that were manually observed to leave behind passengers that are also flagged as such by the estimation method. This value should be as close to 1 as possible.

Table 3.3: Validation of probability and count of left behind passengers, January 31, 2018

| Train | Manual Count |  | Unscaled | Scaled | Model 1 |  |  | Model 3 |  |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: |
|  | Prob. | Count | Vid.Count | Vid.Count | Prob. | Count | Prob. | Count |  |
| 1 | $6.5 \%$ | 2 | 1 | 7 | $3.2 \%$ | 2 | $2.2 \%$ | 2 |  |
| 2 | $0.0 \%$ | 0 | 2 | 15 | $1.0 \%$ | 0 | $1.5 \%$ | 0 |  |
| 3 | $0.0 \%$ | 0 | 2 | 15 | $0.4 \%$ | 0 | $0.9 \%$ | 0 |  |
| 4 | $0.0 \%$ | 0 | 1 | 7 | $0.8 \%$ | 0 | $1.1 \%$ | 1 |  |
| 5 | $0.0 \%$ | 0 | 0 | 0 | $0.3 \%$ | 0 | $0.5 \%$ | 0 |  |
| 6 | $0.0 \%$ | 0 | 0 | 0 | $0.7 \%$ | 0 | $0.8 \%$ | 0 |  |
| 7 | $0.0 \%$ | 0 | 0 | 0 | $0.6 \%$ | 0 | $0.7 \%$ | 0 |  |
| 8 | $0.0 \%$ | 0 | 2 | 15 | $1.2 \%$ | 1 | $1.7 \%$ | 1 |  |
| 9 | $0.0 \%$ | 0 | 2 | 15 | $2.1 \%$ | 1 | $2.2 \%$ | 1 |  |
| 10 | $0.0 \%$ | 0 | 1 | 7 | $0.6 \%$ | 0 | $0.9 \%$ | 0 |  |
| 11 | $15.2 \%$ | 23 | 2 | 15 | $18.2 \%$ | 27 | $7.4 \%$ | 11 |  |
| 12 | $0.0 \%$ | 0 | 1 | 7 | $1.2 \%$ | 1 | $1.3 \%$ | 1 |  |
| 13 | $0.0 \%$ | 0 | 1 | 7 | $2.1 \%$ | 1 | $1.8 \%$ | 1 |  |
| 14 | $0.0 \%$ | 0 | 2 | 15 | $0.8 \%$ | 0 | $1.3 \%$ | 0 |  |
| 15 | $14.4 \%$ | 24 | 3 | 22 | $15.6 \%$ | 19 | $8.4 \%$ | 10 |  |
| 16 | $5.8 \%$ | 5 | 4 | 30 | $7.0 \%$ | 6 | $6.6 \%$ | 6 |  |
| 17 | $14.6 \%$ | 19 | 3 | 22 | $19.5 \%$ | 24 | $9.6 \%$ | 12 |  |
| 18 | $10.6 \%$ | 14 | 10 | 77 | $18.2 \%$ | 20 | $33.9 \%$ | 37 |  |
| 19 | $8.7 \%$ | 9 | 3 | 22 | $6.4 \%$ | 3 | $5.1 \%$ | 2 |  |
| 20 | $2.4 \%$ | 1 | 4 | 30 | $0.9 \%$ | 0 | $2.3 \%$ | 1 |  |
| 21 | $3.5 \%$ | 4 | 3 | 22 | $3.8 \%$ | 3 | $3.8 \%$ | 3 |  |
| 22 | $3.0 \%$ | 3 | 2 | 15 | $1.7 \%$ | 1 | $2.0 \%$ | 2 |  |
| 23 | $0.0 \%$ | 0 | 1 | 7 | $0.8 \%$ | 0 | $1.0 \%$ | 0 |  |
| 24 | $0.0 \%$ | 0 | 3 | 22 | $0.8 \%$ | 0 | $1.6 \%$ | 1 |  |
| 25 | $0.0 \%$ | 0 | 3 | 22 | $1.1 \%$ | 1 | $2.0 \%$ | 2 |  |
| 26 | $2.7 \%$ | 2 | 2 | 15 | $2.7 \%$ | 1 | $2.5 \%$ | 1 |  |
| 27 | $6.7 \%$ | 7 | 2 | 15 | $3.2 \%$ | 3 | $2.8 \%$ | 2 |  |
| 28 | $3.6 \%$ | 2 | 3 | 22 | $0.5 \%$ | 0 | $1.3 \%$ | 1 |  |
| 29 | $3.6 \%$ | 2 | 1 | 7 | $0.8 \%$ | 0 | $1.0 \%$ | 0 |  |
| 30 | $2.7 \%$ | 3 | 2 | 15 | $2.5 \%$ | 2 | $2.4 \%$ | 2 |  |
| Total |  | 120 | 66 | 490 |  | 116 |  | 100 |  |
| MAE |  |  | 3.5 | 12.9 |  | 1.5 |  | 2.8 |  |
| RMSE |  |  | 6.5 | 17.6 |  | 2.4 |  | 5.8 |  |
|  |  |  |  |  |  |  |  |  |  |

Table 3.4: Validating count of left behind passengers using actual number on platform, North Station, January 31, 2018

|  | Measured | Model 1 | Model 2 | Model 3 |
| :---: | :---: | :---: | :---: | :---: |
| Number of left behind passengers | 120 | 137 | 102 | 122 |
| MAE |  | 1.5 | 4.1 | 2.8 |
| RMSE |  | 2.6 | 8.6 | 6.5 |

4. False Detection Rate: The percent of departing trains that are estimated to leave behind passengers but have not, according to manual observations. This value should be as close to 0 as possible.

There is an important distinction to make here, because there are two ways that the model to identify trains leaving behind passengers can be used:

1. to estimate the number of trains that leave behind passengers, in which case we only care about measure 1 ; or
2. to identify which specific trains are leaving behind passengers, in which case measures 2 through 4 are important.

Depending on how the data will be used, application (1) or (2) may be more relevant. For example, application (1) provides an aggregate measure of the number of trains leaving behind passengers. Application (2), on the other hand, is what would be needed to get toward a real-time system for identifying (even predicting) left-behind passengers.

A comparison of the four measures is presented in Table 3.5 for the 30 trains that departed North Station between 3:30pm and 6:30pm on January 31, 2018. Unscaled video counts provide a good estimate of the number of trains that leave behind passengers (measure 1), but suffer from a low detection rate and high false detection rate. Scaled video counts are poor estimators for the occurrence of left-behind passengers

Table 3.5: Validation of estimated occurrence of left-behinds, January 31, 2018

|  | Man.Count | Unscaled Vid. | Scaled Vid. | M1 | M3 |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Total Departing Trains | 30 | 30 | 30 | 30 | 30 |
| Trains with Left-Beh.Pax | 10 | 10 | 27 | 8 | 6 |
| Correct Ident. Rate |  | 0.77 | 0.37 | 0.90 | 0.93 |
| Detection Rate |  | 0.25 | 1 | 0.80 | 0.80 |
| False Alarm Rate |  | 0.33 | 0.70 | 0.05 | 0 |

because they are high enough to trigger too many false detections. The modeled estimates both perform well in approaching the actual number of trains leaving behind passengers. Model 3 has the best performance for measures 2 through 4. It never falsely identifies a train as leaving behind passengers, and it correctly detects most occurrences of passengers being left behind. Like the count estimates above, both Model 1 and Model 3 rely on the scaled video counts to estimate the number of passengers waiting on the platform when the train doors open, so a fusion of TTR records and automated video counts provide the most reliable measures.

### 3.8.2.3 Estimating distribution of experienced waiting times

Another application of the model is to consider the distribution of waiting times implied by the estimated probabilities that passengers are left behind each departing train. From the direct manual counts, a cumulative count of passengers arriving onto the platform and of passengers boarding trains provides a timeseries count of the number of passengers on the platform. If passengers are assumed to board trains in the same order that they enter the platform, the system follows a first-in-first-out (FIFO) queue discipline. Although it is certainly not true that passengers follow FIFO order in all cases, this assumption allows the cumulative count curves to be converted into estimated waiting times for each individual passenger. The FIFO assumption
yields the minimum possible waiting time that each passenger could experience, and the waiting time for each passenger can be represented graphically by the horizontal distance between the cumulative number of passengers entering the platform and boarding trains (see Figure 3.6 for data from November 15, 2017). The yellow curve in Figure 3.12a represents the cumulative distribution of waiting times that are implied by the observed numbers of passengers entering the platform if all passengers on the platform are assumed to be able to board the next departing train. We call this the expected waiting time. The blue curve in Figure 3.12a is the cumulative distribution of waiting times if the number of left-behind passengers are accounted for when trains are too crowded to board. We call this the observed waiting time, because it reflects direct observation of passengers waiting on the platform using manual counts. The distribution indicates the percentage of passengers that wait less than the published headway for a train departure, which is the reliability metric used by the MBTA. For the Orange Line during peak hours, the published headway is 6 minutes ( 360 seconds). Currently, the MBTA is only able to track the expected wait time as a performance metric. The difference between the yellow and blue curves indicates that failing to account for left-behind passengers leads to overestimation of the reliability of the system.

The models developed in this study provide the estimated probability that a passenger is left behind each time the train doors close. In the absence of additional passenger count data, a constant arrival rate is assumed over the course of the rush period, the door closing times from TTR and the probability of passengers being left behind from Model 3 can be used to estimate the cumulative passenger boardings onto trains over time. Under the same FIFO assumptions described above, the distribution of experienced waiting times can be estimated based on train-tracking and video counts. By this process a cumulative distribution of waiting times is estimated


Figure 3.12: Distribution of passenger wait times considering comparison between a) expected (without left-behind passengers) and observed (with actual left-behind passengers), and b) observed and estimated (with estimated left-behind passengers) distributions of waiting times, North Station January 31, 2018

Table 3.6: Comparison of distributions of passengers wait time (seconds)

| Cumulative Probability | Observed | Expected | Abs.Error | Modeled | Abs. Error |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0.1 | 44 | 35 | 9 | 49 | 5 |
| 0.2 | 81 | 69 | 12 | 87 | 6 |
| 0.3 | 118 | 103 | 15 | 124 | 6 |
| 0.4 | 162 | 140 | 22 | 162 | 0 |
| 0.5 | 207 | 188 | 19 | 202 | 5 |
| 0.6 | 254 | 236 | 18 | 246 | 8 |
| 0.7 | 294 | 281 | 13 | 302 | 8 |
| 0.8 | 369 | 344 | 25 | 369 | 0 |
| 0.9 | 457 | 444 | 13 | 466 | 9 |
| 1 | 693 | 693 | 0 | 718 | 25 |
| Total |  |  | 146 |  | 72 |

using probabilities from Model 3 is shown as a red curve in Figure 3.12b, which we call the uniform arrivals modeled wait time. Table 3.6 includes the values of experienced waiting times for the observed, the expected, and the modeled distributions. This table also shows how the accuracy of estimating waiting times can be improved if we consider the actual arrival rate under the same assumptions used to develop the uniform arrivals modeled wait time. We call this distribution the actual arrivals modeled wait time. The Earth Mover's Distance (EMD) is used to measure the difference between the observed distribution and the expected, uniform arrivals and actual arrivals modeled distributions (Rubner et al., 2000). As shown in Table 3.6, the EMD for the expected case is much higher than the EMD for the modeled cases, which indicates that the proposed model reduces errors.

The modeled distributions of waiting times closely approximate the observed distribution. This suggests that the estimated probabilities of passengers being left behind each departing train are consistent with the overall passenger experience. The percentage of passengers experiencing waiting times lower or equal to the 6 minute published headway is $79 \%$ for both the observed and uniform arrivals model curve,
and $77 \%$ for the actual arrivals model curve. The automated count of left behind passengers provides a close approximation of the actual service reliability when applied to the independent data collected on January 31, 2018. The expected distribution, which does not account for left-behind passengers produces an estimate of $81 \%$ of passengers waiting less than 6 minutes. The expected distribution overestimates the reliability of the system by failing to account for the waiting time that left-behind passengers experience.

### 3.9 Summary

This Chapter investigates the potential for measuring the number of left-behind passengers using existing data sources and automated passenger counts derived from existing surveillance video feeds. The analysis of automated passenger counts is based on the implementation of a fast, open-source algorithm called You Only Look Once (YOLO) using existing training sets that identify people as well as other objects. The performance is fast enough that frames from surveillance video feeds could potentially be analyzed in real time.

Following a preliminary analysis of crowding conditions on the MBTA's Orange Line, data collection and analysis focus specifically on northbound trains at North Station during the afternoon peak hours. Data was collected on two typical weekdays and confirmed that overcrowding is a common problem, even on days without disruptions to service. This is an indication that the system is operating very near capacity, and even small fluctuations in headways lead to overcrowded trains that result in left-behind passengers.

Although video counts were not accurate in isolation, the development of models to use automated video counts with automated train-tracking records (Model 3) demonstrate good results for different applications. In predicting the number of trains
leaving behind passengers, the developed models can correctly identify whether or not passengers were left behind for $93 \%$ of the trains. The number of passengers that are left behind during the afternoon rush period can be estimated within $17 \%$ of their actual number using only automated video counts and automatically collected train tracking records. With actual counts of the numbers of passengers on the station platform at each train arrival the model can predict the number of left behind passengers with $2 \%$ of the actual number. Furthermore, the modeled distribution of experienced waiting times reduces the total EMD error by more than $50 \%$ compared to the error of the operator's expected distribution, where left-behind passengers are not considered. This highlights the need of accounting for left-behind passengers when tracking the system's reliability metrics.

## 4 OPERATING COSTS OF ON-DEMAND SYSTEMS

The focus of this Chapter is on non-fixed transit systems, and more specifically on demand responsive paratransit systems, in order to identify challenges and opportunities resulting from the non-uniform distribution of requests. High operating costs of paratransit services, highlight the need for research on identifying strategies for their proper reduction. The research in this chapter is focused on developing a fast and efficient method to identify whether there are trips that should be better served by a TNC and, if yes, which are these trips exactly. The dataset used in this chapter derives from the Massachusetts Bay Transportation Authority (MBTA) paratransit service for the year 2017.

This Chapter is organized as follows. Section 4.1 presents an introduction to the research topic. Section 4.2 describes existing aggregated models that can answer how many paratransit trips, but not which trips specifically, should be better served by TNCs. Section 4.3 includes the methodology that is proposed here in order to fill the aforementioned gap in existing literature. The dataset used to implement the proposed methodology is described in Section 4.4. Section 4.5 presents the results from implementing the existing aggregated models, as well as the proposed methodology using the given dataset. Special considerations related to this research are discussed in Section 4.6. Finally, the summary of this part of the dissertation is included in Section 4.7.

### 4.1 Introduction

The purpose of ADA paratransit is to provide service that complements conventional fixed-route transit for people who are unable to use conventional buses,
subways, or trolleys. According to existing literature presented in Section 2.4, rising ridership with ADA paratransit services poses a challenge due to the high costs of operation and transit agencies are seeking ways to reorganize operations and form partnerships with alternative providers in order to contain costs while meeting rising needs. Taxi companies have provided services to transit agencies under partnering agreements for many years. In similar lines, partnering with TNCs is a potential strategy investigated by transit agencies in an attempt to reduce the high operating costs of the paratransit service. For example, the National Transit Database (NTD) estimates the average cost of a paratrasit trip in the greater Boston area as high as $\$ 52.13$ for 2017. This further enhances the crucial importance of developing methodologies to quantify the value of public transit and TNCs collaboration.

Existing studies have focused on this or relevant topics, implementing different approaches. The aggregated approaches (Rahimi et al., 2018; Turmo et al., 2018) achieve to answer how many paratransit trips should be allocated to taxis or TNCs, but are not detailed enough to determine which trips specifically. More exact methods, as for example Toth and Vigo (1996) and Wong and Bell (2006), can perform well in lower demand levels than the ones often met in big urban centers (e.g. Boston). The use of a heuristic approach in this study allows the implementation of the method to a dataset of more than 3,000 requested trips on a daily basis. The use of such approaches in this research area is supported by Toth and Vigo (1997).

The dataset utilized here comes from the MBTA's ADA paratransit service called "The RIDE", where a Pilot Program allows eligible riders to make subsidized trips with ridesharing companies (Uber, Lyft, and Curb). The dataset includes detailed records for the trips implemented during the year 2017, which we used for developing our proposed method. This study initially investigates the application of existing aggregated models in determining the shift of paratransit trips to TNCs. The
methodology proposed here achieves a more detailed approach in answering which trips specifically should be better served by TNCs. An important aspect for making such decision, is to identify how well a requested trip fits with the other trips in the service area. Thus, our study proposes a more detailed methodology, that accounts for individual trips' characteristics and most importantly for their spatio - temporal relationship with the other requested trips through the quantification of their marginal costs of service.

Finally, it needs to be highlighted that even though existing studies and public debates prove that partnerships between transit agencies and TNCs have the potential to provide large cost savings, we acknowledge that there are several critical challenges that must be considered and addressed including legal requirements for ADA service providers and the potential of inequitable provision of service. This study focuses only on the technical challenge of determining the most cost effective way to organize these arrangements.

### 4.2 Aggregate operations models

Models of aggregated VMT, VHT, and fleet size are based on geometric probability and the resources needed to serve a density of demand over each operator's service area. The models are of the form introduced in Daganzo (1978) and Rahimi et al. (2018). These models are based on simplifying assumptions about the distribution of demand in each service regions and the operating algorithm for serving requested trips. What the model lacks in detail and realism, it makes up for in providing an analytical formula that physically relates explanatory factors to operational outcomes. This approach is valuable, because only two parameters (one for the VMT model and another for the VHT and fleet model) must be calibrated to fit the data. All the other variables are measurable quantities.

The aggregate model builds on the basic operating assumptions for a dial-a-ride system presented in Daganzo (1978). Demand is uniformly distributed within a roughly circular region with area $A$, and conditions do not change significantly within an analysis time period. For this study, we break each day into time periods of length $t_{p}$, within which the demand rate, $\lambda$, and network traffic speed, $v$, are assumed to be constant. At any time, all the demand within a pick-up window of duration $T$ are potential customers to pick-up. Each vehicle is assumed to operate by first picking up the nearest waiting customers until the target vehicle occupancy, $n$, is reached. Then, the vehicle alternates between dropping off the on-board customer with the nearest destination and picking up the next nearest waiting customer. In this way, the number of passengers on-board the vehicle is maintained at a near constant level, and the vehicle is approximately minimizing distance and time traveled by always proceeding to the next nearest stop.

### 4.2.1 Vehicle Miles Traveled (VMT) model

The total VMT operated within a time period is the sum of the distances traveled to pick-up each requested trip and then to drop-off each requested trip. From geometric probability, the average distance to the nearest of $n$ uniformly distributed points within an area of size $A$ is:

$$
\begin{equation*}
E(d \mid n, A)=\frac{r}{2} \sqrt{\frac{A}{n}} \tag{6}
\end{equation*}
$$

where $r$ is a unitless adjustment factor for the network that can be thought of as the ratio between the actual network distance and the straight-line distance. The distance traveled to pick-up a customer is associated with the nearest among $\lambda \mathrm{T}$ potential customers. The drop-off is associated with the nearest among $n$ customers
on-board. Therefore, the total VMT within an analysis period of duration $t_{p}$ is given by

$$
\begin{equation*}
V M T=r_{V M T} \frac{1}{2}\left(\frac{1}{\sqrt{\lambda T}}+\frac{1}{\sqrt{n}}\right) \lambda t_{p} \sqrt{A} \tag{7}
\end{equation*}
$$

where $r_{V M T}$ is the factor that is calibrated to fit the observed data for the region. This model forms a linear relationship between the right-hand side expression and the VMT, so the value of $r_{V M T}$ can be estimated using linear regression.

### 4.2.2 Vehicle Hours Traveled (VHT) model

The model for VHT is based on the VMT model in equation (7) with three important changes. First, the distance traveled is converted to travel time by dividing by the average network speed, $v$. Second, the time required for loading and unloading each passenger, $b$, is added. Finally, the calibration factor is replaced by a new parameter $r_{V H T}$, which allows for the relationship between travel time variables to differ from the relationship between travel distance variables.

$$
\begin{equation*}
V H T=\lambda t_{p}\left[b+r_{V H T} \frac{1}{2 v}\left(\frac{1}{\sqrt{\lambda T}}+\frac{1}{\sqrt{n}}\right) \sqrt{A}\right] \tag{8}
\end{equation*}
$$

In theory, $r_{V H T}=r_{V M T}$ if there is no wasted time or slack in the system schedule. In practice we always expect $r_{V H T}>r_{V M T}$, and the degree to which they differ provides some indication of how efficiently the system is operating compared to an unachievable baseline. For estimation of $r_{V H T}$, it can be useful to rearrange the terms as follows:

$$
\begin{equation*}
\frac{V H T}{t_{p}}-\lambda b=r_{V H T} \frac{1}{2 v}\left(\frac{1}{\sqrt{\lambda T}}+\frac{1}{\sqrt{n}}\right) \lambda \sqrt{A} \tag{9}
\end{equation*}
$$

where the slope relating the right-hand side expression to the left-hand side expression is the calibrated value for $r_{V H T}$.

### 4.2.3 Fleet size (M)

The number of vehicles in operation is closely related to the VHT. Within a time period, operations are assumed to be in roughly steady state conditions, meaning that there are no peaks within each interval. In this case the fleet required during each time period is

$$
\begin{equation*}
M=\frac{V H T}{t_{p}} \tag{10}
\end{equation*}
$$

because each vehicle is assumed to be fully occupied for the entire time period. The required fleet size for a region is the maximum fleet size required over the course of a day, so the busiest time period determines the necessary resources.

### 4.2.4 Total operating and marginal cost

The total costs of operating a paratransit service are based on the magnitude of the operational components that are modeled. These components correspond to VHT, VMT, and M and the total cost model is expressed as follows:

$$
\begin{equation*}
T C=a_{0}+a_{1} V M T+a_{2} V H T+a_{3} M \tag{11}
\end{equation*}
$$

where $a_{0}$ are the fixed costs associated with setting up a paratransit operation in a region, and $a_{1}, a_{2}$, and $a_{3}$ are the incremental cost of each vehicle-hour, vehicle-mile, and vehicle in the fleet. The actual costs to an agency depend on the details of the operating contracts. On some level, however, the underlying costs of operating a DRT service follow a pattern as shown in equation (11).

By replacing VMT, VHT and M in equation (11) with equations (7), (8), and (10), we get an expanded total cost equation. The first derivative of this equation with respect to $\lambda$ is the marginal cost of the paratransit service. Considering $\lambda$ as the
total daily demand, we can express the marginal cost as follows:

$$
\begin{gather*}
M C(\lambda)=\frac{a_{1} t_{p} r_{V M T} \sqrt{A}}{2}\left(\frac{1}{2 \sqrt{\lambda T t_{p}}}+\frac{1}{t_{p} \sqrt{n}}\right)+ \\
\left(a_{2} t_{p}+a_{3}\right)\left[\frac{b}{t_{p}}+\frac{r_{V H T} \sqrt{A}}{2 v}\left(\frac{1}{2 \sqrt{\lambda T t_{p}}}+\frac{1}{t_{p} \sqrt{n}}\right)\right] \tag{12}
\end{gather*}
$$

If a trip can be served by alternative providers at a lower cost, then its allocation to them leads to lower total cost for the paratransit service. By implementing this process repeatedly, we can define the number of trips that should be shifted to taxis or TNCs, if any. More details on allocating trips to paratransit service or taxis/TNCs using the aggregate models mentioned above are included in Turmo et al. (2018). In this study, the authors utilized the aggregate model to identify the number of trips that should be shifted to alternative providers to minimize the combined cost of the system.

### 4.3 Proposed algorithm

The aggregate operations model described above provides unbiased estimates of the total operating parameters associated with serving a level of demand in a service area. These totals are useful for estimating the total monthly or annual costs of operations, but the model is not sensitive to specific variations in the timing and location of requested trips. By its very nature, the aggregate model treats all trips as equivalent components of the total demand $\lambda$.

In order to decide which trips to allocate to alternative providers versus keep on the ADA van service, it is necessary to estimate the marginal cost of each ADA paratransit trip and the corresponding cost of service by the other provider. In order to do this, the specific routing of vehicle must be known so that the incremental
effect on cost of unilaterally eliminating each requested trip can be calculated. The trip records provide information about the actual vehicle routes that are operated each day, but the task of allocating all trips requires the routes can be incrementally re-optimized each time a requested trip is shifted to a taxi or TNC.

The following subsections present a proposed approach to quickly create a routing plan for vehicles based on a set of actual trip requests in the region. Then the marginal cost of each trip is estimated for each trip as a result of this routing and compared against the estimated taxi or TNC fare of the same trip. The trip with the greatest cost benefit for switching is removed from the pool of ADA trips, and the routes are re-optimized. In this manner, trips are incrementally shifted to the alternative provider until no cost savings can be achieved. It is possible that all trips should ultimately be shifted to taxis or TNCs or that some subset of the total ADA demand should shift. This approach is designed in such a way that the algorithm could be run daily as part of the vehicle routing solution.

### 4.3.1 Algorithm to construct representative routes

A fast algorithm is needed to construct the hypothetical vehicle routes, because the procedure will be run iteratively each time a trip is allocated to taxis or TNCs. Such an algorithm for constructing routes is a Greedy Algorithm, which is a heuristic in which each vehicle route is constructed in sequence by choosing among available trips that result in the most efficient route. The paratransit trip configuration considered in this study is illustrated in Figure 4.1. As shown in this figure, the simplifying assumptions made here are that each vehicle is serving one passenger per ride and passengers experience zero waiting times before service. More specifically, the algorithm works as follows:

1. Daily trip data within a region is sorted chronologically by requested time.
2. The first vehicle route starts from the first requested trip of the morning. Assuming the first pick-up is on-time, the arrival time at the drop-off location is estimated based on the straight-line distance factored up by the network circuity factor and divided by the average network speed. Upon drop-off, the vehicle becomes available to serve the next customer.
3. The time to serve each other unserved pick-up request is calculated by adding together the estimated travel time (straight-line distance factored up for network circuity and divided by average speed) and then additional waiting time until the requested pick-up time. Any trips that could only be served with negative waiting time are eliminated as infeasible next pick-ups. The trip with the shortest time from drop-off to pick-up is selected as the next trip in the route.
4. Steps 2 and 3 are repeated until one of two constraints are reached: the duration of the route has reached the maximum length of a shift (if such a constraint is desired), there are no more trips at the end of the day left to serve.
5. Steps 2, 3, and 4 are repeated to construct each route until there are no unserved trip requests left.
6. Using garage location data, the distance and travel time associated with assigning a vehicle to the route from each existing garage is estimated. This distance and time is the dead-head to get the empty vehicle from the garage to its first pick-up and from the last drop-off to the same garage. The cost associated with the distance and time is estimated, and the garage associated with the lowest cost is assumed to supply the vehicle for the route.
7. The daily totals for VMT, VHT, and the required fleet size are calculated from these constructed trips in the same manner used for the actual vehicle routing
plan. The Greedy Algorithm's operating characteristics should be compared with historic operations and proper calibration should be implemented, if possible.


Figure 4.1: Paratransit trip configuration

### 4.3.2 Estimation of marginal cost of each trip

### 4.3.2.1 Marginal cost of each paratransit trip

The total paratransit costs of serving the daily demand are estimated based on equation (11), so they are considered a function of fleet size, VHT and VMT. The marginal cost of each trip is estimated by considering the effect of unilaterally removing the trip on the remaining costs of operations. Each trip falls into one of three cases, each having different degrees of impact on operations and cost.

- Type 1 - Trips that are in a route that contains only that trip are the costliest. Eliminating the trip reduces the required fleet size by 1 vehicle; eliminates the VMT associated with going to pick-up, drop-off, and loaded travel in-between; and eliminates the VHT associated with the route. These trips have a very high marginal cost, because reducing the number of vehicles in the fleet saves a lot of
money. Type 1 trips are associated with the peak demand times during which all other vehicles are occupied, and an additional vehicle must be brought into service to serve a single requested trip.
- Type 2 - Trips that are at the beginning or end of a route have a moderate cost. Eliminating a Type 2 trip does not affect the fleet size, but it does eliminate the VMT associated with serving the trip and reduces the VHT by allowing the vehicle to start operating later or stop operating sooner. Once all of the Type 1 trips have been eliminated, Type 2 trips are the most likely to have high marginal cost.
- Type 3 - Trips that are served in the middle of a route typically have the lowest cost, because eliminating the trip only affects VMT. The fleet size and VHT is unchanged because the vehicle must still be out in operation to serve the preceding and following trip. The effect of removing a Type 3 trip is only the change in VMT associated with deviating the vehicle's route for the pick-up, to carry the passenger, and after drop-off. This saving is offset by the distance that would have been traveled anyway from the previous drop-off to the next pick-up.


### 4.3.2.2 Cost of taxi and TNC trip

The proposed algorithm can be implemented for both taxis and TNCs as potential alternative providers of the paratransit trips. The part that needs update depending on the case is the definition of the alternative service's cost function. Taxis are part of an industry with many private operators providing services. The fares are charged according to a regulated cost function based on distance and time. The average cost
of a paratransit trip served by taxi is given by:

$$
\begin{equation*}
F_{t a x i}=\beta_{0}+\beta_{1} l+\beta_{2} d \tag{13}
\end{equation*}
$$

where $l$ is the length of the trip (in miles) and $d$ is the delayed time experienced (in minutes). Proper average cost coefficients $\beta_{0}(\$)$ for fixed cost, $\beta_{1}(\$ / \mathrm{mi})$ for the trip length and $\beta_{2}(\$ / \mathrm{min})$ for the trip distance can be identified by analyzing historic data or through existing literature and online sources.

The cost of serving a trip by TNC varies depending on the specific service provider, time of day, and length of trip. It is not possible to know exactly what the trip will cost, because prices can fluctuate in real-time in response to the relative supply and demand (dynamic or surge pricing). The basic TNC fare is relatively consistent. A TNC cost function is the following:

$$
\begin{equation*}
F_{T N C}=\max \left\{f_{\min }, \gamma_{0}+\gamma_{1} l+\gamma_{2} t\right\} \tag{14}
\end{equation*}
$$

where $l$ is the trip length (in miles) and $t$ is the trip time (in minutes). TNC fares are usually structured so that a minimum amount, $f_{\min }(\$)$, is charged no matter how short or fast the trip is. Proper average cost coefficients $\gamma_{0}(\$)$ for fixed cost, $\gamma_{1}$ $(\$ / \mathrm{mi})$ for the trip length and $\gamma_{2}(\$ / \mathrm{min})$ for the trip time are available online for every TNC and for different regions. As in the case of taxis, average values available online could be replaced by a more detailed analysis of trip costs based on historic data, if such data are available. In most cases, cost data are not available for this type of transportation service.

### 4.3.3 Procedure to allocate trips to paratransit or taxi/TNC

Equipped with a method to estimate the marginal cost of each trip on the ADA paratransit service and the cost of the subsidy to serve it with a taxi or TNC, the trips with the greatest benefit of shifting to another provider can be identified. The procedure for optimally allocating trips is as follows.

1. Group all of the requested ADA paratransit trips in a region into routes using the algorithm described in Section 4.3.1.
2. Identify the trip with the greatest estimated cost saving associated with a switch to service with a taxi or TNC using the cost calculations presented in Section 4.3.2. The net marginal costs, $M C_{n e t}$, are calculated as shown below:

$$
\begin{equation*}
M C_{n e t}=M C_{p}-M C_{t a x i / T N C} \tag{15}
\end{equation*}
$$

where $M C_{p}$ is the marginal cost of the trip if served by the paratransit vans and $M C_{t a x i / T N C}$ the marginal cost of the trip if served by a taxi/TNC.
3. Eliminate the trip from the pool of requested ADA paratransit trips and repeat steps 1 and 2. Each time updating the total cost estimate for the ADA paratransit operations and adding the cumulative cost of all of the trips shifted to alternative provider. This process can be repeated until there are no trips remaining on the ADA paratransit service.

In order to implement the proposed methodology, both historic and daily acquired data are required. In terms of historic data, the average network speed (mph) and the average loading and unloading times (min) are needed. These values can be estimated either on a daily or a time-period-specific basis, depending on the impact that
this is expected to have on the performance of the model. Proper cost coefficients for both the paratransit and the taxi/TNC operation need to be identified. If garages are considered in the route scheduling, then their location coordinates (latitude, longitude) should be known. Regarding the demand related information, the operators should know the exact requested time of the trip (hh:mm:ss), as well as the origin and the destination coordinates (latitude, longitude). Each trip's expected length and duration can be estimated either analytically (e.g., straight line distance from coordinates properly calibrated) or through available tools.

In practice, the total cost to the agency is minimized when it is no longer possible to save money by transferring trips from ADA paratransit to the taxi/TNC. Although it may appear at the first iteration that there are many ADA trips with very low marginal cost, this incremental approach shows how this cost increases as other trips are removed. As Type 2 trips are removed from a route, formerly Type 3 trips become new Type 2 trips. Eventually, when one trip is left in the route, it becomes a costly Type 1 trip. This means that the marginal cost of each trip depends on all of the other demand that is served. Trips that appear to be very cost efficient with one set of demand may become very costly as the trips around are shifted to alternative providers.

The final challenge is that it may not be possible to shift all trips to TNCs either because the vehicles are not accessible to some customers or because some customers are reluctant to use an alternative service provider. In this case, the same procedure is implemented with the difference being that only feasible trips are actually eliminated from the pool of requested ADA paratransit trips and shifted to TNCs. The process of shifting trips must then stop when no feasible trips remain. Figure 4.2 summarizes the proposed algorithm in a flow chart format.


Figure 4.2: Flow chart of proposed methodology

### 4.4 Study site

There are two main types of data on which the models and analyses in this study are based. First, customer records from the MBTA provide demographic information
about each eligible paratransit customer which can be used to associate travel patterns with personal characteristics such as age and type of disability. Second, detailed records from each ADA paratransit trip served include the locations and times of each passenger pick-up and drop-off as well as identification of the vehicle or route that served each trip. This data not only shows the temporal and spatial distribution of ADA paratransit trips, it can also be used to reconstruct the vehicle routes, which reveals the operations associated with serving the demand.

### 4.4.1 Description

The MBTA operates public transit services throughout Greater Boston, Massachusetts, including buses, light rail, heavy rail, commuter rail, electric trolleybuses, and ferries. The RIDE is generally available to customers with eligible disabilities and between the hours of 5 AM and 1 AM daily. It is typical of many ADA paratransit services across the United States in that operation of vehicles is provided by private operators under contract to the MBTA. During the year 2017, four different providers operated the system under contract. In this study we analyse the results from the three largest of them. The service area is divided in three subregions (North, West, South) which overlap forming a shared area. For the purposes of this study, we name "Provider 1" the provider that served North region before June 2017, "Provider 2" the one who operated in West region during the entire year and "Provider 3" the one that operated partly in South region before June and fully on North and South region after June. The shared area is considered in all three cases. Figure 4.3 illustrates the operating areas for each provider where garages are represented by black squares.

Although the ADA only requires that paratransit service be made available within $3 / 4$ of a mile of MBTA bus and subway stops, the MBTA makes The RIDE available to customers throughout 58 towns and cities in Greater Boston. This is common


Figure 4.3: MBTA paratransit service's monthly pick-up requests and garage locations for many agencies, because the $3 / 4$ of a mile boundary can exclude many important origins and destinations in a region, limiting access for customers who may not have other options for travel. A distinction is made in the fares charged:

1. Local ADA one-way fare for trips with origin and destination within $3 / 4$ mile of an MBTA bus or subway stop is $\$ 3.15$.
2. Premium one-way fare for trips with an origin and/or destination further than $3 / 4$ mile from an MBTA bus or subway stop is $\$ 5.25$.

The MBTA is implementing a pilot that allows customers to perform subsidized trips with TNCs. The fare policy considered in this study is to charge the first $\$ 2$ to the pilot participant and pay the next $\$ 40$ of fare. Since the travel time and network distance have been calculated for every requested trip and reported in the trip database, estimation of the subsidy for each trip is a straightforward calculation
using equation (14) and subtracting $\$ 2$ for each trip. The respective parts of the proposed methodology can be very easily adjusted to this fare policy.

### 4.4.2 ADA paratransit customer data

The database of eligible customers for The RIDE contains records for 40,721 individuals. Personal identifying information is not necessary for the analysis of this study, but the following data fields were available:

1. Customer ID - A unique number is assigned to each eligible ADA customer. This ID allows us to track the trips that each individual makes and relate those trips to other customer characteristics.
2. Date of Birth - The customer's date of birth allows us to calculate age, which has the potential to be an explanatory factor for travel behavior.
3. Home ZIP Code - The zip code for each customer's registered home address provides an indication of where customers reside and where many of their trips are likely to start or end.
4. Disability - The qualifying disability or disabilities associated with customer are recorded, and these have the potential to be explanatory factors for travel behavior.
5. Equipment - In addition to customer disabilities, the type or types of equipment that the customer uses is listed. This includes mobility devices such as wheelchair, power chair, scooter, walker, cane, etc. This is also the field where specific vehicle requirements are listed, such as requirement of a lift or service only with a van. This field is particularly important for identifying which cus-
tomers are ambulatory and which customers require a Wheelchair Accessible Vehicle (WAV).

### 4.4.3 The RIDE trip records

In addition to records about each customer, the MBTA maintains a database of all The RIDE trips. These records include a detailed accounting of where and when each customer travelled, and which vehicle or route was used to serve them. For this study, the MBTA provided the research team with all 4,012,592 trip records from January 2016 (prior to the Pilot's start in October 2016) through March 2018. Each trip record includes the following data that is used in the analysis:

1. Trip ID - Each trip is uniquely identified by an ID.
2. Customer ID - The ID for the customer requesting the trip allows each trip to be linked to the specific customer characteristics in the customer data table.
3. Trip Date - The calendar date of each requested, scheduled, and served trip.
4. Subscription - Customers that make regular trips (e.g., to and from work) are able to request their trip as a subscription rather than having to call in the same request over and over again. This data field indicates the ID of the associated subscription, if applicable.
5. Provider - Each trip is served by a private operator that works under contract with the MBTA. This field indicates which provider serves the trip. This provides an indication of the region in which the trip is assigned, because each of the three regions is initially served by a different provider. Some reorganization during the time period of observation has resulted in changing geographic coverage for each provider.
6. Pick-up Location - The address and latitude/longitude coordinates of the requested pick-up location are recorded. This is used (along with the drop-off location) to determine if the trip is within the required ADA service area or in the broader "premium" service region in which the ADA does not require service.
7. Drop-off Location - The address and latitude/longitude coordinates of the requested drop-off location. This is used along with the pick-up location to categorize trips.
8. Origin-Destination Network Distance - The estimated driving distance from the pick-up location to the drop-off location is recorded, assuming the trip can be served as a direct ride without intermediate stops. Ultimately most trips are served directly in this manner, but some vehicles are routed to share multiple rides, so the actual distance travelled by a customer may be somewhat greater.
9. Estimated Trip Time - Based on the location and time of day of the pick-up and drop-off, a travel time estimate is generated by the scheduling software for a direct trip following the network distance above. This is a travel time estimate that may be greater or less than the actual travel time for the passenger.
10. Requested Pick-up Time - This is the time that the customer initially requested to be picked up by The RIDE.
11. Promised Pick-up Time - This is the time that The RIDE offered to the customer during the booking process. Customers are expected to be prepared to board the vehicle from 5 minutes before to 15 minutes after the promised time.
12. Arrival Time at Pick-up - This is the time that the vehicle arrived at the pickup address. As described above, the vehicle is intended to arrive between 5
minutes before to 15 minutes after the promised pick-up time. Any arrival after this time window is considered to be late.
13. Departure Time from Pick-up - This is the time that the vehicle departs the pick-up location. The difference between the departure time and the arrival time at pick-up is the time that the driver waits for the customer to get ready and to get into the vehicle.
14. Arrival Time at Drop-off - This is the time that the customer actually arrives at his or her destination. The difference between the arrival time at dropoff and the departure time from pick-up is the time that the customer spends traveling in the vehicle, including any intermediate stops. For trips that are served without intermediate stops, this elapsed time can be used with the origindestination network distance to calculate the average speed of the vehicle in the network.
15. Vehicle ID or Route ID - Depending on the month, the data set includes a field for the vehicle ID or route ID. Within a day, all trips with a common vehicle/route ID can be grouped to identify the actual vehicle routing. By linking together trips in this way, the actual operations of all The RIDE vehicles can be deduced in terms of the total number of vehicles operating, VHT, and VMT.

### 4.4.4 Relevant explanatory variables

In order to develop and calibrate the aggregated models, as well as to perform the proposed methodology, all necessary explanatory variables related with the operation of the paratransit service need to be calculated. For the purpose of this analysis, we consider time periods of length $t_{p}=3$ hours, which results in 5 time periods per day:
$6 A M-9 A M ; 9 A M-12 P M ; 12 P M-3 P M ; 3 P M-6 P M$; and $6 P M-9 P M$. Very few trips are completed outside of these hours, and they are not considered neither for the calibration of the aggregate operations modeling nor the proposed methodology. Annual average values by operator, day of week, and time of day are included in Table 4.1. Even though the average values used in our study were calculated on a monthly and daily basis, the values included in the table present the expected magnitudes of each variable for each provider.

The percentage of total time by number of passengers onboard is shown in Figure 4.4. The three providers are similar, and vehicles in all regions spend most of their time without any passengers onboard at all. Although the vans are observed to carry as many as 8 passengers, the vehicles are rarely loaded with more than one passenger at a time. This observation highlights that the simplifying assumption of developing a routing strategy with maximum one passenger onboard is not far from what is in reality implemented by the MBTA.


■ Provider 1 - Provider 2 Provider 3

Figure 4.4: Percentage of total time by number of passengers onboard for MBTA paratransit service

The unloading time is not observed, because only records of vehicle arrival time at the drop-off location are available. A working assumption is that the unloading time is one third as long as loading, because drivers do not have to wait for customers to get ready and come out to the vehicle. Without including the shared area, the service area, $A$, for the North is $211.8 \mathrm{mi}^{2}$, for South is $330.1 \mathrm{mi}^{2}$ and for West $216.6 \mathrm{mi}^{2}$. The shared area is $64 \mathrm{mi}^{2}$. Finally, for The RIDE, the policy is to pick-up customers within a 20 minute window, $T$, from 5 minutes before the scheduled pick-up time to 15 minutes after.

### 4.4.5 Observed operational outputs

Considering a vehicle's travel between consecutive stops as a segment, the straightline distance associated with each segment is calculated based on the difference of latitude and longitude of the coordinates. Trip segments that correspond to a single customer's travel directly from pick-up to drop-off have a corresponding network distance reported in the data set. By comparing the straight-line distance and the network distance for these segments, a network circuity factor can be estimated. When multiplied by the straight-line distance, this factor provides an estimate of the actual network distance traveled. Figure 4.5 illustrates the estimation of this factor for each provider's service area. Using these factors for all trip segments that their network distance is not reported in the dataset, the calculation of VMT could be completed. A summary of the average VMT and VHT per three-hour time period by operator, day of week, and time of day is summarized in Table 4.2.

Table 4.1: Average values of relevant explanatory variables

| Time Period | Provider 1 |  | Provider 2 |  | Provider 3 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Weekday | Weekend | Weekday | Weekend | Weekday | Weekend |
|  | Average Loading/Unloading Time, b (min/passenger) |  |  |  |  |  |
| $6 \mathrm{AM}-9 \mathrm{AM}$ | 5.03 | 5.25 | 7.48 | 8.82 | 6.33 | 7.40 |
| $9 \mathrm{AM}-12 \mathrm{PM}$ | 5.37 | 5.65 | 8.13 | 8.91 | 6.95 | 7.22 |
| $12 \mathrm{PM}-3 \mathrm{PM}$ | 6.19 | 6.27 | 9.26 | 8.55 | 8.28 | 7.70 |
| $3 \mathrm{PM}-6 \mathrm{PM}$ | 6.59 | 6.62 | 9.30 | 9.62 | 8.10 | 8.41 |
| $6 \mathrm{PM}-9 \mathrm{PM}$ | 6.76 | 6.93 | 9.38 | 10.13 | 9.00 | 9.13 |
|  | Average Vehicle Occupancy, $n$ (passengers/vehicle) |  |  |  |  |  |
| $6 \mathrm{AM}-9 \mathrm{AM}$ | 1.46 | 1.24 | 1.41 | 1.20 | 1.47 | 1.26 |
| $9 \mathrm{AM}-12 \mathrm{PM}$ | 1.35 | 1.33 | 1.27 | 1.22 | 1.31 | 1.28 |
| $12 \mathrm{PM}-3 \mathrm{PM}$ | 1.38 | 1.29 | 1.35 | 1.23 | 1.40 | 1.29 |
| $3 \mathrm{PM}-6 \mathrm{PM}$ | 1.49 | 1.31 | 1.41 | 1.22 | 1.50 | 1.30 |
| $6 \mathrm{PM}-9 \mathrm{PM}$ | 1.28 | 1.26 | 1.25 | 1.18 | 1.28 | 1.27 |
| Average Network Speed, v (miles/hour) |  |  |  |  |  |  |
| $6 \mathrm{AM}-9 \mathrm{AM}$ | 15.56 | 22.06 | 15.13 | 20.21 | 16.03 | 21.85 |
| $9 \mathrm{AM}-12 \mathrm{PM}$ | 17.83 | 18.69 | 15.57 | 16.50 | 17.60 | 18.48 |
| $12 \mathrm{PM}-3 \mathrm{PM}$ | 17.29 | 17.70 | 15.56 | 15.96 | 16.74 | 17.41 |
| $3 \mathrm{PM}-6 \mathrm{PM}$ | 14.69 | 18.19 | 13.05 | 16.67 | 14.01 | 17.86 |
| $6 \mathrm{PM}-9 \mathrm{PM}$ | 19.36 | 20.80 | 16.54 | 19.42 | 18.83 | 20.91 |
| Average Demand, $\lambda$ (trips/hour) |  |  |  |  |  |  |
| $6 \mathrm{AM}-9 \mathrm{AM}$ | 128 | 43 | 146 | 41 | 191 | 71 |
| $9 \mathrm{AM}-12 \mathrm{PM}$ | 183 | 92 | 165 | 72 | 256 | 131 |
| $12 \mathrm{PM}-3 \mathrm{PM}$ | 181 | 82 | 188 | 70 | 262 | 119 |
| $3 \mathrm{PM}-6 \mathrm{PM}$ | 118 | 56 | 126 | 50 | 173 | 81 |
| $6 \mathrm{PM}-9 \mathrm{PM}$ | 35 | 27 | 35 | 23 | 52 | 41 |



Figure 4.5: Network versus straight-line distance for a) Provider 1, b) Provider 2, c) Provider 3 service area

### 4.5 Results

### 4.5.1 Aggregate model of The RIDE paratransit operations

The calibration of VMT and VHT using equations (7) and (9) resulted in $r_{V M T}$ and $r_{V H T}$ factors, which are included in Table 4.3. Note that in all cases the value of $r_{V H T}>r_{V M T}$, and the difference is greater during time periods when demand is lower. Also, the model fit as represented by the $R^{2}$ values is lower for $r_{V H T}$ than for $r_{V M T}$, especially at times with lower demand. This is an indication of greater variability in the data, which leads to greater uncertainty in model estimates. Regarding the fleet size, it is directly related to VHT, so the same model outcomes are used to estimate the number of vehicles using equation (10).

Table 4.2: Average VMT and VHT by time of day, 2017

| Time Period | Provider 1 |  | Provider 2 |  | Provider 3 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Weekday | Weekend | Weekday | Weekend | Weekday | Weekend |
| Average Vehicle Miles Traveled per Time Period, VMT (veh-mi) |  |  |  |  |  |  |
| $6 \mathrm{AM}-9 \mathrm{AM}$ | 2399 | 857 | 2267 | 644 | 3713 | 1455 |
| $9 \mathrm{AM}-12 \mathrm{PM}$ | 3093 | 1581 | 2299 | 1008 | 4378 | 2262 |
| $12 \mathrm{PM}-3 \mathrm{PM}$ | 3118 | 1478 | 2814 | 1051 | 4757 | 2205 |
| $3 \mathrm{PM}-6 \mathrm{PM}$ | 2091 | 1072 | 1976 | 817 | 3324 | 1641 |
| $6 \mathrm{PM}-9 \mathrm{PM}$ | 705 | 618 | 572 | 442 | 1125 | 975 |
| Average Vehicle Hours Traveled per Time Period, VHT (veh-hr) |  |  |  |  |  |  |
| 6 AM - 9 AM | 419.6 | 165.9 | 460.4 | 167.8 | 617.1 | 277.1 |
| $9 \mathrm{AM}-12 \mathrm{PM}$ | 538.6 | 268.8 | 481.2 | 209.4 | 799.7 | 394.6 |
| $12 \mathrm{PM}-3 \mathrm{PM}$ | 529.4 | 270.1 | 541.9 | 232.7 | 880.8 | 444.0 |
| $3 \mathrm{PM}-6 \mathrm{PM}$ | 405.1 | 201.3 | 469.1 | 194.9 | 646.2 | 324.3 |
| $6 \mathrm{PM}-9 \mathrm{PM}$ | 151.2 | 101.3 | 148.1 | 96.1 | 269.6 | 181.9 |

Without detailed cost information from the MBTA's operators, it is necessary to make cost estimates based on data from other operators. For the purposes of this study, we use cost factors estimated from the PVTA in Springfield, Massachusetts (Turmo et al., 2018):

- Cost per Vehicle Mile of Operation, $a_{1}=0.518 \$ / \mathrm{veh} \cdot \mathrm{mile}$;
- Cost per Vehicle Hour of Operation, $a_{2}=19.89 \$ / \mathrm{veh} \cdot \mathrm{hour} ;$
- Cost per Vehicle, $a_{3}=150.81 \$ /$ veh $\cdot$ day or $55,046 \$ /$ veh $\cdot$ year (fleet size cost)

Regarding the TNC cost function, we identified online the cost coefficients for one of the currently available TNCs in the area of Boston. The respective TNC cost function is described as:

$$
\begin{equation*}
F_{T N C}=\max \{6.85,3.95+0.88 l+0.36 t\} \tag{16}
\end{equation*}
$$

where $l$ is the trip length (in miles) and $t$ is the trip time (in minutes). This TNC's

Table 4.3: Modeled values of $r_{V M T}$ and $r_{V H T}$ by time of day, 2017

| Time Period | Provider 1 |  | Provider 2 |  | Provider 3 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | r |  | $R^{2}$ | r | $R^{2}$ | r |
| Modeled value of $r_{V M T}$ |  |  |  |  |  |  |
| 6 AM - 9 AM |  |  |  |  |  |  |
| 9 AM - 12 PM | 1.08 | 0.98 | 0.88 | 0.99 | 0.77 | 0.99 |
| 12 PM - 3 PM | 1.09 | 0.98 | 0.94 | 0.87 | 0.98 | 0.99 |
| 0.75 | 0.99 |  |  |  |  |  |
| 3 PM - 6 PM | 1.11 | 0.97 | 1.06 | 0.98 | 0.84 | 0.99 |
| 6 PM - 9 PM | 1.24 | 0.88 | 0.98 | 0.87 | 0.90 | 0.95 |
| All times | 1.11 | 0.99 | 0.92 | 0.97 | 0.78 | 0.98 |
| Modeled value of $r_{V H T}$ |  |  |  |  |  |  |
| 6 AM - 9 AM | 1.95 | 0.92 | 1.71 | 0.94 | 1.32 | 0.94 |
| 9 AM - 12 PM | 1.92 | 0.93 | 1.50 | 0.96 | 1.33 | 0.97 |
| 12 PM - 3 PM | 1.85 | 0.94 | 1.52 | 0.96 | 1.37 | 0.96 |
| 3 PM - 6 PM | 1.89 | 0.94 | 1.69 | 0.95 | 1.32 | 0.96 |
| 6 PM - 9 PM | 2.60 | 0.85 | 2.11 | 0.87 | 2.12 | 0.92 |
| All times | 1.91 | 0.95 | 1.59 | 0.95 | 1.36 | 0.95 |

fares are structured so that a minimum of $\$ 6.85$ is charged no matter how short or fast the trip is. The proposed algorithm is structured such that any TNC or taxi cost can be easily incorporated and further studied. Our purpose is not to evaluate the performance of a specific TNC, but to identify the efficiency of our method in allocating trips, no matter who is the alternative provider.

By implementing the results of our analysis into the marginal cost function derived from the aggregate models, we get the graph shown in Figure 4.6. Data from January 23, November 14 and October 17 were respectively used for Provider 1, Provider 2 and Provider 3. Comparing the marginal cost of the paratransit service with respect to demand shifted to TNC to the TNC costs associated with the requested trips, we observe that there is no point of intersection between the two systems (i.e., paratransit service and TNCs). Thus, a combined system would not be efficient. Moreover, since the TNC costs are lower than the paratransit service's, it is evident that the allocation
of all requested trips to TNCs would be beneficial, if that was practically feasible.


Figure 4.6: Marginal cost of the paratransit service with respect to demand shifted to TNC and the TNC cost for a) Provider 1, b) Provider 2, c) Provider 3 service area

### 4.5.2 Optimized allocation of trips to paratransit and TNCs

The first step in this part of the study is to implement the routing algorithm for several weekdays and for all providers and then compare the estimated VHT and VMT with their actual values during those days. Although the Greedy Algorithm does not exactly match the observed operations, the model produces estimates that are proportional. Figure 4.7 shows the results of calibrating the outputs of the Greedy Algorithm for the first 10 weekdays of January 2017, for North region (provider 1). The $R^{2}$ is higher than $90 \%$ in both cases, indicating a good fit. Similar results are achieved for the other two providers. Such calibration allows for having more realistic operating values for our method's implementation, partly reducing the errors caused by the simplifying assumptions of having maximum one passenger onboard and that passengers do not experience waiting times before service.


Figure 4.7: Calibrating greedy algorithm's outputs for a) VHT and b) VMT in the North Region during January 2017

The implementation of the entire optimal allocation methodology for each provider is shown in Figure 4.8 (blue curve). The horizontal axis indicates the cumulative number of trips shifted to TNCs and the vertical axis is the total agency cost, including subsidies paid for TNC trips. The days selected for application are the same as the ones used in Figure 4.6 for every provider. The costs drop most dramatically for
the first few trips as inefficient routes serving peak demand are eliminated. For the particular dates selected, and the cost parameters used, the agency costs continue to decline until all demand has been shifted to TNCs. In this case, the lowest cost is achieved by shifting all trips from ADA paratransit to TNCs, thus confirming the conclusion from the aggregate models implementation presented above.

Also in this figure, the effect of shifting trips strategically to TNCs is compared with alternative patterns. First, the effect of allocating trips using their TNC cost as criterion is investigated (orange curve). As we can see, the costs still decline, but not as efficiently as using the estimated marginal costs. Second, the effect on total cost of randomly shifting trips was calculated for 10 realizations for Provider 1. The mean and $95 \%$ confidence interval based on these realizations is shown in gray. It is noteworthy that all cases have a general downward trend in cost. Therefore, total costs are expected to decline with increasing utilization of TNCs, at least with the estimated cost parameters.

### 4.5.2.1 Marginal cost of paratransit trips

The proposed algorithm is first implemented on the entire set of requested ADA trips for January 23, 2017, in the North region. By sequentially generating vehicle routes, the marginal cost of each trip is estimated and compared with the fare that would be charged if the trip were served by a TNC (based on TNC's fare structure). Figure 4.9 shows the distribution of the net marginal cost of each trip based on the estimated cost savings from shifting the trip away from The RIDE, $M C_{p}$, offset by the estimated cost of the TNC fare, $M C_{T N C}$. The net marginal costs are calculated as shown in equation (15).

A positive value in Figure 4.9 indicates that the marginal cost of paratransit


Figure 4.8: Comparison of costs by incrementally shifting trips to TNC under different patterns for a) Provider 1, b) Provider 2, c) Provider 3 service area


Figure 4.9: Distribution of estimated marginal costs for all trips, North, January 23, 2017
operations exceeds the expected TNC fare, and shifting the trip would save money. A negative value indicates that the expected TNC fare would exceed the marginal operating cost. The greatest values are for the small number of Type 1 trips for which an extra vehicle is needed to serve a single trip. Many trips with negative net marginal costs are the Type 3 trips within a route, which can be served at relatively low cost by the ADA fleet, because the vehicles are already out on the road.

By the proposed algorithm, the costliest trip should be shifted to a TNC, and then the routing process must be recalculated to estimate the new marginal costs. Therefore, trips with low (or negative) net marginal cost at the first iteration may become more beneficial contenders for shifting to TNCs as the routes change.

### 4.5.2.2 Shifted trips characteristics

It is also useful to look at the characteristics of the trips that are shifted. For example, the distribution of shifted trips by requested pick-up time is shown in Fig-
ure 4.10. Each curve in the figure shows the distribution of trip start times after a number of trips have been shifted to TNCs in the optimized order. The curve for all trips represents that existing case that all demand is served by The RIDE, and this curve exhibits two distinct peaks: a late morning peak at 11 AM and an afternoon peak at 3 PM. The first 250 trips to be shifted from The RIDE to TNCs are mostly Type 1 trips from the peak of the peak and Type 2 trips from the end of the day. The effect of removing these trips is to flatten the peaks and drop demand faster at the end of the day (as shown by the curve labeled " 250 Removed"). As trips are sequentially removed, the resulting demand pattern for The RIDE is a more uniform distribution, which allows vehicles to be used more consistently throughout the day.


Figure 4.10: Distribution of remaining ADA paratransit trips by time of day, Provider 1, January 23, 2017

A second analysis of the shifted trips is to look at the geographic locations of shifted trips within the region. Figure 4.11 shows a series of maps of the North and Shared regions, served by Provider 1 during January 2017. Each map shows the towns with requested trip pickups, and the colors indicate the percentage of requested trips
that are selected to shift to TNCs. Gray color is used to indicate towns with no pickup requests. There is not an obvious geographic pattern to the trips being removed, because trips in the suburbs and in the city center are selected for removal at each stage. In general, there seems to be a trend to eliminate suburban trips sooner than Boston city center trips. This is expected, because the requested trips in the suburbs tend to be longer in distance and more spread apart, which makes them costlier to serve with the ADA vans.


Figure 4.11: Percent of trips shifted to TNCs per town considering a) 100 first removed trips, b) 200 first removed trips, c) 300 first removed trips, d) 400 first removed trips, and e) 500 first removed trips for Provider 1, Jan. 23, 2017

### 4.6 Discussion

### 4.6.1 Special equipment

It is not always possible (or desirable) to shift all trips to taxis or TNCs. Part of this may be due to general attitudes or preferences regarding the modes, but the evidence suggests that customers with wheelchairs, power chairs, scooters, or other devices requiring a Wheelchair Accessible Vehicle (WAV) with a lift are unable or uncomfortable using a taxi or TNC. Applying the same procedure for optimally allocating trips to TNCs while leaving all wheelchair and lift customers on ADA paratransit, the red curve in Figure 4.12 shows the sequence of changing costs. During the day of January 23, 2017 that is presented in this figure, 250 out of the total 1,800 customers in North region used heavy equipment and, thus, could potentially have difficulties in using alternative providers.


Figure 4.12: Change in cost by incrementally shifting trips with greatest net marginal cost to TNC, North, January 23, 2017

Overall the pattern in the red curve is similar to the blue curve, with a steep initial
decline in agency costs associated with eliminating the most inefficient routes during peak demand. Then the cost savings accrue more slowly and the effect of shifting trips levels off before all of the feasible trips have been shifted. The prevailing pattern is still that costs are minimized when as many trips as possible are shifted to TNCs (although this may not necessarily happen in all regions with all demand patterns). In this case, because there are some customers that must always be served by the ADA paratransit van fleet, there is a small number of general trips that can be more efficiently served by the vans in combination with the other trips than shifting to TNCs.

Other concerns regarding users with heavy equipment refer to the higher loading and unloading times that they might require. Our efforts in this study were put in developing a fast and efficient model that would consider trip requests as individuals, in terms of temporal (request time) and spatial (origin and destination locations) trip characteristics. The proposed method, however, offers the flexibility to the operators to incorporate more individual user characteristics (e.g., different loading times), if required. Also, concerns may rise about the curb-to-curb (and not door-to-door) service that taxis and TNCs offer. Similar to customers with heavy equipment, customers that require door-to-door service could be eliminated from the trip requests that are examined for allocation to alternative providers.

As previously mentioned, current regulations do not allow customers to be directly assigned to TNCs, so the choice of using TNCs or the paratransit service is up to the customer. If trips that should be better served by taxis/TNCs are identified, however, then proper incentives can be defined in order to achieve the expected benefits from their shift to alternative providers. The proposed model aims at quantifying the potential benefits from such an operating strategy in order to reduce the very high operating costs of paratransit services and deploy ride-hailing services as more
strategic partners for public transit agencies.

### 4.6.2 Taxi/TNC pricing

The application presented in Section 4.5, considers some average cost coefficients for TNCs that are available online. These values, however, do not reflect changes that might occur within the day due to high levels of demand, for example (i.e., surge pricing). Also, these values are expected to change over greater time periods (e.g., semester). Figure 4.13 presents the changes in the trip allocation as a result of multiplying the time and distance cost coefficients by a multiplier, m, similar to how surge pricing works (although, surge pricing occurs during specific time periods within a day).


Figure 4.13: Change in trip allocation by increasing the value of TNC cost coefficients, North, January 23, 2017

Our purpose here is to identify the sensitivity of the proposed model's outputs related to the taxi/TNC costs. As we observe from this figure, the overall conclusion is intuitive. After the first (more costly) trips are reallocated, as $m$ increases the
total daily operating cost decreases with a lower rate (for $m=1.25$ to 1.75 ), then presents an almost flat rate ( $m=2.00$ ) and eventually increases ( $m=3.00$ to 5.00 ). A TNC with almost double the values of the cost coefficients considered in this study is an alternative provider almost equivalent with the MBTA's paratransit service, in terms of operating costs. The fare subsidies of maximum $\$ 40$ and all other pilot characteristics as described in Section 4.4.1 are still applied.

### 4.6.3 Environmental impacts

The operation of TNCs is often associated with environmental concerns due to the operation of more vehicles within a city network and the respective increase of emissions. According to existing literature, vehicle emissions are mostly related with VMT (Lyman et al., 2019) and they can be calculated by multiplying VMT with a properly calibrated emission factor. This factor depends on vehicle technology and network speed. For example, the average network speed for provider 1 is 17 mph , which corresponds to an emission factor of around $530 \mathrm{gCO}_{2}$-eq/veh $\cdot \mathrm{mi}$, for light duty automobiles. The graph for estimating this factor is included in Figure 4 of Lyman et al. (2019) with data from the California Air Resources Board EMFAC model. This is a macroscopic emission model that relates average emission rate to average speed with a u-shaped curve.

Vehicle speed and technology are expected to be similar between paratransit and TNC vehicles. Thus, the investigation of the environmental impacts of assigning paratransit trips to TNCs should focus on the effect of the strategy on the vehicle miles traveled. Figure 4.14 shows the change in total VMT as trips are assigned to TNCs for North region during January 23, 2017. In order to calculate the distance associated with serving the trips by TNCs, the distance needed for the TNC to pick a passenger up is required. The available dataset offers the distance needed to drop
passengers off (i.e., trip length) and existing literature offers insights on how these two types of distances are related in the case of TNCs. More specifically, Wenzel et al. (2019) estimate the distance needed by TNC vehicles to pick-up a passenger equal to $21 \%$ of the requested trip length, based on a study in Austin, Texas. According to San Francisco County Transportation Authority (2017), the authors estimate this distance to be equal to $26 \%$ of the trip length. The estimated value from San Francisco was used for the analysis illustrated in Figure 4.14.


Figure 4.14: Total VMT of serving paratransit trips in coordination with TNCs, North, January 23, 2017

Figure 4.14 shows that as trips are assigned to TNCs, the total VMT of the paratransit trips keeps reducing. This means that the total emissions decrease as a result of this strategy. This observation depends on several assumptions. First, the algorithm assumes that one passenger is served per trip by the paratransit vehicles. As discussed above, this assumption is not far from reality for the paratransit services of Boston. A second assumption is that TNCs and paratransit vehicles use the same type of vehicle. The two vehicle types are not expected to be very different especially
when the paratransit service uses sedans, but heavier multipassenger vans would make the emission factor greater for paratransit than for the TNC. Finally, the adopted relationship between TNC distance to pick-up and drop-off might be slightly different for the case study of Boston that is analyzed here. Figure 4.15 shows how the environmental impact of such strategy changes for different pick-up distance factors, $p$, ranging from $20 \%$ to $100 \%$ of the distance to drop-off. This figure shows that for $p>60 \%$ the effects of this strategy become harmful for the environment if all trips are assigned to TNCs.


Figure 4.15: Total VMT of serving paratransit trips in coordination with TNCs for various pick-up distance factors, North, January 23, 2017

### 4.7 Summary

The focus of this Chapter is to address challenges associated with high operating costs of on-demand services. The strategy analyzed here is the allocation of paratransit trips to alternative providers, such as taxis and TNCs. The relationship between public transit and ride-hailing systems can lead to benefits that are yet to
be identified. Our study aims to offer insights in the ways that taxis or TNCs can complement the public transit operations, by developing a flexible tool that could be easily adjusted to the specific needs of a paratransit agency in order to assist optimized decision making. A first approach to determine whether paratransit trips should be shifted to taxis/TNCs or not, includes the implementation of existing aggregated models. However, those models are lacking because not all paratransit trips have the same impact on operations and operating costs.

This study investigates the optimal allocation of trips between conventional ADA paratransit service and TNCs using data from the MBTA's ADA paratransit service called "The RIDE". MBTA implements a pilot that allows paratransit eligible passengers to perform subsidized trips with alternative providers, in an attempt to reduce their high operating costs. Although The RIDE is not currently structured in a way to assign riders directly to TNCs, this could be a potential future operating strategy that requires careful investigation. Using operating cost coefficients from existing literature, the results indicate that for all the studied regions within MBTA service area, all trips should be better served by TNCs. The implementation of the more detailed algorithm that is proposed through this dissertation confirmed this result. Apart from the number of trips to be assigned to taxis or TNCs, however, this algorithm orders the trips from the most to the the least costly, which could benefit the decision-making processes of the service operator. For the days of operation that were investigated here, the expected cost savings are approximately $48 \%$ for Provider 1, $53 \%$ for Provider 2, and $47 \%$ for Provider 3.

The proposed algorithm is developed to estimate the marginal cost of each paratransit trip in the context of the vehicle routings so trips can be incrementally reassigned to alternative providers when the costs make it advantageous to do so. The routing model is based on some simplifying assumptions to maintain its development
and implementation as fast and efficient as possible. These assumptions refer to vehicle occupancy of maximum one passenger and zero waiting times from customers before they are served. Expected deviations of the estimated VHT and VMT values from such assumptions are partly eliminated by performing proper calibrations. Regarding average occupancy particularly, it is proved that the assumption of only one passenger served per ride is quite realistic, at least for the case study considered here.

## 5 EFFICIENT IMPLEMENTATION OF FLEXIBLE SYSTEMS

The inefficiencies of fixed route (e.g., low user satisfaction) and on-demand services (e.g., high operating costs) can be avoided through the implementation of flexible services. This Chapter focuses on flexible systems which combine elements from both fixed and non-fixed systems. There are challenges, however, regarding which are the operational and demand related characteristics of a service area that can assure a successful implementation of such a service. This chapter proposes the development of a hybrid transit service, which could take the form of conventional fixed route, fully flexible route deviation or an intermediate form where vehicles that operate on a fixed corridor can deviate within a flexible region to serve passengers curb-tocurb. The resulted type of service is based on the operational characteristics and the expected demand levels for a given service area. The system's decision variables of flexible region and station spacing at a location $x$ are optimized considering continuous approximation approach with the objective of minimizing the total generalized costs from such a service.

The Chapter is organized as follows. Section 5.1 presents an introduction to the research topic. Section 5.2 describes the operation of the proposed service. The model development for the route deviation services is explained in Section 5.3, whereas Section 5.4 describes the cost modeling. The results of the optimization process are presented in Section 5.5. Section 5.6 includes numerical implementation of the proposed models considering numerical values as close to real case studies as possible. The benefits from implementing the hybrid service rather than conventional fixed route or fully flexible route deviation systems in different case studies are shown in Section 5.7. Section 5.8 attempts to simplify the optimized formula for station
spacing through making it independent from the formula for optimal flexible region at a location $x$. Section 5.9 includes sensitivity analysis for three important input values. Section 5.10 compares the performance of the analytical models with the results of a simulated case study. Finally, Section 5.11 presents the summary of this chapter.

### 5.1 Introduction

Current economic trends and population growth patterns pose challenges for the operation of fixed route systems, whereas demand responsive systems are often associated with high operating costs. There are a number of flexible transit services as an intermediate system between conventional fixed-route and demand responsive transit services, which leads to the improved efficiency of transit systems. Flexible route systems are preferable in areas with demand density that is too low to support fixed route systems. The ability of flexible transit services to adapt to customer demands also makes it suitable for serving passengers with a disability. Changing demand for transit services, including disruptions due to the COVID-19 pandemic, has created a need for alternative public transit systems that accommodate the need for user mobility and agency cost reduction associated with low transit demand.

Flexible transit systems can be designed under different service configurations according to service area characteristics and demand levels. It is thus important to properly identify the service areas where such systems may be effective, as well as the type of flexible transit that is most appropriate. There are many variations of flexible route services, and it is not uncommon for similar types of services to be referred to by different names, since individual transit agencies do not follow a standard naming practice. According to Koffman (2004), there are four elements of service design that could assist in defining the type of flexible service: a) where vehicles operate; b)
boarding and alighting locations; c) schedule; and d) advance notice requirements. The same study suggests that the flexible transit services can be broadly categorized as: 1) route deviation; 2) point deviation; 3) demand-responsive connector; 4) request stops; 5) flexible-route segments; and 6) zone route.

Existing literature includes flexible transit modeling approaches, such as analytical methods, simulation, empirical analysis, and stochastic processes. The study proposed here analyzes a hybrid fixed route transit system with elements of flexible services. More specifically, continuous approximation techniques are implemented to identify the optimal boundaries in a given corridor for providing flexible services in the form of route deviation. The proposed flexible hybrid service is compared with conventional fixed route service and fully flexible route deviation within the same corridor. The proposed model for flexible transit is expected to be beneficial in areas where the best transit solution lies between the fixed route and the full flexible route systems.

### 5.2 System description

The service area considered in this study is rectangular with length $L$ and width $W$. Vehicles are assumed to travel within the corridor on a rectilinear street network. The basic model of a fixed route transit service is a straight-line corridor in the middle of the service area, with one end being a major terminal station. A typical configuration for this network is given in Figure 5.1a. Demand in the corridor is assumed to follow a many-to-one pattern in which trips with uniformly distributed origins are all destined for the terminal and trips that originate at the terminal are destined for uniformly distributed points in the service area. The one-way demand in the corridor is the number of passengers trip origins per area per time, $Q$, which is assumed to be uniformly distributed over space and time. The vehicle average
speed, $V$, accounts for stopping times and traffic delays. Vehicles operate on uniform headway, $H$, and vehicles are assumed to be large enough that passenger capacity is not a binding constraint. The stop spacing, $S(x)$, can vary across the fixed route corridor as a continuous function of the distance from the start of the route $x$.


Figure 5.1: Examples of system configuration for a) conventional fixed route, b) flexible with route deviation, c) hybrid fixed with route deviation

Users are assumed to travel from a location within the service area to a terminal station, or vice-versa. The terminal station is assumed to connect the service area with a city center or other transportation hub. Thus, it is considered that passengers only board the vehicle as it moves towards the terminal station and they only alight in the opposite direction. Two types of users are analyzed:

- Curb-to-curb users - system users that request curb-to-curb service either for their pick-up or drop-off and will be served by a vehicle that is routed to the requested stop.
- Fixed stop users - system users that use only the fixed stops that are served by the flexible system.

Flexible services may involve only one or both the types of users presented above. Examples of curb-to-curb requests include users that are eligible for ADA paratransit or other passengers that want to avoid the efforts associated with accessing a fixed stop and waiting at a transit stop rather than their own private space. Such phenomena are expected to increase substantially during the ongoing COVID-19 pandemic, since public transit users aim to reduce their risk of infection to the extent possible. Alternatively, curb-to-curb requests could be assigned on a first-come-first-served basis to the first $a(\%)$ of trips requested, based on the number of users that can be served curb-to-curb during a single trip time.

The modeling approach presented in this study assumes that all users are served as they desire, either curb-to-curb or at fixed stops. Thus, the factors that could lead to rejection of service (e.g., vehicle seating capacity) are considered negligible. Both types of demand are perfectly inelastic, which means that they are not affected by the quality of service. This study focuses specifically on a model of flexible service using route deviation, as described in the following section.

### 5.3 Modeling route deviation

A vehicle starts its trip from the terminal station and serves customers in a given corridor at fixed stops or by deviating to serve the curb-to-curb demand, which makes up a fraction $a \in[0,1]$ of the total demand. The locations of fixed stops are defined in terms of the stop spacing at location $x, S(x)$. The curb-to-curb users are assumed to request their pick-ups or drop-offs with sufficient advanced notice that the vehicle routing can be scheduled and determined prior to dispatch. The route has a longitudinal length $L$, which is the length of the corridor. For each requested stop, the vehicle travels a lateral distance, $d$, to pick-up or drop-off the curb-to-curb requests and then the same distance, $d$, to return to the main route. The expected distance of a uniformly distributed requested stop from the main route is $W / 4$. The vehicle does not backtrack to serve curb-to-curb demand. The remaining $(1-a)$ portion of total demand is associated with passengers that walk to the nearest fixed stop and wait at that location for service. A typical configuration of a flexible system with route deviation is shown in Figure 5.1b.

The focus of this study is to optimize the operation of a transit system in order to identify when and where flexible service will be more beneficial for both agency and users. The resulting system is a hybrid system between a conventional fixed route and a flexible route deviation system. An example of such a system's configuration is given in Figure 5.1c. The red dashed line indicates the flexible region where the vehicles may deviate from the fixed corridor to serve the curb-to-curb requested demand. The width of the flexible area around a point $x$ along the fixed corridor is $A(x)$, where $A(x) \in[0, W]$. The expected deviation is $A(x) / 4$.

Here we present calculations for distributed demand and vehicle operations in a corridor heading toward the terminal. The reverse direction, with distributed desti-
nations for passengers heading away from the terminal, is symmetric. The number of passengers boarding each vehicle per unit distance traveled in the corridor is the product of the demand rate, the headway since the last vehicle, and the corridor width, $Q H W$. Of this total demand, the number of passengers with request stop service is $a Q H A(x)$, where the width of the flexible service area can vary as a function of the location in the corridor, $x$.

Vehicle distance and travel time can be calculated by integrating across the incremental vehicle distance and time required for the transit vehicle to traverse a distance $d x$ at any location $x$. The total distance and time required to traverse the corridor is obtained by integrating the incremental values over the length $L$. The one-directional value is then doubled to obtain the distance and travel time associated with a cycle of travel from the terminal back to the terminal.

The vehicle distance is the sum of longitudinal distance traveled along the corridor and the lateral distance traveled to serve each requested stop:

$$
\begin{equation*}
V M T=2 \int_{0}^{L}\left(1+a Q H A(x) \frac{A(x)}{2}\right) d x \tag{17}
\end{equation*}
$$

The first term is the longitudinal distance traveled per unit length of the corridor; the total longitudinal distance is $2 L$ per cycle. The second term is the product of the expected number of passengers with request stop service per unit length of the corridor and the expected lateral distance per request stop, which is twice $A(x) / 4$.

The cycle time, $C$, includes the travel times for the longitudinal and lateral travel at speed $V$. It also includes dwell time for three kinds of stops: the dwell time at fixed stops, $\tau^{f}$; the dwell time at requested stops, $\tau^{r}$; and the dwell time at the terminal station, $\tau^{t}$. Fixed stops have spacing $S(x)$, as defined above, so the expected number of fixed stops per unit length of corridor is $1 / S(x)$. The number of requested stops
per unit length of the corridor is the same as the expected number of passengers with request stop service, because each request trip is served individually. The vehicle stops once at the terminal. As a result, the cycle time is given by:

$$
\begin{equation*}
C=2 \int_{0}^{L}\left(\frac{1}{V}+a Q H A(x) \frac{A(x)}{2 V}+\tau^{f} \frac{1}{S(x)}+a Q H A(x) \tau^{r}\right) d x+\tau^{t} \tag{18}
\end{equation*}
$$

### 5.4 Modeling costs

The continuous approximation approach is adopted here to determine the optimal width of the flexible service area, $A(x)$, as well as the optimal spacing between fixed stops, $S(x)$. Both characteristics are treated as continuous functions of the distance, $x$, from the edge of the corridor. Specifically, $A(x)$ can be implemented as a continuous function, and $S(x)$ is approximated by a continuous function. Like the formulation for $V M T$ and $C$, the analysis is focused on the costs associated with the cycle of vehicle traversing the corridor from the terminal to the end and back.

### 5.4.1 Agency costs

The agency cost per vehicle cycle, $A C$, consists of three parts: costs attributed to vehicle distance traveled, costs attributed to vehicle hours of operation, and costs associated with the fleet size. Each of these costs is calculated by multiplying a cost coefficient by the corresponding value,

$$
\begin{equation*}
A C=a_{V M T} V M T+a_{V H T} V H T+a_{M} M \frac{H}{O} \tag{19}
\end{equation*}
$$

where $a_{V M T}$ is the cost per vehicle distance traveled, $a_{V H T}$ is the cost per vehicle time operated, $a_{M}$ is the daily capital cost per vehicle, and $M$ is the number of vehicles in the fleet. The vehicle distance traveled per cycle, $V M T$, and the cycle time, $C$, are
given by equations (17) and (18). The fleet size is considered to be constant for this analysis, so its cost must be spread over the number of vehicle cycle operated within a daily period of operations. If the daily operating hours are denoted by $O$ and the service headway is $H$, then there are $O / H$ vehicle cycles operated per day.

### 5.4.2 User costs

User costs include costs associated with walking, waiting, and riding as experienced by the users. Like the analysis of vehicle operations and agency costs, the user costs can be calculated by integrating the incremental user cost associated with each unit length across the corridor. As a result, the total daily user cost is the sum of these components, weighted by corresponding user cost coefficients: $a_{W K}$ for time spent walking per vehicle cycle, $W K$; $a_{W T}$ for time spent waiting per vehicle cycle, $W T$; and $a_{R}$ for time spent riding per vehicle cycle, $R$.

$$
\begin{equation*}
U C=a_{W K} W K+a_{W T} W T+a_{R} R \tag{20}
\end{equation*}
$$

The models for each of these components of time spent by users are presented in the subsections below.

### 5.4.2.1 Walking

Passengers that receive request stop service do not experience walking time, so the remaining demand $Q H(W-a A(x))$ per unit length of the corridor must walk to the nearest fixed transit stop. On average, this is $W / 4$ in the direction perpendicular to the main corridor and $S(x) / 4$ along the corridor. The walking speed is assumed to be $V_{W K}$. The total walking time for all users served in a vehicle cycle is thus given
by

$$
\begin{equation*}
W K=2 \int_{0}^{L}\left(Q H(W-a A(x)) \frac{W+S(x)}{4 V_{W K}}\right) d x \tag{21}
\end{equation*}
$$

### 5.4.2.2 Waiting

All transit users, either served curb-to-curb or at fixed stops, are expected to experience waiting time equal to half the headway. The total waiting time for all passengers served in a vehicle cycle is simply the product of the demand, $2 Q H W$, and the average waiting time, $H / 2$.

$$
\begin{equation*}
W T=Q H^{2} W \tag{22}
\end{equation*}
$$

### 5.4.2.3 Riding

The expected riding time is calculated based on the incremental riding time experienced by all passengers on board a vehicle as it traverses a unit length of the corridor at location $x$. The number of passengers onboard the vehicles is the cumulative number of passengers that have boarded since the beginning of the line. It is useful to think of this in terms of a vehicle trip from the edge of the corridor that starts empty and picks up passengers en route to the terminal. By the time the vehicle reaches location $x$, there are $Q H W x$ passengers onboard. Each of these passengers experiences travel time associated with longitudinal and lateral vehicle distance as well as loss time per fixed and requested stop. The incremental travel time per unit length of the corridor for all passengers is the product of $Q H W x$ and the incremental vehicle travel time, which is the integrand of equation (18). Therefore, total riding time for a vehicle cycle, $R$, has a similar structure to the expression for cycle time,
$C$.

$$
\begin{array}{r}
R=2 \int_{0}^{L} Q H W x\left[\frac{1}{V}+a Q H A(x) \frac{A(x)}{2 V}+\tau^{f} \frac{1}{S(x)}+a Q H A(x) \tau^{r}\right] d x  \tag{23}\\
+Q H W L \tau^{t}
\end{array}
$$

In order to estimate the total riding costs, the dwell time at the terminal should also be considered. For a fixed corridor of length $L$, there are $2 Q H W L$ passengers that each experience half of the dwell time at the terminal, $\tau^{t} / 2$.

### 5.4.3 Total weighted generalized costs

The total generalized for a day of flexible transit operations, $G C$, is the sum of agency costs, $A C$, and user costs, $U C$, weighted by $w_{A C}$ and $w_{U C}$, respectively.

$$
\begin{equation*}
G C=w_{A C} A C+w_{U C} U C \tag{24}
\end{equation*}
$$

This cost depends on the size of the flexible region, $A(x)$, and the fixed stop spacing, $S(x)$, which can be designed as functions of $x$.

The total daily generalized cost, $T G C$, is then calculated by multiplying the cost per cycle by the number of vehicle cycles that are operated in a day, $O / H$.

$$
\begin{equation*}
T G C=G C \frac{O}{H} \tag{25}
\end{equation*}
$$

The objective in this study is to minimize $T G C$ with respect to $A(x)$ and $S(x)$ for given $O$ and $H$ in order to achieve the optimal performance for the hybrid transit system studied here. The respective analysis is presented in the following section.

### 5.5 Optimal spacing and flexible region

Given that the duration and daily operations, $O$, and the service headway, $H$, are treated as exogenous values in this analysis, the minimization of $T G C$ is equivalent to minimizing $G C$. Thus, the optimization problem that this study addresses is the following:

$$
\begin{equation*}
\min _{A(x), S(x)} G C \tag{26}
\end{equation*}
$$

subject to

$$
\begin{align*}
& 0 \leq A(x) \leq W  \tag{27}\\
& 0 \leq S(x) \leq 2 L \tag{28}
\end{align*}
$$

The constraints on $A(x)$ ensures that the flexible region is always a subset of the corridor. The stop spacing is constrained to $2 L$, which would be the case if there were only a stop at the terminal, thereby forcing any customer that does not receive request stop service to walk all the way to their destination.

To facilitate the optimization, it is useful to note that in equations (17), (18), (21), (22), and (23), which are the inputs to equation (24), the decision variables, $A(x)$ and $S(x)$, only appear within the integrand. This integrand containing the terms with decision variables can be rewritten as:

$$
\begin{array}{r}
w_{A C}\left[a_{V M T}\left(\frac{a Q H}{2} A(x)^{2}\right)+a_{V H T}\left(\frac{a Q H}{2 V} A(x)^{2}\right)+\frac{\tau^{f}}{S(x)}+a Q H \tau^{r} A(x)\right]+ \\
w_{U C}\left[a_{W K}\left(\frac{Q H}{4 V_{W K}}(W-a A(x))(W+S(x))\right)\right]+  \tag{29}\\
w_{U C}\left[a_{R} Q H W x\left(\frac{a Q H}{2 V} A(x)^{2}\right)+\frac{\tau^{f}}{S(x)}+a Q H \tau^{r} A(x)\right]
\end{array}
$$

The value of the continuous approximation formulation is that we can now focus on
identifying the values of $A(x)$ and $S(x)$ that minimize the integrand at any $x$, and the results are functions that minimizes the integral, and thus the generalized cost.

Expression (29) is not quite separable with respect to $S(x)$ and $A(x)$, because the term associated with walking cost includes $(W-a A(x))(W+S(x))$, which combined both decision variables. This combined term prevents us from being able to derive a closed form analytical solution for the optimal values for $S(x)$ and $A(x)$ at any location $x, S^{*}(x)$ and $A^{*}(x)$. We note that expression (29) is convex in $S(x)$ if $A(x)$ is treated as given, and it is convex in $A(x)$ if $S(x)$ is treated as given. Therefore, a closed form for the optimal stop spacing at each location, $S^{*}(x)$, can be expressed in terms of $A(x)$ by solving the first order conditions for expression (29) with respect to $S(x)$; i.e., setting the first derivative equal to zero and solving for $S(x)$.

$$
\begin{equation*}
S^{*}(x)=2\left[\frac{V_{W K} \tau^{f}\left(w_{A C} a_{V H T}+w_{U C} a_{R} Q H W x\right.}{w_{U C} a_{W K} Q H(W-a A(x))}\right]^{0.5} \tag{30}
\end{equation*}
$$

Likewise, a closed form for the optimal size of the flexible service area at each location, $A^{*}(x)$, can be expressed in terms of $S(x)$ by solving the first order conditions with respect to $A(x)$.

$$
\begin{equation*}
A^{*}(x)=\frac{V}{V_{W K}} \frac{w_{U C} a_{W K}(W+S(x))-4 w_{U C} a_{R} Q H W x \tau^{r} V_{W K}-4 w_{A C} a_{V H T} \tau^{r} V_{W K}}{w_{U C} a_{R} Q H W x+w_{A C} a_{V M T} V+w_{A C} a_{V H T}} \tag{31}
\end{equation*}
$$

Equations (30) and (31) are directly applicable to cases where one of the two decision variables is exogenous. For example, equation (30) provides the optimal fixed stop spacing for a system in which an agency has already decided how big the flexible service area should be (e.g., $A(x)=1.5$ miles to satisfy minimum ADA requirements). Similarly, equation (31) defines the optimal size of the flexible service area for a transit agency that may not want to move the stop locations of a fixed route service that has
already been designed.
The more complex case is to optimize both decision variables, $S(x)$ and $A(x)$, simultaneously, because each depends on the other. A computational approach can be implemented to identify the fixed point solution satisfying equations (30) and (31). This can be solved substituting the expression for $S^{*}(x)$ in equation (30) into equation (31) to obtain an expression with only $A(x)$ terms. The optimal value, $A^{*}(x)$, that satisfies the equation can be identified by iterating through potential values of $A(x) \in(0, W)$ for increments of $x$. A numerical solution can be obtained quickly with a computer. Once $A^{*}(x)$ has been identified, $S^{*}(x)$ is given by equation (30).

Finally, it is necessary to confirm that the available fleet size, $M$, is sufficient for the designed service operation. Although it is theoretically possible to make $M$ a variable that depends on design variables, the reality is that flexible transit service in low density corridors typically operates at such long headways that only a small number of vehicles are ever needed. Therefore, $M$ is treated as an input parameter. The fleet size must be at least large enough to sustain the headway, $H$, with the cycle time, $C$.

$$
\begin{equation*}
M \geq \frac{C}{H} \tag{32}
\end{equation*}
$$

### 5.6 Numerical analysis

We now present a numerical analysis to illustrate application of the model to realistic corridors. Optimal values of $A^{*}(x)$ and $S^{*}(x)$ are calculated every 0.001 miles in order to provide a high-resolution representation optimized functions. The input values for the numerical examples presented here are summarized in Table 5.1.

The fundamental assumption for user costs is that walking should have higher cost coefficients than waiting and riding and the two latter are considered equal. Insights on the transit user cost coefficients can be found in Wardman (2004). The

Table 5.1: Input values

| Parameter | Value | Units |
| :---: | :---: | :---: |
| Percent of Curb-to-Curb Demand, $a$ | 0.5 | unitless |
| Fleet Cost Coefficient, $a_{M}$ | 100 | $\$ / \mathrm{veh}$ |
| Riding Cost Coefficient, $a_{R}$ | 10 | $\$ / \mathrm{veh}$ |
| VHT Cost Coefficient, $a_{V} H T$ | 20 | $\$ / \mathrm{veh} . \mathrm{hr}$ |
| VMT Cost Coefficient, $a_{V} M T$ | 20 | $\$ / \mathrm{veh} . \mathrm{mi}$ |
| Walking Cost Coefficient, $a_{W} K$ | 20 | $\$ / \mathrm{hr}$ |
| Waiting Cost Coefficient, $a_{W} T$ | 10 | $\$ / \mathrm{hr}$ |
| Vehicle Headway, $H$ | 1 | $\mathrm{hr} / \mathrm{veh}$ |
| Operational Hours, $O$ | 18 | $\mathrm{hr} / \mathrm{day}$ |
| Cruising Speed, $V$ | 25 | mph |
| Walking Speed, $V_{w} k$ | 3 | mph |
| Weighting Factor for AC, $w_{A} C$ | 1 | unitless |
| Weighting Factor for UC, $w_{U} C$ | 1 | unitless |
| Dwell Time at Fixed Stops, $\tau^{f}$ | 0.008 | $\mathrm{hr} / \mathrm{stop}$ |
| Dwell Time at Curb-to-Curb Stops, $\tau^{r}$ | 0.005 | $\mathrm{hr} / \mathrm{stop}$ |
| Dwell Time at Terminal Stop, $\tau^{t}$ | 0.010 | $\mathrm{hr} / \mathrm{stop}$ |

magnitudes considered here for agency costs are derived from existing literature for the paratransit services in New Jersey and Greater Boston Area, which are considered the worst case scenario, since demand responsive operations in large cities tend to be made more expensive by the high costs of labor. For more details on the agency cost coefficients the readers are referred to Rahimi et al. (2018) and Turmo et al. (2018).

Real-world flexible service areas where vehicles deviate their route to serve customers as needed can be identified in existing literature. In Zheng et al. (2018a) Route 289 in a suburban area of Zhengzhou City, China, is evaluated for an implementation of point and route deviation services. A single service vehicle is considered for a service area of $W=1$ mile and $L=3$ miles. The demand density ranges from 4 to $17 \mathrm{pax} / \mathrm{mi}^{2} / \mathrm{hr}$. The MTA Line 646 in Los Angeles County is used in several case studies of flexible system (Qiu et al., 2015; Zheng et al., 2018b). The service area has a width of $W=1$ mile and length of $L=10$ miles, with one operating
service vehicle. In Zheng et al. (2018b) demand ranging from 0.8 to $2.8 \mathrm{pax} / \mathrm{mi}^{2} / \mathrm{hr}$ is considered. In Qiu et al. (2015) slightly higher demand levels are considered for the same service area and a corridor of size $W=2$ and $L=5$ miles is evaluated. A third real-world case study for flexible systems is the Plymouth Area Link in Greater Attleboro Taunton Regional Transit Authority in Massachusetts, USA, which operates the Manomet/Cedarville Deviated Link, where two vehicles operate on a fixed corridor of $L \approx 8$ miles, with a headway $H=1$ hour, which is a common headway for such services. The vehicles are allowed to deviate to serve passengers within $3 / 4$ miles of the fixed route, indicating a service area of width $W=1.5$ miles.

In the remaining analyses, the magnitudes of $W, L$, and $Q$ are based on values in existing literature to investigate the implementation of the proposed method under different service area scenarios. For input values with no clear indications from existing literature, a sensitivity analysis is performed to investigate their impact on the proposed flexible transit service.

### 5.6.1 Optimal decision variables

Figure 5.2 shows the flexible region boundaries for $W \in\{1,2,3\}$ miles for a service area of length $L=10$ miles. Since $A^{*}(x)$ and $S^{*}(x)$ depend only on cumulative demand up to $x$, shorter corridors are represented by the same figures, just truncated to $L<10$. The horizontal line in the middle of each service area represents the fixed route corridor. Three cases of demand density per direction are investigated in this figure, $Q \in\{2.5,5,7.5\} \mathrm{pax} / \mathrm{mi}^{2} / \mathrm{hr}$. The three shaded areas represent the flexible regions in each case, colored with grey, blue, and red, respectively. For all $W$, the lower value of $Q$ leads to a greater flexible service region and the flexible region gets smaller as $W$ increases. Station locations are also shown for each demand density by black, blue, and red dots, respectively. The station spacing increases with $x$, because
greater vehicle occupancy increases the generalized cost of stopping. For more details on determining the station location from a continuous function of spacing between stations, the readers are referred to Wirasinghe et al. (1977).

(b)

(c)

Figure 5.2: Service area configuration for a) $W=1$, b) $W=2$, and c) $W=3$

### 5.6.2 Optimized system cost components

Figures 5.3 and 5.4 show that increasing $W, L$, and $Q$ increases all of the cost components. The costs associated with walking have the greatest impact, and the costs associated with VMT have the least impact. Fleet size costs shown in Figure 5.4 a present a step increase of one vehicle after $x=4$. The maximum fleet size for all scenarios investigated here is equal to two vehicles. Although fleet size costs and waiting costs are independent of the optimization process, their values offer insights to the relative magnitudes of the components of the generalized costs. In the case of VHT and VMT shown in Figure 5.4b and Figure 5.4c, it is apparent that the increase of $Q$ has a lower effect on costs, compared with the increase of $W$.

### 5.7 Comparison between fixed, hybrid and route deviation system costs

Table 5.2 compares the benefit of the optimized hybrid system with a fixed route and the fully flexible service. The agency cost components considered in optimizing the hybrid service are the VHT and VMT costs. Three corridor lengths are considered, $L \in\{3,5,10\}$ miles. The percent benefit, $B^{S}(\%)$, from implementing hybrid transit $(H T)$ is

$$
\begin{equation*}
B^{S}(\%)=\frac{C^{S}-C^{H T}}{C^{S}} 100 \% \tag{33}
\end{equation*}
$$

where $C^{S}$ represents the cost of system $S$, with $S \in(F R, R D)$ for fixed route and route deviation, respectively. Fixed route service has the lower agency costs among all three systems, so the benefit of hybrid service is negative. Route deviation has the highest agency costs, so the benefit of hybrid service is positive.

The user benefits associated with the hybrid system compared to the fixed route

(a)

(b)


|  | $\mathrm{W}=1, \mathrm{Q}=2.5$ |
| :---: | :---: |
|  | $\mathrm{W}=2, \mathrm{Q}=2.5$ |
|  | $\mathrm{W}=3, \mathrm{Q}=2.5$ |
|  | $\mathrm{W}=1, \mathrm{Q}=5$ |
|  | $\mathrm{W}=2, \mathrm{Q}=5$ |
|  | $\mathrm{W}=3, \mathrm{Q}=5$ |
|  | $\mathrm{W}=1, \mathrm{Q}=7.5$ |
|  | $\mathrm{W}=2, \mathrm{Q}=7.5$ |
|  | $\mathrm{W}=3, \mathrm{Q}=7.5$ |

(c)

Figure 5.3: Daily user costs of a) walking, b) waiting, and c) riding

(a)

(b)


| - | $\mathrm{W}=1, \mathrm{Q}=2.5$ |
| :--- | :--- |
| - | $\mathrm{W}=2, \mathrm{Q}=2.5$ |
| - | $\mathrm{W}=3, \mathrm{Q}=2.5$ |
| -- | $\mathrm{W}=1, \mathrm{Q}=5$ |
| -- | $\mathrm{W}=2, \mathrm{Q}=5$ |
| -- | $\mathrm{W}=3, \mathrm{Q}=5$ |
| $\square$ | $\mathrm{~W}=1, \mathrm{Q}=7.5$ |
| - | $\mathrm{W}=2, \mathrm{Q}=7.5$ |
| - | $\mathrm{W}=3, \mathrm{Q}=7.5$ |

Figure 5.4: Daily agency costs of a) fleet size, b) VHT, and c) VMT
are shown in Figure 5.5. The user costs of walking and riding affect the optimization process and are considered here. The user benefits range from 0 to $35 \%$ for all combinations of service areas and demand densities. Smaller service areas and lower demand densities lead to greater user benefits from the implementation of hybrid systems compared with fixed route. Comparing with full route deviation systems, the implementation of the hybrid transit has a user benefit of up to $\sim 80 \%$, with some cases having a small loss (i.e., $\leq 5 \%$ for small areas and low demand densities). This loss is due to the effect of agency costs in the optimization process for the hybrid service.

Under the COVID-19 pandemic and the resulting decrease in transit ridership, it is noteworthy that for any one of the service areas studied here, there is a significant increase in users' benefits with a hybrid system as the demand density decreases. The hybrid system is also more beneficial for users than full route deviation systems, especially for $W>1$. Finally, the agency loss associated with the hybrid system compared to conventional fixed route is slightly affected by falling demand. For these reasons, the proposed hybrid system has the potential for many beneficial applications in low density communities or in areas where demand has dropped significantly due to the pandemic.

### 5.7.1 Optimal percent flexibility

The percent flexibility for the services of a given area can be calculated considering the results of implementing equation (31) and the dimensions of the service area, as shown in equation (34). Thus, it can be defined as the percentage of the service area that is covered by the flexible region. For the inputs presented in Table 5.1, the optimal percent flexibility for a service area of length $L=10$ miles is shown in Figure 5.6, for different values of headway and service area width. This figure shows


Figure 5.5: Percent user benefits from implementing hybrid transit instead of fixed route for a) $Q=2.5$, b) $Q=5$, c) $Q=7.5$, and route deviation for d) $Q=2.5$, e) $Q=5$, and f) $Q=7.5$
that lower headways, demand densities and widths of service area lead to greater flexibility. Such graphs can be easily constructed and provide guidance for transit agencies to make decisions such as whether or not to implement flexible services in a corridor.

$$
\begin{equation*}
\text { flexibility }(\%)=\frac{\int_{0}^{L} A(x) d x}{W L} * 100 \tag{34}
\end{equation*}
$$

Table 5.2: Percent agency benefit from implementing the optimized hybrid transit instead of Fixed Route (FR) and Route Deviation (RD)

| $Q=2.5$ (pax/sq.mi/hr) |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Benefits(\%) | $\mathrm{W}=1$ |  |  | $\mathrm{W}=2$ |  |  | $\mathrm{W}=3$ |  |  |
|  | $\mathrm{L}=3$ | $\mathrm{L}=5$ | $\mathrm{L}=10$ | $\mathrm{L}=3$ | $\mathrm{L}=5$ | $\mathrm{L}=10$ | $\mathrm{L}=3$ | $\mathrm{L}=5$ | $\mathrm{L}=10$ |
| FR | -26.3 | -19.6 | -12.0 | -52.1 | -35.9 | -20.4 | -77.7 | -52.4 | -29.3 |
| RD | 11.4 | 18.0 | 25.5 | 44.3 | 52.3 | 60.0 | 63.2 | 70.1 | 76.3 |
| $Q=5$ (pax/sq.mi/hr) |  |  |  |  |  |  |  |  |  |
| Benefits(\%) | $\mathrm{W}=1$ |  |  | $\mathrm{W}=2$ |  |  | $\mathrm{W}=3$ |  |  |
|  | $\mathrm{L}=3$ | $\mathrm{L}=5$ | $\mathrm{L}=10$ | $\mathrm{L}=3$ | $\mathrm{L}=5$ | $\mathrm{L}=10$ | $\mathrm{L}=3$ | $\mathrm{L}=5$ | $\mathrm{L}=10$ |
| FR | -28.8 | -19.8 | -11.2 | -53.5 | -35.5 | -19.5 | -78.7 | -51.9 | -28.4 |
| RD | 31.7 | 38.7 | 45.6 | 65.3 | 71.1 | 76.2 | 79.0 | 83.3 | 87.0 |
| $Q=7.5$ (pax/sq.mi/hr) |  |  |  |  |  |  |  |  |  |
| Benefits(\%) | $\mathrm{W}=1$ |  |  | $\mathrm{W}=2$ |  |  | $\mathrm{W}=3$ |  |  |
|  | $\mathrm{L}=3$ | $\mathrm{L}=5$ | $\mathrm{L}=10$ | $\mathrm{L}=3$ | $\mathrm{L}=5$ | $\mathrm{L}=10$ | $\mathrm{L}=3$ | $\mathrm{L}=5$ | $\mathrm{L}=10$ |
| FR | -28.8 | -19.2 | -10.5 | -53.2 | -34.8 | -19.0 | -78.2 | -51.1 | -28.0 |
| RD | 45.0 | 51.3 | 57.3 | 74.9 | 79.3 | 83.1 | 85.4 | 88.5 | 91.0 |

### 5.8 Optimization of station spacing based on fixed route and route deviation systems

We now consider the effect of the the size of the flexible region, $A(x)$ on the optimal fixed stop spacing and the costs of the system. Specifically, we are interested in the two extreme cases: $A(x)=0$, which is a fixed route system, and $A(x)=W$, which is a route deviation system. Although $S^{*}(x)$, as calculated in equation (30) is sensitive to the value of $A(x)$ used, $A^{*}(x)$ from equation (30) is not greatly affected whether $S^{*}(x, A(x)=0)$ or $S^{*}(x, A(x)=W)$ is used. Figure 5.7a shows that the optimized fixed stop spacing for the hybrid transit, $S_{H T}^{*}(x)$ lies between the optimized station spacings for fixed route, $S_{F R}^{*}(x)$, and route deviation, $S_{R D}^{*}(x)$. In this case, $S_{H T}^{*}(x)$ overlaps $S_{R D}^{*}(x)$ for locations $x$ where $A^{*}(x)=W$ and then moves towards $S_{F R}^{*}(x)$. Although the optimized spacings differ depending on the type of service, the optimized flexible regions that result from implementing each of the three optimal spacings are


$$
-\mathrm{W}=1 \quad \mathrm{~W}=2 \quad \mathrm{~W}=3
$$

(a)



- $\mathrm{W}=1$ - $\mathrm{W}=2$ - $\mathrm{W}=3$
(b) $-W=1-W=2-W=3$
(c)

Figure 5.6: Optimal percent flexibility of a service area with length $\mathrm{L}=10$ (miles) and headway equal to a) 0.5 hours/veh, b) 1 hours/veh, and c) 1.5 hours/veh
very similar, as shown in Figure 5.7b.
The difference in cost is more important than that difference in the design variables, because it is the generalized cost of the system that we seek to minimize. The percent change in cost for implementing either the fixed route or full route deviation system relative to the optimized hybrid system is given by

$$
\begin{equation*}
\Delta(\%)=\frac{C\left(S_{T}^{*}\right)-C\left(S_{H T}^{*}\right)}{C\left(S_{H T}^{*}\right)} 100 \% \tag{35}
\end{equation*}
$$

where $C\left(S_{T}^{*}\right)$ is the cost of implementing the optimal station spacing for system $T$,


Figure 5.7: Optimized decision variable of a) station spacing, and b) flexible region for a corridor with $W=2, L=10$, and $Q=5$
with $T \in\{F R, R D\}$, and $C\left(S_{H T}^{*}\right)$ the cost of implementing the optimal spacing for the hybrid service. The costs considered in this analysis are those that are included in the optimization process, namely walking, riding, $V H T$ and $V M T$ costs.

This analysis shows that the effect of different optimized station spacings on the user and agency costs is always small; less than $2 \%$ for the cases presented in Figure 5.8. As a result, it is acceptable to approximate the joint optimization of $A(x)$ and $S(x)$ by implementing equations (30) and (31) independently. Although the optimized station spacing might differ based on what system is considered in its optimization, the optimal flexible region and the resulting operating costs are not severely impacted.

### 5.9 Sensitivity analysis

### 5.9.1 Effect of headway

The headway of service has a significant effect on the system design and cost, because it determines the number of passengers served by each vehicle and the number

(a)

(c)

(b)

> S* $^{\text {for }}$ FR $-W=1$ S $^{\star}$ for FR $-W=2$ S $^{\star}$ for $\mathrm{FR}-\mathrm{W}=3$ S $^{\star}$ for RD $-W=1$ S $^{\star}$ for RD $-W=2$ S $^{\star}$ for RD $-W=3$

Figure 5.8: Percent cost difference for a) $Q=2.5$, b) $Q=5$, and c) $Q=7.5$
of vehicles needed in the fleet. To facilitate the analysis in this study, $H=1 \mathrm{hr}$ was used as an exogenous value in accordance with many real-world flexible systems. We now consider the effect of varying $H \in(0.1,2) \mathrm{hrs}$ on the optimized design variables and the resulting costs. Figure 5.9a and Figure 5.9b show the effect of $H$ on $S^{*}(x)$ and $A^{*}(x)$ for a service area with $W=2$ miles, $L$ up to 10 miles, and $Q=5 \mathrm{pax} / \mathrm{mi}^{2} / \mathrm{hr}$. Greater $H$ leads to smaller the flexible regions and shorter stop spacing as the system
more closely resembles fixed route.
Table 5.3 shows that as $H$ increases, daily user costs increase significantly for any percentage of demand served curb-to-curb, $a$. Lower $H$ is associated with greater impact of $a$ on user costs. The costs included in Table 5.3, refer to all types of user and agency costs in order to offer an overview of the overall cost magnitudes.

### 5.9.2 Effect of flexible service demand

The percentage of demand receiving request stop service within the flexible region, $a$, affects the distance and time traveled to serve the requested stops. Figure 5.9c shows that $S^{*}(x)$ increases with $a$. For the extreme case of $a=1.00$, the fixed spacing tends to infinity for $x \leq 0.22$ miles, $A^{*}(x)=W$ in this range so no passengers use fixed stops. Further along the corridor, $A^{*}(x)$ drops, increasing the number of passengers using fixed stops. Grey lines in Figure 5.9c and Figure 5.9d are associated

Table 5.3: User and agency costs per day for different headways and percent of demand served curb-to-curb

| $\begin{gathered} \mathrm{H}=0.5 \\ h r / v e h \end{gathered}$ | User Costs (\$/day) |  |  | Agency Costs (\$/day) |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\mathrm{a}=0.25$ | $\mathrm{a}=0.5$ | $a=0.75$ | $\mathrm{a}=0.25$ | $\mathrm{a}=0.5$ | $\mathrm{a}=0.75$ |
| $\mathrm{L}=3 \mathrm{mi}$ | 7,599.9 | 7,260.9 | 6,908.1 | 722.3 | 832.0 | 947.8 |
| $\mathrm{L}=5 \mathrm{mi}$ | 13,984.5 | 13,591.2 | 13,183.4 | 968.8 | 1,087.6 | 1,212.7 |
| $\mathrm{L}=10 \mathrm{mi}$ | 33,509.7 | 33,066.5 | 32,608.2 | 1,738.5 | 1,864.4 | 1,996.6 |
| $\mathrm{H}=1$ | User Costs (\$/day) |  |  | Agency Costs (\$/day) |  |  |
| hr/veh | $\mathrm{a}=0.25$ | $\mathrm{a}=0.5$ | $\mathrm{a}=0.75$ | $\mathrm{a}=0.25$ | $\mathrm{a}=0.5$ | $a=0.75$ |
| $\mathrm{L}=3 \mathrm{mi}$ | 10,435.2 | 10,246.7 | 10,053.2 | 471.5 | 529.4 | 589.5 |
| $\mathrm{L}=5 \mathrm{mi}$ | 18,659.78 | 18,458.4 | 18,252.0 | 792.9 | 852.8 | 914.9 |
| $\mathrm{L}=10 \mathrm{mi}$ | 42,730.3 | 42,525.6 | 42,315.9 | 1,375.7 | 1,436.4 | 1,499.3 |
| H | User Costs (\$/day) |  |  | Agency Costs (\$/day) |  |  |
| hr/veh | $\mathrm{a}=0.25$ | $\mathrm{a}=0.5$ | $\mathrm{a}=0.75$ | $\mathrm{a}=0.25$ | $a=0.5$ | $a=0.75$ |
| $\mathrm{L}=3 \mathrm{mi}$ | 13,192.0 | 13,064.4 | 12,934.2 | 483.6 | 522.5 | 562.6 |
| $\mathrm{L}=5 \mathrm{mi}$ | 23,226.6 | 23,096.3 | 22,963.4 | 764.0 | 803.4 | 844.1 |
| $\mathrm{L}=10 \mathrm{mi}$ | 51,800.2 | 51,669.9 | 51,537.1 | 1,452.2 | 1491.7 | 1,532.4 |

with increments of $a$ from 0 to 1 , with a step of 0.1 . Figure 5.9 d shows that the optimal flexible region is very insensitive to $a$. Only when $a=0.00$ does it have no impact on costs. Therefore, advanced knowledge of the percent of users served with request stops is not necessary for identifying $A^{*}(x)$ is due to $S^{*}(x)$.

### 5.9.3 Effect of cost weights

These weights can control the relative effect that agency costs and user costs have on the optimal values for the two decision variables. Figure 5.9e and Figure 5.9 f show the effects of changing user cost weights from 0.1 to 1 with a step of 0.1 and 1 to 10 with a step of 1 . When a cost weight is examined the other is considered equal to one.

Figure 5.9 e shows that station distance is decreased as user costs are weighed more heavily against agency costs. Intuitively, this could be attributed to walking costs, which are reduced as user costs have a higher impact on the total generalized costs. The change in $S^{*}(x)$ is greater for $0.1<w_{U C}<1$ and much lower for $1<w_{U C}<10$. At $x \approx 0.20$ miles, the lines that correspond to $w_{U C}>1$ overlap, indicating that station spacing is independent of the weight of user costs. This location is the point where the optimal flexible region boundaries reach their maximum value (i.e., $W=2$ in this case). At some locations $x$, the optimal value for $A^{*}(x)$ is bounded by the feasibility condition that $A^{*}(x) \leq W$ and the optimal spacing is estimated based on this bounded value of $A^{*}(x)$. These points are at $x=0.12$ for $w_{U C}=1$ and at $x=0.41$ for $w_{U C}=10$. For $w_{U C}=0$, the optimal value for station spacing goes to infinity. Similarly, for $w_{U C}=0$, the optimal flexible region boundaries go to 0 . Figure 5.9 f shows the increase of flexible region boundaries as the weight of user costs increase. Again, for $0.1<w_{U C}<1$, the boundaries present a greater rate of increase compared to the respective changes in $1<w_{U C}<10$.


Figure 5.9: Optimal decision variable of a) $S^{*}(x)$ for various headways, $\left.H, \mathrm{~b}\right) A^{*}(x)$ for various headways, $H$, c) $S^{*}(x)$ for various percentages, $a$, d) $A^{*}(x)$ for various percentages, $a$, e) $S^{*}(x)$ for various weights, $\left.w_{U C}, \mathrm{f}\right) A^{*}(x)$ for various weights, $w_{U C}$

The effect of the agency cost weight, $w_{A C}$, on the two decision variables is the opposite of the ones described for $w_{U C}$ sensitivity analysis. More specifically, an
increase in $w_{A C}$ leads to an increase in the station spacing and decrease in the flexible region boundaries. At the same locations $x$ as in the case of $w_{U C}$, there are overlaps between the lines of station spacing when $A^{*}(x)$ is bounded by $W$. The overlap in this case occurs when the agency costs are undervalued. Overall, undervaluing the agency costs has a smaller effect on the two optimized decision variables than overvaluing it. However, undervaluing the weight of user costs has a greater effect on the two decision variables than overvaluing them, even if this overvalue is as great as 10 times up.

### 5.10 Simulation

The simulation process adopted in this study is developed using the R programming language and aims to evaluate the assumptions made for the development of the analytical model. The output of the simulation algorithm is the scheduling of vehicles in terms of times of arrival at the fixed and curb-to-curb stops, as well as the costs that result from their operation. The demand is generated as a Poisson process. The algorithm serves curb-to-curb passengers following a first-come, first-served pattern and the vehicles don't backtrack. The input data considered in the simulation process are the ones used for the analytical model implementation and are summarized in Table 5.1.

### 5.10.1 Optimal flexible region and station spacing

The optimal station spacing and the optimal flexible region are estimated based on the respective analytical models presented in equations (30) and (31). The demand density value required for the calculations is the expected demand density for the service area during a given day. The following subsection describes in detail how the demand is generated in this simulation.

### 5.10.2 Simulated demand

Demand is simulated as a Poisson process with random trip requests occurring throughout the day and across the service corridor according to the average demand density. Every trip request includes the following information:

- x-axis coordinates of trip's origin (or destination) bounded by the length of the service corridor, such that $x \in[0, L]$
- y-axis coordinates of trip's origin (or destination) bounded by the width of the service area, such that $y \in[0, W]$
- time of the request, $t$, expressed in minutes from the beginning of the operation period, such that $t \in[0,60 O]$, where $O$ denotes the operational hours per day

The next step is to identify which of the generated trip requests lie within the flexible region borders. From the eligible trips, $a \%$ are randomly identified to be served curb-to-curb. The percentage of trips served curb-to-curb is assumed to be a constant number throughout the day, and this may represent the percentage of customers that are eligible for curb-to-curb service (e.g., passengers with disabilities, senior citizens, etc.). This demand simulation process is performed for each direction separately.

### 5.10.3 Simulated scheduling

The scheduling process starts with considering the first cycle of service for a vehicle to traverse the length of the corridor and back. The duration of this cycle indicates the number of vehicles that will be needed to maintain a constant headway, $H$, as shown in equation (36) (Vuchic, 2007). For instance, if the first cycle time is greater than $H$, then a second vehicle is needed. The number of required vehicles will then
determine the times that each one of them starts from the terminal. For example, in the case of $n$ operating vehicles, the first will begin from the terminal every $n H$ hours starting at time 0 , whereas the second vehicle will begin from the terminal every $n \mathrm{H}$ hours starting at time $H$ hours.

$$
\begin{equation*}
n_{v e h}=\left\lceil\frac{C}{H}\right\rceil \tag{36}
\end{equation*}
$$

An alternative approach, especially in the cases that first cycle duration is close to the headway (or integer multipliers of it), is to account for many simulated cases of first run duration in order to calculate a confidence interval. More specifically, multiple simulations of the first run give a sample of first run duration with the respective average value and standard deviation which can lead to the maximum bound of the confidence interval considering t-distribution. Demand density per direction can be generated through Poisson distribution for each run. The construction of such confidence intervals is described later in this dissertation in Section 5.10.5.

After the starting times for each vehicle are determined, each vehicle is considered separately and the algorithm "visits" every location $x$ on the service corridor considering an appropriate space step $\Delta x$, starting from direction 1 and then considering direction 2. The time that each vehicle, $m \in[1, M]$, is at a location $x, t_{x}^{m}$, can be calculated by the time that it was at location $x-1, t_{x-1}^{m}$, plus the travel time $\Delta x / V$ needed to reach location $x$, as shown below.

$$
\begin{equation*}
t_{x}^{m}=t_{x-1}^{m}+\frac{\Delta x}{V} \tag{37}
\end{equation*}
$$

Then the algorithm checks whether there are requested curb-to-curb trips in the interval $\Delta x$ or not. If there is a trip request from a passenger $p$ at location $x$, then the algorithm checks if the requested time of service, $t_{p}^{r}$, is earlier than the time that
is needed for the vehicle to deviate and serve the request, $t_{p}^{d}$, assuring that the service is not before the requested time. If the curb-to-to-curb trip can be served, then the vehicle deviates to serve it and the time that vehicle departs location $x$ is updated accordingly. The service criterion is shown in equation (38).

$$
\begin{equation*}
t_{p}^{r} \leq t_{x-1}^{m}+\frac{\Delta x}{V}+t_{p}^{d} \tag{38}
\end{equation*}
$$

The next step is to take into account whether there is a fixed stop at this location or not. If there is a fixed stop, then it is assumed to be served after the curb-to-curb stops at this location. It is also assumed that if there is more than one curb-to-curb request to be served at the same time, then the vehicle serves them in order of request. Each trip in the simulation is served independently, which means that the vehicle deviates from the service corridor to serve the curb-to-curb trip and then drives back to the service corridor. Even though trips served consecutively at the same location could have combined service (e.g., after leaving the first trip's location $\left(x, y_{1}\right)$ the vehicle could move to the next curb-to-curb trip's location $\left.\left(x, y_{2}\right)\right)$, this is not done due to consistency with the analytical model. This will only make a difference when the demand density for curb-to-curb trips is very high.

The scheduling process is repeated for every vehicle for as long as the starting time from the terminal (i.e., the starting time of a new cycle) is lower than $O$. If a new cycle commences, then it will be completed, even if the operation time exceeds $O$.

### 5.10.4 Simulated costs

The main outputs of this algorithm are the user and agency costs that result from the simulation. These costs are then compared with the analytical estimations to
evaluate their performance.

### 5.10.4.1 Walking costs

The algorithm inputs include the exact origin (or destination) locations of the simulated users of the system in the form of coordinates, $\left(x_{p}, y_{p}\right)$. Thus, the distance between every passenger's, $p$, origin (or destination) location and the location of fixed stops can reveal which fixed stop is associated with the shortest walking time. This is the station that the passenger is assumed to use for service. The distance, $d_{p, s}$, between every passenger's origin (or destination) location and a fixed stop, $s$, with coordinates $\left(x_{s}, y_{s}\right)$ is calculated from:

$$
\begin{equation*}
d_{p, s}=\left|x_{p}-x_{s}\right|+\left|y_{p}-y_{s}\right| \tag{39}
\end{equation*}
$$

For clarification, the direction that all passengers alight, the closest stop to the passenger's destination is the one that they alight. In the opposite direction, the closest stop is the stop that the passengers will wait until they board. Passengers served curb-to-curb have zero walking costs.

The sum of all passengers' walking times multiplied by the cost coefficient results in the simulated walking costs, as shown in equation (40).

$$
\begin{equation*}
W K=a_{w k} \sum_{p=1}^{P} d_{p, s^{\prime}} \tag{40}
\end{equation*}
$$

where $P$ is the total number of passengers and $s^{\prime}$ is the closest stop.

### 5.10.4.2 Waiting costs

After identifying the station at which each passenger is served, $s^{\prime}$, the next step is to identify the first vehicle, $m^{\prime}$, that serves this station right after the passenger's request time. This identification is performed by simply checking the output of the scheduling process as described in Section 5.10.3. More specifically, in this part of the algorithm the code checks what is the smallest positive time difference between the time that each vehicle, $m$, arrives at the fixed stop location of interest and the requested time. The vehicle and the time of arrival that correspond to this value are the service vehicle, $m^{\prime}$, and the service time for the passenger, $t_{p}^{s}$, respectively.

The difference between the trip request time, $t_{p}^{r}$, and the time of service, $t_{p}^{s}$, is the respective waiting time, $t_{p}^{w}$ for the passenger, $p$, as shown in equation (41).

$$
\begin{equation*}
t_{p}^{w}=\left|t_{p}^{r}-t_{p}^{s}\right| \tag{41}
\end{equation*}
$$

For those waiting for a vehicle to board, the code assumes that they arrive at the fixed stop at the requested time and they wait for the first vehicle to arrive. The time of service is the time of their boarding. For those who alight, the code assumes that they alight at the time that the first vehicle arrives at the closest stop to their destination, which is some time after the request time. The time of service here is the time they alight. For passengers served curb-to-curb, the waiting time is calculated in a similar manner, considering the time they are served at the requested locations.

The sum of all passengers' waiting times multiplied by the cost coefficient gives the simulated waiting costs, as shown below.

$$
\begin{equation*}
W T=a_{w t} \sum_{p=1}^{P} t_{p}^{w} \tag{42}
\end{equation*}
$$

### 5.10.4.3 Riding costs

Here we consider direction 1 as the direction that passengers board at the terminal and alight at distributed destinations, and direction 2 as the direction that passengers board at distributed origins and alight at the terminal. Regarding direction 1, the code identifies the last time that the vehicle of service, $m^{\prime}$, departed the terminal before the passenger's request time, $t_{m^{\prime}}^{t, 1}$. In the opposite direction, the passengers are assumed to board the service vehicle, so the code identifies the first time that the service vehicle arrives at the terminal right after the passenger's request time, $t_{m^{\prime}}^{t, 2}$. Considering that all passengers boarded their vehicle on time, the riding time in both cases is the difference between the boarding and alighting times, as shown in equations (43) and (44) for direction 1 and 2, respectively.

$$
\begin{align*}
& r_{p}^{1}=t_{m^{\prime}}^{t, 1}-t_{p}^{s}  \tag{43}\\
& r_{p}^{2}=t_{p}^{s}-t_{m^{\prime}}^{t, 2} \tag{44}
\end{align*}
$$

The sum of all passengers' riding times multiplied by the cost coefficient gives the simulated riding costs.

$$
\begin{equation*}
R=a_{r}\left(\sum_{p=1}^{P_{1}} r_{p}^{1}+\sum_{p=1}^{P_{2}} r_{p}^{2}\right) \tag{45}
\end{equation*}
$$

where $P_{1}$ is the number of passengers in direction 1 and $P_{2}$ the number of passengers in direction 2.

### 5.10.4.4 Fleet costs

The number of vehicles, $M$, required to serve the daily demand is determined
as described in Section 5.10.3. The respective fleet size costs are then calculated by $a_{M} M$.

### 5.10.4.5 VHT costs

In order to measure the VHT, the code considers the duration of each cycle performed by all vehicles. The calculation is the following:

$$
\begin{equation*}
V H T=\sum_{m=1}^{M} \sum_{c=1}^{C_{m}} t_{c}^{m} \tag{46}
\end{equation*}
$$

where $t_{c}^{m}$ is the duration of cycle $c$ of vehicle $m$ and, $C_{m}$ is the number of cycles for vehicle $m$.

The sum of all cycles duration multiplied by the cost coefficient leads to the simulated VHT costs:

$$
\begin{equation*}
C_{V H T}=a_{v h t} V H T \tag{47}
\end{equation*}
$$

### 5.10.4.6 VMT costs

The total miles traveled by the vehicles in this simulation depend on the number of vehicles, $M$, the length of the fixed corridor, $L$, the number of cycles that each vehicle performs, $C_{m}$, and the distance needed to deviate and serve each curb-to-curb request, $\Delta y_{p}^{d}$, as shown below:

$$
\begin{equation*}
V M T=2 L \sum_{m=1}^{M} C_{m}+\sum_{p=1}^{P^{\prime}} 2 \Delta y_{p}^{d} \tag{48}
\end{equation*}
$$

where $P^{\prime}$ is the number of trips served curb-to-curb. The respective costs are:

$$
\begin{equation*}
C_{V M T}=a_{v m t} V M T \tag{49}
\end{equation*}
$$

### 5.10.5 Comparison between simulation and analytical modeling

Figure 5.10 presents the comparison between the analytical model and the confidence interval including lower bounds (LB) and upper bounds (UB) that result from running 50 simulations. The confidence interval, $C I$, is calculated by

$$
\begin{equation*}
C I=\left(\bar{X}-t \frac{s . d .}{\sqrt{N}}, \bar{X}+t \frac{s . d .}{\sqrt{N}}\right) \tag{50}
\end{equation*}
$$

where $\bar{X}$ represents the mean value of each simulated cost component, s.d. is the standard deviation, $N$ is the sample size and $t$ is the t-distribution value. In this case, the sample size is equal to 50 with 49 degrees of freedom, so $t=2.01$.

The case study considered here is $W=2(\mathrm{mi}), L=10(\mathrm{mi})$ and $Q=5$ ( $\mathrm{pax} / \mathrm{mi}^{2} / \mathrm{hr}$ ) per direction, and the operational hours are equal to 8 . All other input values are as in Table 5.1. It is reminded that the system's flexible region and station spacing are fixed in each run and they are optimized for the expected demand density per direction, $Q=5\left(\mathrm{pax} / \mathrm{mi}^{2} / \mathrm{hr}\right)$. The simulation, however, allows in each run the demand density per direction to be determined by Poisson. The results shown in Figure 5.10 do not include the first and last cycle of the day, to account for the algorithm's "warm-up" and "cool-down" times.

As shown in Figure 5.10, for this case study it is confirmed that the analytical model's results are always statistically equivalent to the simulation, since the analytical values are always within the simulation confidence intervals. The same holds for most case studies investigated here, but there are cases that the analytical value is out
of the confidence interval. Such deviations can be attributed to differences between the two methodologies. Such differences refer, for example, to the stochasticity associated with the simulated demand distribution and to the continuous approximation for variables that should be integer in reality (e.g., number of stations).


Figure 5.10: Comparison between analytical model and simulation costs of a) users and b) agency

### 5.11 Summary

This chapter proposes a method for optimizing station spacing and flexible region boundaries for a hybrid transit service. These outputs are independent from the total
length, $L$, of the service area. The user costs associated with waiting and the agency costs associated with fleet size do not affect the optimized design variables. The fleet size is calculated indirectly based on the decision variables. As the respective figures reveal, the station spacing becomes greater as the distance $x$ from the edge of the fixed corridor increases (i.e., as the distance from the terminal decreases). For all cases of $W$, the lower value of $Q$ leads to a greater flexible service region and the flexible region tends to zero sooner as the service area width, $W$ increases.

When comparing the proposed hybrid system with fixed route systems, the greatest area and demand density considered here correspond to the lowest benefit from the hybrid transit, whereas the smallest area with the lowest demand density leads to the greatest benefit. These results confirm that the fixed route service is more beneficial for the operations in larger areas with higher demand, rather than in small areas of low demand, such as suburban and rural communities. It is noteworthy that since the resulting optimized flexible region narrows with $x$, the respective benefits or costs always decrease as $L$ increases. Although the optimized station spacing might differ based on the type of system that is considered in its optimization, the optimal flexible region and the resulting operating costs are not severely impacted. Thus, a service area could switch from fixed route or full deviation to hybrid service within a day, adjusting to any level of demand and maintaining the same station spacing and infrastructure without negative impacts on the operational costs.

The sensitivity analyses performed in this study investigated the effects of different input values on the model performance. For the input values considered in this study, the percent demand of users served curb-to-curb, $a$, has a very small effect on the optimal flexible region boundaries, which becomes even less important as $x$ increases. Thus, a lack of advance knowledge of $a$ is not detrimental when optimizing the design of the proposed hybrid transit service. Regarding the headway, the lowest value
considered here has the greatest decrease of user costs as $a$ increases from 0.25 to 0.75 . Thus, transit services where low headways are maintained should expect greater user benefit from implementing the proposed hybrid system. Finally, a simulation analysis verified the results of the analytical models with randomly distributed origins and destinations.

## 6 CONCLUSIONS, TRANSFERABILITY AND FUTURE EXTENSIONS

This Chapter aims at presenting the objectives, methods, results and conclusions from the research associated with this dissertation. Comments on the transferability of the dissertation outputs and future extensions are included here. This dissertation addresses challenges and opportunities associated with transit demand, in order to improve transit efficiency through the use of optimization techniques and technology. The analysis presented here includes all three types of public transit systems, namely fixed route, on-demand and flexible services. Crowding phenomena that result in left behind passengers are studied for a subway system, and models to quantify their numbers and their effects on passenger waiting times are proposed. Paratransit is studied as a specific implementation of demand responsive transit service. This dissertation proposes a tool to optimize strategic coordination between paratransit services and taxi/TNCs in order to reduce the high operating costs of the service. Regarding flexible transit, a hybrid system is optimized here. This system has operational characteristics from both fixed route and flexible route deviation services and offers insights regarding the service area characteristics that are more suitable for flexible services.

This Chapter is organized as follows. Section 6.1 describes the conclusions from this dissertation, focusing of fixed route, on-demand and flexible services separately. Section 6.2 offers guidelines and insights on how the outputs of this dissertation can be transfered to other study sites. Finally, Section 6.3 describes the ways that the content of this dissertation could be extended in the future.

### 6.1 Conclusions

### 6.1.1 Crowding on fixed route systems

This part of the dissertation highlights the challenges that non-uniform spatiotemporal distribution of demand poses in fixed route transit, since it leads to overcrowded geographic locations (i.e., stations) during specific time periods within a day (i.e., peak hours). The proposed method aims to address this challenge by quantifying its impact on the performance of transit systems considering various data sources. More specifically, the objective of this analysis is to present a method for measuring passengers that are left behind overcrowded trains in transit stations without records of exiting passengers. Existing literature includes a study performed by Miller et al. (2018) that also addresses this challenging case. Based on the results obtained here, the method proposed in this dissertation performs better at low crowding levels, whereas the study by Miller et al. (2018) performs better at high crowding levels, proving the complementarity between the two works.

The proposed methodology uses archived data with automatic video counts as inputs to estimate the total number of left behind passengers during peak demand periods. Video counts were not proven accurate in isolation, but the development of logistic regression models that combine automated video counts with automated train-tracking records demonstrated good results for different applications. Logistic regression was selected since it allows probabilities associated with discrete outcomes (i.e., passengers boards or passengers is left behind) to be estimated based on measurable inputs. With the proposed methodology, the number of left behind passengers can be estimated within $17 \%$ of their actual number. The work performed in this area also revealed the effects of accounting for left behind passengers on the estimation of the current reliability metric used by transit agencies: the experienced waiting
times. The study site that was used to implement this method is the MBTA's subway Orange line. The results proved the need of accounting for left-behind passengers when tracking a transit system's reliability metrics. Finally, it is noteworthy that although there are limitations to any single data source, the potential for improving performance metrics through data fusion and modeling continues to grow.

### 6.1.2 Operating costs of on-demand systems

This study investigates the optimal allocation of trips between conventional ADA paratransit service and TNCs, in an attempt to reduce their high operating costs. In the case of on-demand services, the non-uniform spatio-temporal distribution of demand is associated with both challenges and opportunities that need to be taken into account when studying such systems. For example, a trip that happens to fall along the path of an otherwise empty vehicle can be served at very little cost to the agency. On the other extreme, a requested trip at the edge of the service area during early, late, or peak hours might require an additional vehicle to be put into service to drive out to serve the requested trip at great cost. Such detailed characteristics of a trip cannot be captured by an aggregated approach, whereas existing exact methods cannot be efficiently used for large datasets. In order to fill the respective gap in existing literature, a heuristic-based algorithm was developed to estimate the fleet size and vehicle operations required to serve a set of demanded paratransit trips each day.

The proposed algorithm is developed to estimate the marginal cost of each paratransit trip in the context of the vehicle routings so trips can be incrementally reassigned to alternative providers when the costs make it advantageous to do so. The routing model is based on some simplifying assumptions to make its development and implementation as fast and efficient as possible. The errors introduced by these
assumptions were partially eliminated through calibrations. Focusing on the specific study site of MBTA's paratransit service, the analysis of the removed trips shows that the first trips to be removed are the trips requested during the peak of the peak hours as well as the trips at the end of the day. The expected cost savings from this strategy for the transit agency can be as high as $53 \%$. As explained in the discussion section of Chapter 4, assigning all trips to TNCs may not be feasible for many reasons, such as the case of passengers not willing or not being able to switch to TNCs due to heavy equipment. In such a case, expected benefits for the operating agency from implementing the allocation strategy were still proven to be high. It is noteworthy that the total expected emissions of the paratransit service did not increase due to this strategy. For TNC cost coefficients equal to almost double the ones considered here, some trips still remain at the operation of the paratransit service. The implementation of the proposed method for the MBTA paratransit service did not prove any geographical inequities of user service.

### 6.1.3 Efficient implementation of flexible systems

In the area of flexible services, this dissertation investigates the level of transit service flexibility within a given area that would lead to the optimal costs for both users and agencies. Various types of flexible systems have been implemented but there is still room for improvement, both in terms of operation and design. This dissertation focuses on a new hybrid transit system, with elements of both fixed route and route deviation services. The level of flexibility is determined through continuous approximation and optimization techniques. The numerical analysis performed here adopts input values based on existing flexible service areas and reveals the behavior of the modeling approach under various case scenarios. Demand is a significant factor in this analysis, since it has a major effect on the results from implementing the
optimization formulas of flexible region size and station spacing, which are the main outputs of the proposed optimization approach.

Findings from this part of the dissertation confirm that the fixed route service becomes more beneficial as greater areas with greater demand are considered. As expected, the agency costs are always the lowest for the fixed route services, followed by the agency costs of the hybrid service and the full route deviation services. The hybrid service proposed here always has the lowest user costs compared to the other two services, leading to user benefits of up to $35 \%$ when compared with fixed route and $80 \%$ when compared with full flexible services. There is the exception of small areas with low demand density where the proposed hybrid service has a small loss for the users (less than 5\%) compared to full route deviation. These losses are attributed to the effect of agency costs in the optimization process. It is noteworthy that the percentage of passengers served curb-to-curb is found to not play an important role in determining the operational costs in this study. The investigation of headways revealed that service areas with low headways should expect greater user benefits from the proposed hybrid service. Finally, the benefits from the analyzed hybrid system as transit demand decreases is promising for the implementation of such systems during and after the COVID-19 pandemic when the density of transit demand has dropped in many communities to levels that no longer support fixed route transit service.

### 6.2 Transferability

The proposed methodology for fixed route systems is developed to estimate the number of left behind passengers at a transit station when trains are too crowded for them to board. The method relies primarily on train tracking records and surveillance video feeds automatically collected and recorded. Additional archived data that describe inferred travel patterns based on farecards are used to identify stations with
expected crowding conditions on the vehicles. If a transit agency knows in advance which are their most crowded stations, then this part of the analysis could be ignored. Manual counts considered in this analysis are used for model estimation and validation only. Consequently, the proposed methodology here requires only the automatically collected input data for further implementation at the station used here as a case study. In order for this methodology to be transferred to other transit systems, the first requirement is that the data sources described in Section 3.4 are available. It is reminded that manual counts need to be collected only once per station and after that the methodology depends only on the automatically collected input data.

In the case of on-demand services, the proposed methodology can be used as a decision making tool for transit agencies regarding which paratransit trips should be better served by an alternative provider rather than the agency vehicles. Implementation of this method only requires the daily trip requests in order for the models to determine which trips to assign to alternative providers, if any. In order for this methodology to be transferred to other case studies, the respective authorities should have detailed trip records for their paratransit operation for a significant time period (e.g., a semester). After proper processing, these data could provide the operational and network related input values that are required here for the development of the heuristic algorithm. After the algorithm that compares marginal trip costs with TNC (or taxi) costs is calibrated to account for the new case study's operational costs, the only required input for implementing this method is the set of daily trip requests. The required information about these requests is described in Section 4.4. It is highlighted that the heuristic algorithm developed here can be easily adjusted to account for any specific alternative provider, so the respective cost functions and/or cost coefficients should be properly updated.

Flexible services described in Chapter 5 can be optimized using the proposed
methodology of operating a hybrid service with elements of both fixed route and route deviation services. The proposed models to design the optimal flexible region and the optimal station spacing have been evaluated using a simulation method. This simulation method could also be used as an assessment tool for agencies that want to evaluate their performance under unexpected levels of demand after designing their service area using optimized decision variables considering the expected demand. Input values considered here are chosen from existing literature with the aim of using hypothetical case studies as close to reality as possible. To implement the proposed analytical models as well as the simulation approach, a transit agency would only need to have the input values presented in Section 5.6. This study can also serve as a reference for agencies to identify areas of additional data need.

### 6.3 Future extensions

There are a number of ways that the work presented here to address crowding phenomena of fixed route systems could be extended. One approach would be to implement and evaluate the developed models over more days. In terms of passenger flow data, the ODX model has some known drawbacks given existing limitations, such as lack of tap-out farecard data or passenger counters on trains. In systems without these limitations, the developed models could achieve higher accuracy. The methodology presented here could also be combined with the previous study by Miller et al. (2018) in order to improve the overall process for estimating left behind passengers in subway systems without tap-out. Comparing the two studies, Miller et al. (2018) achieves higher accuracy for very crowded conditions, whereas the method proposed in this dissertation performs better when there are few passengers left behind. The automated object detection presented in this study could also be combined with the model proposed by Miller et al. (2018) as part of its real-time implementation in
case of special events where real-time AFC is not available. In the area of image processing, alternative object detection tools could be tested in order to identify the one with the best performance.

In the area of on-demand services, this dissertation could be expanded in many different ways in the future. Individual passenger characteristics (e.g., loading/unloading times) could replace the average values included in the algorithm. Dynamic (or surge) pricing for TNCs could be incorporated in an attempt to achieve more accurate estimations of TNC trip costs. The allocation of paratransit trips to taxis or TNCs could also be investigated from the user side as well. Surveys could reveal the user perception towards such a strategy at a given study site and an improved algorithm that accounts for the user preferences could be developed. For example, if such a survey revealed that passengers that belong in a specific age group are not willing to use alternative providers, then these users could easily be eliminated from the proposed algorithm's implementation. Moreover, the social costs of having more vehicles driving around city centers as a potential result of such a strategy could be investigated. Such costs refer, for example, its effects on traffic congestion in urban centers. Also, ways to incorporate the environmental impacts in the algorithm as part of the trip allocation process could be studied, to add more sustainability elements in this research. In the case that the allocation strategy is satisfactory for users, the effects of potential induced demand need to be studied.

Finally, regarding flexible transit systems, the analysis performed in this dissertation could be extended to account for zone strategies associated with the operation of the hybrid service. Moreover, if real data are available, the model's performance could be tested and calibrated, if needed. Elements of alternative types of flexible services (e.g., point deviation) could be modeled and incorporated in the hybrid system's operation as a replacement of route deviation. Similarly to the extensions described
above for the on-demand services, such a strategy of vehicles deviating to serve passengers curb-to-curb can have environmental impacts which could be incorporated in the optimization process. Finally, since flexible transit systems respond to users needs in order to improve user experience from public transit, surveys on user preference toward proposed hybrid services could reveal the potential from implementing such services in real-life case studies. Such surveys could also assist in making the user cost coefficients better adjusted to users perceptions.

Public transit systems serve the needs of numerous users around the world and their efficient operation is associated with the successful completion of their daily activities. Acknowledging the importance of public transit for modern societies, this dissertation focuses on addressing challenges related with the performance of public transit systems, with studies ranging from fixed route to flexible and full on-demand systems. Outputs include methods and tools that can be implemented by transit agencies to improve the quality of their service and thus assure their successful operation. Focusing on both agencies and users, future work will enhance these outputs to provide additional innovative and sustainable solutions to current and future transit inefficiencies.

## A APPENDIX

A list of acronyms included in this dissertation is given below in alphabetical order:

ADA - Americans with Disabilities Act
AFC - Automatic Fare Collection
AIC - Akaike Information Criterion
APC - Automatic Passenger Count
AVL - Automated Vehicle Location
COCO - Common Objects in Context
DARP - Dial-a-Ride Problem
DART - Dial-A-Ride Transit
DRT - Demand Responsive Transportation
EMD - Earth Mover's Distance
EMFAC - Emission Factors
FIFO - First-In-First-Out
ITS - Intelligent Transportation Systems
MBTA - Massachusetts Bay Transportation Authority
MTA - Metropolitan Transportation Authority
NTST - National Transit Summaries and Trends
ODX - Origin-Destination-Transfer
PVTA - Pioneer Valley Transit Authority
R-CNN - Region Based Convolutional Neural Network
SCAT - Sarasota County Area Transit
SDP - Service Delivery Policy
SVM - Support Vector Machine

TCQSM - Transit Capacity and Quality of Service Manual TCRP - Transit Cooperative Highway Research Program TNC - Transportation Network Company

TRB - Transportation Research Board
TTR - Train Tracking Records
VHT - Vehicle Hours Traveled
VMT - Vehicle Miles Traveled
VRPTW - Vehicle Routing Problem with Time Windows
YOLO - You Only Look Once
WAV - Wheelchair Accessible Vehicle

## References

Aex, R.P., 1975. Demand responsive transit and the integration of $\mathrm{d} / \mathrm{r}$ systems with traditional transit. Transportation 4, 419-428.

Agard, B., Morency, C., Trépanier, M., 2006. Mining public transport user behaviour from smart card data. IFAC Proceedings Volumes 39, 399-404.

Amirgholy, M., Gonzales, E.J., 2016. Demand responsive transit systems with timedependent demand: user equilibrium, system optimum, and management strategy. Transportation Research Part B: Methodological 92, 234-252.

Ansari, S., Başdere, M., Li, X., Ouyang, Y., Smilowitz, K., 2018. Advancements in continuous approximation models for logistics and transportation systems: 19962016. Transportation Research Part B: Methodological 107, 229-252.

Attanasio, A., Cordeau, J.F., Ghiani, G., Laporte, G., 2004. Parallel tabu search heuristics for the dynamic multi-vehicle dial-a-ride problem. Parallel Computing 30, 377-387.

Batarce, M., Muñoz, J.C., de Dios Ortúzar, J., 2016. Valuing crowding in public transport: Implications for cost-benefit analysis. Transportation Research Part A: Policy and Practice 91, 358-378.

Batarce, M., Muñoz, J.C., de Dios Ortúzar, J., Raveau, S., Mojica, C., Ríos, R.A., 2015. Use of mixed stated and revealed preference data for crowding valuation on public transport in Santiago, Chile. Transportation Research Record 2535, 73-78.

Bongiovanni, C., Kaspi, M., Geroliminis, N., 2019. The electric autonomous dial-aride problem. Transportation Research Part B: Methodological 122, 436-456.

Broome, K., Worrall, L., Fleming, J., Boldy, D., 2012. Evaluation of flexible route bus transport for older people. Transport Policy 21, 85-91.

Buehler, R., Pucher, J., 2012. Demand for public transport in Germany and the USA: An analysis of rider characteristics. Transport Reviews 32, 541-567.

Bullard, R.D., Johnson, G.S., Torres, A.O., 2004. Highway Robbery: Transportation Racism \& New Routes to Equity. South End Press.

Burkhardt, J., Doherty, J., Rubino, J., Kim, J., 2008. Accessible community transportation in our nation: A survey on the use of taxis in paratransit programs. National Aging and Disability Transportation Center, Washington, DC .

Burkhardt, J.E., 2010. Potential cost savings from taxi paratransit programs, in: 12th international conference on mobility and transport for elderly and disabled persons, (TRANSED 2010), Hong Kong.

Camacho, T.D., Foth, M., Rakotonirainy, A., 2012. Pervasive technology and public transport: Opportunities beyond telematics. IEEE Pervasive Computing 12, 18-25.

Campbell, R., 2012. Simple player. Computer Software. Berkeley, California.
Carlson, R.C., 1976. Anatomy of a systems failure: Dial-a-ride in Santa Clara County, California. Transportation 5, 3-16.

Cats, O., West, J., Eliasson, J., 2016. A dynamic stochastic model for evaluating congestion and crowding effects in transit systems. Transportation Research Part B: Methodological 89, 43-57.

Chavis, C., Gayah, V.V., 2017. Development of a mode choice model for general purpose flexible-route transit systems. Transportation Research Record 2650, 133141.

Chen, J., Liu, Z., Wang, S., Chen, X., 2018. Continuum approximation modeling of transit network design considering local route service and short-turn strategy. Transportation Research Part E: Logistics and Transportation Review 119, 165188.

Cheng, D., Wang, J., Wei, X., Gong, Y., 2015. Training mixture of weighted SVM for object detection using EM algorithm. Neurocomputing 149, 473-482.

Chevrier, R., Canalda, P., Chatonnay, P., Josselin, D., 2006. Comparison of three algorithms for solving the convergent demand responsive transportation problem, in: 2006 IEEE Intelligent Transportation Systems Conference, IEEE. pp. 10961101.

Chia, D., 2008. Policies and practices for effectively and efficiently meeting ADA paratransit demand. TCRP Report 74. Transportation Research Board, Washington, DC.

Commins, N., Nolan, A., 2011. The determinants of mode of transport to work in the Greater Dublin Area. Transport Policy 18, 259-268.

Coniglio, C., Meurie, C., Lézoray, O., Berbineau, M., 2017. People silhouette extraction from people detection bounding boxes in images. Pattern Recognition Letters 93, 182-191.

Cordeau, J.F., Laporte, G., 2003. The Dial-a-Ride Problem (DARP): Variants, modeling issues and algorithms. Quarterly Journal of the Belgian, French and Italian Operations Research Societies 1, 89-101.

Cordeau, J.F., Laporte, G., 2007. The dial-a-ride problem: models and algorithms. Annals of Operations Research 153, 29-46.

Cox, T., Houdmont, J., Griffiths, A., 2006. Rail passenger crowding, stress, health and safety in britain. Transportation Research Part A: Policy and Practice 40, 244-258.

Daganzo, C.F., 1978. An approximate analytic model of many-to-many demand responsive transportation systems. Transportation Research 12, 325-333.

Daganzo, C.F., 1984. The distance traveled to visit n points with a maximum of c stops per vehicle: An analytic model and an application. Transportation Science 18, 331-350.

Davey, J.A., 2007. Older people and transport: Coping without a car. Ageing \& Society 27, 49-65.

Deflorio, F.P., 2011. Simulation of requests in demand responsive transport systems. IET Intelligent Transport Systems 5, 159-167.

Deng, J., Berg, A.C., Li, K., Fei-Fei, L., 2010. What does classifying more than 10,000 image categories tell us?, in: European Conference on Computer Vision, Springer. pp. 71-84.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 248-255.

Desrosiers, J., Dumas, Y., Soumis, F., 1986. A dynamic programming solution of the large-scale single-vehicle dial-a-ride problem with time windows. American Journal of Mathematical and Management Sciences 6, 301-325.

Dessouky, M., Ordóñez, F., Quadrifoglio, L., 2005. Productivity and cost-effectiveness of demand responsive transit systems. Citeseer.

Diana, M., Dessouky, M.M., Xia, N., 2006. A model for the fleet sizing of demand responsive transportation services with time windows. Transportation Research Part B: Methodological 40, 651-666.

Diec, J., Coxon, S., De Bono, A., 2010. Designing a public train station shelter to minimise anti-social behaviour and crime in Melbourne's metropolitan rail environment, in: 33rd Australasian Transport Research Forum (ATRF), Canberra.

Douglas, N., Karpouzis, G., 2005. Estimating the cost to passengers of station crowding, in: 28th Australasian Transport Research Forum (ATRF), Sydney.

Drabicki, A., Kucharski, R., Cats, O., Fonzone, A., 2017. Simulating the effects of real-time crowding information in public transport networks, in: 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), IEEE. pp. 675-680.

Eboli, L., Mazzulla, G., 2012. Performance indicators for an objective measure of public transport service quality. European Transport 51, 1-21.

Ellis, E.H., 2016. Use of Taxis in Public Transportation for People with Disabilities and Older Adults. Project J-7, Topic SB-26.

Errico, F., Crainic, T.G., Malucelli, F., Nonato, M., 2013. A survey on planning semiflexible transit systems: Methodological issues and a unifying framework. Transportation Research Part C: Emerging Technologies 36, 324-338.

Evans, G.W., Wener, R.E., 2007. Crowding and personal space invasion on the train: Please don't make me sit in the middle. Journal of Environmental Psychology 27, 90-94.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The Pascal visual object classes (VOC) challenge. International Journal of Computer Vision 88, 303-338.

Fan, Y., Guthrie, A., Levinson, D., 2016. Waiting time perceptions at transit stops and stations: Effects of basic amenities, gender, and security. Transportation Research Part A: Policy and Practice 88, 251-264.

Farber, S., Ritter, B., Fu, L., 2016. Space-time mismatch between transit service and observed travel patterns in the Wasatch Front, Utah: A social equity perspective. Travel Behaviour and Society 4, 40-48.

Faroqi, H., Mesbah, M., Kim, J., 2019. Comparing sequential with combined spatiotemporal clustering of passenger trips in the public transit network using smart card data. Mathematical Problems in Engineering 2019.

Federal Highway Administration, 2002. Status of the nation's highways, bridges, and transit: 2002 conditions and performance report, Chapter 14.

Fielding, G.J., Glauthier, R.E., Lave, C.A., 1977. Development of performance indicators for transit. Institute of Transportation Studies, University of California. Technical Report.

Figliozzi, M.A., 2008. Planning approximations to the average length of vehicle routing problems with varying customer demands and routing constraints. Transportation Research Record 2089, 1-8.

Figliozzi, M.A., 2009. Planning approximations to the average length of vehicle routing problems with time window constraints. Transportation Research Part B: Methodological 43, 438-447.

Flusberg, M., 1976. An innovative public transportation system for a small city: The Merrill, Wisconsin, case study. Transportation Research Record 606, 54-59.

Fu, L., 1999. Improving paratransit scheduling by accounting for dynamic and stochastic variations in travel time. Transportation Research Record 1666, 74-81.

Fu, L., 2002a. Planning and design of flex-route transit services. Transportation Research Record 1791, 59-66.

Fu, L., 2002b. Scheduling dial-a-ride paratransit under time-varying, stochastic congestion. Transportation Research Part B: Methodological 36, 485-506.

Fu, L., 2002c. A simulation model for evaluating advanced dial-a-ride paratransit systems. Transportation Research Part A: Policy and Practice 36, 291-307.

Furth, P.G., Hemily, B., Muller, T.H., Strathman, J.G., 2006. Using archived AVLAPC data to improve transit performance and management. TRB National Research Council, Washington, D.C.

Gattoufi, S., Oral, M., Kumar, A., Reisman, A., 2004. Content analysis of data envelopment analysis literature and its comparison with that of other OR/MS fields. Journal of the Operational Research Society 55, 911-935.

Gilbert, G., Dajani, J.S., 1975. Measuring the performance of transit service. Center for Policy Analysis, Institute of Policy Sciences and Public Affairs, Duke University.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587.

Gong, Y., Liu, Y., Lin, Y., Yang, J., Duan, Z., Li, G., 2012. Exploring spatiotemporal characteristics of intra-urban trips using metro smartcard records, in: 2012 20th International Conference on Geoinformatics, IEEE. pp. 1-7.

Gonzales, E.J., Keklikoglou, A., Sipetas, C., 2018. Measuring Left-Behinds on Subway. Technical Report.

Guenther, K.W., Authority, A.A.T., 1976. Demand-responsive transportation in Ann Arbor: Operation. Transportation Research Record 608, 20-25.

Hasan, S., Schneider, C.M., Ukkusuri, S.V., González, M.C., 2013. Spatiotemporal patterns of urban human mobility. Journal of Statistical Physics 151, 304-318.

Haywood, L., Koning, M., 2015. The distribution of crowding costs in public transport: New evidence from Paris. Transportation Research Part A: Policy and Practice 77, 182-201.

Haywood, L., Koning, M., Monchambert, G., 2017. Crowding in public transport: Who cares and why? Transportation Research Part A: Policy and Practice 100, 215-227.

Heathington, K.W., Davis Jr, F.W., Middendorf, D.P., Brogan, J.D., 1974. Demandresponsive transportation systems in the private sector. Transpn Res. Record 522, 46-55.

Hensher, D.A., Rose, J.M., Leong, W., Tirachini, A., Li, Z., 2013. Choosing public transport-incorporating richer behavioural elements in modal choice models. Transport Reviews 33, 92-106.

Hörcher, D., Graham, D.J., Anderson, R.J., 2017. Crowding cost estimation with large scale smart card and vehicle location data. Transportation Research Part B: Methodological 95, 105-125.

Hu, N., Legara, E.F., Lee, K.K., Hung, G.G., Monterola, C., 2016. Impacts of land use and amenities on public transport use, urban planning and design. Land Use Policy 57, 356-367.

Jaw, J.J., Odoni, A.R., Psaraftis, H.N., Wilson, N.H., 1986. A heuristic algorithm for the multi-vehicle advance request dial-a-ride problem with time windows. Transportation Research Part B: Methodological 20, 243-257.

Karlaftis, M.G., 2004. A dea approach for evaluating the efficiency and effectiveness of urban transit systems. European Journal of Operational Research 152, 354-364.

Karlaftis, M.G., Tsamboulas, D., 2012. Efficiency measurement in public transport: are findings specification sensitive? Transportation Research Part A: Policy and Practice 46, 392-402.

Kim, K.M., Hong, S.P., Ko, S.J., Kim, D., 2015. Does crowding affect the path choice of metro passengers? Transportation Research Part A: Policy and Practice 77, 292-304.

Kim, M.E., Levy, J., Schonfeld, P., 2019. Optimal zone sizes and headways for flexible-route bus services. Transportation Research Part B: Methodological 130, 67-81.

Kittelson \& Associates, Inc., 1999. Transit capacity and quality of service manual (1st ed.). TCRP Project A-15. TRB, National Research Council, Washington, D.C.

Kittelson and Associates Inc, Brinckerhoff, P., Inc, K.G., Institute, T.A.T., Arup, 2013. Transit service and quality of service manual. Transportation Research Board of the National Academies.

Koffman, D., 2004. Operational experiences with flexible transit services. 53, Transportation Research Board, Washington, DC.

Kraus, M., 1991. Discomfort externalities and marginal cost transit fares. Journal of Urban Economics 29, 249-259.

Lee, S.G., Hickman, M., Tong, D., 2013. Development of a temporal and spatial linkage between transit demand and land-use patterns. Journal of Transport and Land Use 6, 33-46.

Lem, L.L., Li, J.L., Wachs, M., 1994. Comprehensive transit performance indicators. UCTC No. 225. University of California, Berkeley.

Lerman, S., Wilson, N.H., 1974. Analytic model for predicting dial-a-ride system performance. Transportation Research Board Special Report 147, 48-53.

Lewis, D., Evans, T., Koffman, D., 1998. Impact of reliability on paratransit demand and operating costs. Transportation Planning and Technology 21, 323-346.

Lézoray, O., Grady, L., 2012. Image processing and analysis with graphs: theory and practice. CRC Press.

Li, B., Zhang, J., Zhang, Z., Xu, Y., 2014. A people counting method based on head detection and tracking, in: 2014 International Conference on Smart Computing, IEEE. pp. 136-141.

Li, Z., Hensher, D.A., 2011. Crowding and public transport: A review of willingness to pay evidence and its relevance in project appraisal. Transport Policy 18, 880-887.

Li, Z., Hensher, D.A., 2013. Crowding in public transport: a review of objective and subjective measures. Journal of Public Transportation 16, 6.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common Objects in Context, in: European Conference on Computer Vision, Springer. pp. 740-755.

Lundberg, U., 1976. Urban commuting: Crowdedness and catecholamine excretion. Journal of Human Stress 2, 26-32.

Luo, D., Cats, O., van Lint, H., 2019. Can passenger flow distribution be estimated solely based on network properties in public transport systems? Transportation 47, 2757-2776.

Lyman, C., Campbell, N., Gonzales, E.J., Christofa, E., 2019. Modeling the effect of new commuter bus service on demand and impact on greenhouse gas emissions: Application to Greater Boston. Transportation Research Record 2673, 125-138.

Ma, X., Liu, C., Wen, H., Wang, Y., Wu, Y.J., 2017. Understanding commuting patterns using transit smart card data. Journal of Transport Geography 58, 135145.

Ma, X., Wu, Y.J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transportation Research Part C: Emerging Technologies 36, 1-12.

Ma, Z., Koutsopoulos, H.N., Chen, Y., Wilson, N.H., 2019. Estimation of denied boarding in urban rail systems: alternative formulations and comparative analysis. Transportation Research Record 2673, 771-778.

Mamun, S.A., Lownes, N.E., Osleeb, J.P., Bertolaccini, K., 2013. A method to define public transit opportunity space. Journal of Transport Geography 28, 144-154.

Manley, E., Zhong, C., Batty, M., 2018. Spatiotemporal variation in travel regularity through transit user profiling. Transportation 45, 703-732.

MBTA, 2010. Service Delivery Policy. Massachusetts Bay Transportation Authority, Boston, MA.

MBTA, 2017. Service Delivery Policy. Massachusetts Bay Transportation Authority, Boston, MA.

Miller, E., Sánchez-Martínez, G.E., Nassir, N., 2018. Estimation of passengers left behind by trains in high-frequency transit service operating near capacity. Transportation Research Record 2672, 497-504.

Miller, P., de Barros, A.G., Kattan, L., Wirasinghe, S., 2016. Public transportation and sustainability: A review. KSCE Journal of Civil Engineering 20, 1076-1083.

Mukherjee, S., Saha, B., Jamal, I., Leclerc, R., Ray, N., 2011. A novel framework for automatic passenger counting, in: 2011 18th IEEE International Conference on Image Processing, IEEE. pp. 2969-2972.

Mulley, C., Nelson, J.D., 2009. Flexible transport services: A new market opportunity for public transport. Research in Transportation Economics 25, 39-45.

Nakanishi, Y., Kim, K., Ulusoy, Y., Bata, A., 2003. Assessing emergency preparedness of transit agencies: Focus on performance indicators. Transportation Research Record 1822, 24-32.

Neutens, T., Delafontaine, M., Scott, D.M., De Maeyer, P., 2012. A GIS-based method to identify spatiotemporal gaps in public service delivery. Applied Geography $32,253-264$.

Newell, G.F., 1973. Scheduling, location, transportation, and continuum mechanics: some simple approximations to optimization problems. SIAM Journal on Applied Mathematics 25, 346-360.

Nourbakhsh, S.M., Ouyang, Y., 2012. A structured flexible transit system for low demand areas. Transportation Research Part B: Methodological 46, 204-216.

Noursalehi, P., Koutsopoulos, H.N., Zhao, J., 2019. Predictive decision support for real-time crowding prediction and information generation, in: 98th Annual Meeting of Transportation Research Board, Washington, D.C.

Nuzzolo, A., Crisalli, U., Comi, A., Rosati, L., 2016. A mesoscopic transit assignment model including real-time predictive information on crowding. Journal of Intelligent Transportation Systems 20, 316-333.

Osuna, E., Newell, G., 1972. Control strategies for an idealized public transportation system. Transportation Science 6, 52-72.

Ozer, B., Wolf, M., 2014. A train station surveillance system: Challenges and solutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 638-643.

Pei, M., Lin, P., Ou, J., 2019. Real-time optimal scheduling model for transit system with flexible bus line length. Transportation Research Record 2673, 800-810.

Pelletier, M.P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19, 557-568.

Pew Research Center, 2015. Who relies on public transit in the U.S. https://www.pewresearch.org/fact-tank/2016/04/07/who-relies-on-public-transit-in-the-u-s/.

Phillips, J.K., 2004. An application of the balanced scorecard to public transit system performance assessment. Transportation Journal 43, 26-55.

Potts, J.F., Marshall, M.A., Crockett, E.C., Washington, J., 2010. A guide for planning and operating flexible public transportation services. Transit Cooperative Research Program (TCRP) Report 140. Transportation Research Board, Washington, DC.

Qiu, F., Li, W., Haghani, A., 2015. A methodology for choosing between fixed-route and flex-route policies for transit services. Journal of Advanced Transportation 49, 496-509.

Quadrifoglio, L., Dessouky, M.M., Ordóñez, F., 2008. A simulation study of demand responsive transit system design. Transportation Research Part A: Policy and Practice 42, 718-737.

Quadrifoglio, L., Li, X., 2009. A methodology to derive the critical demand density for designing and operating feeder transit services. Transportation Research Part B: Methodological 43, 922-935.

Rahimi, M., Amirgholy, M., Gonzales, E.J., 2018. System modeling of demand responsive transportation services: Evaluating cost efficiency of service and coordinated taxi usage. Transportation Research Part E: Logistics and Transportation Review 112, 66-83.

Rahimi, M., Gonzales, E.J., 2015. Systematic evaluation of zoning strategies for demand responsive transit, in: Transportation Research Board 94th Annual Meeting.

Raveau, S., Guo, Z., Muñoz, J.C., Wilson, N.H., 2014. A behavioural comparison of route choice on metro networks: Time, transfers, crowding, topology and sociodemographics. Transportation Research Part A: Policy and Practice 66, 185-195.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788.

Van der Reis, A.P., 1983. Black commuting in Pretoria: Attitudes towards crowding. Technical Report. National Institute for Transport Road Research, Special Report BCP 9, South Africa.

Rietveld, P., 2005. Six reasons why supply-oriented indicators systematically overestimate service quality in public transport. Transport Reviews 25, 319-328.

Ritzinger, U., Puchinger, J., Hartl, R.F., 2016. Dynamic programming based metaheuristics for the dial-a-ride problem. Annals of Operations Research 236, 341-358.

Ronald, N., Thompson, R., Winter, S., 2015. Simulating demand-responsive transportation: a review of agent-based approaches. Transport Reviews 35, 404-421.

Ruan, M., Lin, J., 2009. An investigation of bus headway regularity and service performance in chicago bus transit system, in: Transport Chicago, annual conference.

Rubner, Y., Tomasi, C., Guibas, L.J., 2000. The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40, 99-121.

San Francisco County Transportation Authority, T., 2017. Today: A profile of San Francisco transportation network company activity.

Sánchez-Martínez, G.E., 2017. Inference of public transportation trip destinations by using fare transaction and vehicle location data: Dynamic programming approach. Transportation Research Record 2652, 1-7.

Scott, D., Horner, M., 2008. Examining the role of urban form in shaping people's accessibility to opportunities: an exploratory spatial data analysis. Journal of Transport and Land Use 1, 89-119.

Scott, R.A., 2010. Demand responsive passenger transport in low-demand situations. Research Report,425, Wellington: NZ Transport Agency.

Shen, C.W., Quadrifoglio, L., 2012. Evaluation of zoning design with transfers for paratransit services. Transportation Research Record 2277, 82-89.

Shi, Z., Zhang, N., Liu, Y., Xu, W., 2018. Exploring spatiotemporal variation in hourly metro ridership at station level: The influence of built environment and topological structure. Sustainability 10, 4564.

Sipetas, C., Keklikoglou, A., Gonzales, E.J., 2020. Estimation of left behind subway passengers through archived data and video image processing. Transportation Research Part C: Emerging Technologies 118, 102727.

Sonka, M., Hlavac, V., Boyle, R., 2014. Image processing, analysis, and machine vision. Cengage Learning.

Stanley, R.G., 2004. Performance-based measures in transit fund allocation. 56, Transportation Research Board.

Stasko, T., Levine, B., Reddy, A., 2016. Time-expanded network model of train-level subway ridership flows using actual train movement data. Transportation Research Record 2540, 92-101.

Stein, D.M., 1978. Scheduling dial-a-ride transportation systems. Transportation Science 12, 232-249.

Sundstrom, E., 1978. Crowding as a sequential process: Review of research on the effects of population density on humans. Human Response to Crowding. Hillsdale, NJ.

Talley, W.K., 1986. A comparison of two methodologies for selecting transit performance indicators. Transportation 13, 201-210.

Tang, C., Ceder, A., Ge, Y.E., 2018. Optimal public-transport operational strategies to reduce cost and vehicle's emission. PloS one 13, e0201138.

Tellez, O., Vercraene, S., Lehuédé, F., Péton, O., Monteiro, T., 2018. The fleet size and mix dial-a-ride problem with reconfigurable vehicle capacity. Transportation Research Part C: Emerging Technologies 91, 99-123.

Tirachini, A., 2019. Ride-hailing, travel behaviour and sustainable mobility: an international review. Transportation 47, 2011-2047.

Tirachini, A., Hensher, D.A., Rose, J.M., 2013. Crowding in public transport systems: Effects on users, operation and implications for the estimation of demand. Transportation Research Part A: Policy and Practice 53, 36-52.

Tirachini, A., Sun, L., Erath, A., Chakirov, A., 2016. Valuation of sitting and standing in metro trains using revealed preferences. Transport Policy 47, 94-104.

Torkjazi, M., Huynh, N., 2019. Effectiveness of dynamic insertion scheduling strategy for demand-responsive paratransit vehicles using agent-based simulation. Sustainability 11, 5391.

Toth, P., Vigo, D., 1996. Fast local search algorithms for the handicapped persons transportation problem, in: Meta-Heuristics. Springer, pp. 677-690.

Toth, P., Vigo, D., 1997. Heuristic algorithms for the handicapped persons transportation problem. Transportation Science 31, 60-71.

Transportation Research Board, 2003. TCRP Report 88: A Guidebook for Developing a Transit Performance-Measurement System. TRB of the National Academies, Washington, D.C.

Trépanier, M., Morency, C., Agard, B., 2009. Calculation of transit performance measures using smartcard data. Journal of Public Transportation 12, 5.

Turmo, V., Rahimi, M., Gonzales, E.J., Armstrong, P., 2018. Evaluating potential demand and operational effects of coordinated americans with disabilities act paratransit and taxi service. Transportation Research Record 2672, 686-697.

Turner, S., Corbett, E., O’Hara, R., White, J., 2004. Health and safety effects of rail crowding: Hazard identification. Rail Safety and Standards Board. HSL Report RAS/04/12.

Tuttle, D., Eaton, K., 2012. Optimizing accessible taxi service to augment traditional public transit services in Delaware. DCT 222. Institute of Public Administration, University of Delaware.

Tyrinopoulos, Y., Antoniou, C., 2008. Public transit user satisfaction: Variability and policy implications. Transport Policy 15, 260-272.

Velastin, S.A., Boghossian, B.A., Vicencio-Silva, M.A., 2006. A motion-based image processing system for detecting potentially dangerous situations in underground railway stations. Transportation Research Part C: Emerging Technologies 14, 96113.

Viola, P., 1993. Feature-based recognition of objects, in: Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision, pp. 60-64.

Vitt, J., Bauer, H., Canty, E., Golob, T., Heathington, K., 1970. Determining the importance of user-related attributes for a demand-responsive transportation system. Highway Research Record 318, 50-65.

Vuchic, V.R., 2007. Urban transit systems and technology. John Wiley \& Sons.
Wardman, M., 2004. Public transport values of time. Transport policy 11, 363-377.
Weiner, R., 2008. Integration of paratransit and fixed-route transit services. volume 76. Transportation Research Board, Washington, DC.

Wenzel, T., Rames, C., Kontou, E., Henao, A., 2019. Travel and energy implications of ridesourcing service in Austin, Texas. Transportation Research Part D: Transport and Environment 70, 18-34.

Wilson, N.H.M., Weissberg, R.W., Hauser, J., 1976. Advanced dial-a-ride algorithms research project (No. R76-20 Final RPt). MIT, Boston.

Wirasinghe, S.C., Hurdle, V.F., Newell, G.F., 1977. Optimal parameters for a coordinated rail and bus transit system. Transportation Science 11, 359-374.

Wong, K., Bell, M.G., 2006. Solution of the dial-a-ride problem with multidimensional capacity constraints. International Transactions in Operational Research 13, 195-208.

Yan-yan, C., Ning, C., Yu-yang, Z., Ke-han, W., Wei-wei, Z., 2014. Pedestrian detection and tracking for counting applications in metro station. Discrete Dynamics in Nature and Society 2014.

Yao, D., Xu, L., Li, J., 2019. Evaluating the performance of public transit systems: A case study of eleven cities in China. Sustainability 11, 3555.

Yu, C., He, Z.C., 2017. Analysing the spatial-temporal characteristics of bus travel demand using the heat map. Journal of Transport Geography 58, 247-255.

Yu, L., Chen, Q., Chen, K., 2019. Deviation of peak hours for urban rail transit stations: A case study in xi'an, china. Sustainability 11, 2733.

Zhang, S.J., Chen, S.Y., 2014. Design \& implementation of real-time broadcast system of urban public transport crowding index based on the internet of things, in: Advanced Materials Research, Trans Tech Publ. pp. 1753-1758.

Zhang, X., Xu, Y., Tu, W., Ratti, C., 2018. Do different datasets tell the same story about urban mobility - a comparative study of public transit and taxi usage. Journal of Transport Geography 70, 78-90.

Zheng, Y., Li, W., Qiu, F., 2018a. A methodology for choosing between route deviation and point deviation policies for flexible transit services. Journal of Advanced Transportation 2018.

Zheng, Y., Li, W., Qiu, F., 2018b. A slack arrival strategy to promote flex-route transit services. Transportation Research Part C: Emerging Technologies 92, 442455.

Zhou, Z.H., 2012. Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.

Zhu, Y., Koutsopoulos, H.N., Wilson, N.H., 2017. Inferring left behind passengers in congested metro systems from automated data. Transportation Research Procedia 23, 362-379.

Zhu, Y., Koutsopoulos, H.N., Wilson, N.H., 2018. Inferring left behind passengers in congested metro systems from automated data. Transportation Research Part C: Emerging Technologies 94, 323-337.


[^0]:    This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

[^1]:    Anna Nagurney, Member

[^2]:    ${ }^{1}$ All stations from Tufts Medical Center through Haymarket and the northbound platform at North Station on the Orange Line ( 11 platforms), three out of four Blue Line stations in downtown Boston (5 platforms), and all northbound platforms for the Red Line from South Station to Porter (8 platforms) meet this criterion.

