Designing Next-generation Transportation Systems with Emerging Vehicle Technologies

Zhiwei Chen
University of South Florida

Follow this and additional works at: https://digitalcommons.usf.edu/etd

Scholar Commons Citation
https://digitalcommons.usf.edu/etd/8525

This Dissertation is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact digitalcommons@usf.edu.
Designing Next-generation Transportation Systems with Emerging Vehicle Technologies

by

Zhiwei Chen

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Civil Engineering Department of Civil and Environmental Engineering College of Engineering University of South Florida

Major Professor: Xiaopeng Li, Ph.D.
Amy L. Stuart, Ph.D.
Deb Niemeier, Ph.D.
Fred Mannering, Ph.D.
Hadi Charkhgard, Ph.D.

Date of Approval:
November 29, 2020

Keywords: Modular Autonomous Vehicles, System Design, Inequality, Discrete Optimization, Continuous Approximation, Agent-based Simulation

Copyright © 2020, Zhiwei Chen
Dedication

I dedicate this dissertation to my beloved parents and sisters.
Acknowledgments

I sincerely thank my advisor Dr. Xiaopeng Li for his guidance and assistance during my studies at USF. I am also very thankful for my dissertation committee members Dr. Amy L. Stuart, Dr. Deb Niemeier, Dr. Fred Mannering, and Dr. Hadi Charkgard for their valuable advices during the preparation of this dissertation. I learned the importance of finding scientific questions and of viewing transportation problems from an interdisciplinary perspective from them. Thank everyone in the Connected & Autonomous Transportation Systems (CATS) Lab. I cherish the time we spent together in the past years.
# Table of Contents

List of Tables .................................................................................................................. iv

List of Figures .................................................................................................................. v

Abstract .......................................................................................................................... vii

Chapter 1: Introduction ................................................................................................. 1  
  1.1 Demand Supply Asymmetry in Transportation Systems ............................................ 1  
  1.2 Embracing a Modular Autonomous Vehicle Era ....................................................... 2  
  1.3 How Will MAVs Change Transportation Systems? .................................................. 3  
  1.4 Research Goal and Scientific Questions ................................................................ 4  
  1.5 Dissertation Overview and Contributions ............................................................... 6  

Chapter 2: Literature Review ......................................................................................... 8  
  2.1 Operational Design Paradigm .................................................................................. 8  
  2.2 Continuous Approximation Method ....................................................................... 10  
  2.3 Discrete Optimization Method .............................................................................. 13  
  2.4 Equity Studies on Emerging Vehicle Technologies ................................................... 15  

Chapter 3: Shuttle Systems ......................................................................................... 18  
  3.1 Overview ................................................................................................................ 18  
  3.2 Problem Statement ............................................................................................... 21  
  3.3 Theoretical Property Analysis ............................................................................... 26  
  3.4 Continuous Model ............................................................................................... 30  
     3.4.1 Model Formulation ....................................................................................... 30  
     3.4.2 Analytical Solution ...................................................................................... 33  
     3.4.3 Discretization Method ............................................................................... 35  
  3.5 Discrete Model ..................................................................................................... 37  
     3.5.1 Model Formulation ...................................................................................... 38  
     3.5.2 Valid Inequalities ....................................................................................... 40  
     3.5.3 Dynamic Programming Algorithm .............................................................. 43  
  3.6 Numerical Experiments ......................................................................................... 47  
     3.6.1 Case Study 1: Batong Line in Beijing Subway System .................................... 47  
     3.6.1.1 Computation Performance .................................................................... 49  
     3.6.1.2 Effectiveness of MAV Shuttle Systems ................................................ 53  
     3.6.1.3 Sensitivity Analysis ............................................................................... 56  
     3.6.2 Case Study 2: Future MAV Services in Tampa ............................................. 59  
  3.7 Chapter Summary ................................................................................................. 62
Chapter 4: Corridor Systems

4.1 Overview ..................................................................................64
4.2 Problem Statement ..................................................................67
4.3 Theoretical Property Analysis ..................................................74
4.4 Continuous Model ...................................................................77
  4.4.1 Model Formulation ...............................................................78
  4.4.2 Analytical Solution ...............................................................80
  4.4.3 Discretization Method ............................................................83
4.5 Discrete Model .........................................................................85
  4.5.1 Model Formulation ...............................................................85
  4.5.2 Solution Space Reduction Properties ..................................88
  4.5.3 Branch and Bound Algorithm .............................................92
    4.5.3.1 Algorithm Overview .......................................................92
    4.5.3.2 Lower Bound ...............................................................98
    4.5.3.3 Upper Bound ..............................................................100
4.6 Numerical Experiments ............................................................103
  4.6.1 Case Study 1: A Hypothetical Transit Corridor in Mandl’s Network ......103
    4.6.1.1 Computation Performance ...........................................105
    4.6.1.2 Effectiveness of the MAV Corridor Systems ....................109
  4.6.2 Case Study 2: MAV Service in Batong Line, Beijing Subway System ....114
4.7 Chapter Summary .....................................................................118

Chapter 5: Inequality Impacts of AV Systems ..................................119
5.1 Overview ..................................................................................119
5.2 Methods ...................................................................................122
  5.2.1 Methodological Framework .................................................122
    5.2.1.1 AV System Simulation ......................................................122
    5.2.1.2 Transportation Outcome Metrics ...................................125
    5.2.1.3 Inequality Analysis Methods ..........................................127
  5.2.2 Case Study Application ........................................................129
    5.2.2.1 Study Area .......................................................................129
    5.2.2.2 Data Sources ...................................................................131
    5.2.2.3 Study Scenarios ..............................................................132
5.3 Results and Discussions ............................................................133
  5.3.1 Summary Statistics ...............................................................133
  5.3.2 Inequality Analysis from the Horizontal Perspective ................135
  5.3.3 Inequality Analysis from the Vertical Perspective ...................141
  5.3.4 Limitations ............................................................................151
5.4 Chapter Summary .....................................................................153

Chapter 6: The Road Ahead ............................................................155
6.1 Summary of Results .................................................................155
  6.1.1 Computational and Analytical Tools .....................................155
  6.1.2 Properties of Optimal MAV System Design .........................156
  6.1.3 Performance of MAV Systems ............................................157
  6.1.4 Inequality Impacts of AV Systems .......................................159
6.2 Future Research Directions

6.2.1 Transmodal Systems
6.2.2 Stochastic Demand
6.2.3 Simulating MAVs in a Metropolitan Area
6.2.4 Equity-aware System Design

References

Appendix A: Copyright Permissions
Appendix B: Proofs of Lemmas and Propositions
List of Tables

Table 3.1: Statistics of solution quality metrics.................................................................51
Table 3.2: Statistics of computational time metrics............................................................53
Table 3.3: System performance metrics of different demand scenarios...............................54
Table 4.1: Solution quality of different solution methods ....................................................108
Table 4.2: Computational time of different solution methods..............................................109
Table 5.1: Summary statistics of the study data and transportation outcome metrics........135
Table 5.2: Gini index values for different transportation outcomes .....................................136
Table 5.3: Summary statistics of transportation outcomes by population subgroups........145
List of Figures

Figure 1.1: MAV prototypes in industry ................................................................. 3
Figure 1.2. Summary of dissertation contents ......................................................... 6
Figure 3.1: A motivating example for MAV shuttle systems ...................................... 19
Figure 3.2: Cumulative passenger counts in MAV shuttle systems ............................. 22
Figure 3.3: Oversaturated and unsaturated periods ................................................... 28
Figure 3.4: Revised preferred virtual arrival demand curve in MAV shuttle systems ........ 31
Figure 3.5: Discretization of $h^*(t)$ ......................................................................... 36
Figure 3.6: Cumulative passenger counts in MAV shuttle systems in a discrete setting .... 38
Figure 3.7: Batong Line and the converted shuttle system .............................................. 49
Figure 3.8: Passenger demand in the converted shuttle system in Batong Line ............... 49
Figure 3.9: Optimal MAV system design for Batong Line ........................................... 55
Figure 3.10: Sensitivity analysis of performance metrics to different input parameters of Batong Line ......................................................................................... 58
Figure 3.11: Route information (left, source: Google Maps) and arrival demand rate (right) in the hypothetical shuttle transportation system in Tampa ......................................... 59
Figure 3.12: Optimal design for future MAV services in Tampa ...................................... 61
Figure 3.13: Performance metrics under various demand scenarios ............................... 61
Figure 4.1: An illustrative example for different operational modes ............................. 65
Figure 4.2: The corridor system (top) and cumulative passenger counts (bottom) ........... 68
Figure 4.3: An example for cumulative arrival and departure curves decomposition ........... 71
Figure 4.4: Sub-region division for the feasible region of $h(t)$ for unit-time problems .......... 81
Figure 4.5: Framework of the expedited B&B algorithm .......................................................... 93

Figure 4.6: Mandl’s network and the hypothetical corridor ..................................................... 104

Figure 4.7: Sensitivity analysis of Case study 1 ....................................................................... 112

Figure 4.8: Passenger demand rate during the unsaturated period of Batong line in Beijing Subway system ................................................................................................................. 115

Figure 4.9: Optimal design for Batong line in Beijing Subway System with a minimum dispatch headway of 9 mins (top row), 6 mins (middle row), and 3 mins (bottom row) .......................................................... 115

Figure 4.10: Sensitivity analysis of Case study 2 ................................................................. 117

Figure 5.1: Methodological framework for analyzing the inequality impacts of AVs .......... 123

Figure 5.2: Geographic extent of Tampa Bay study region (source: Google Maps) ............. 130

Figure 5.3: Spatial distribution of accessibility ..................................................................... 137

Figure 5.4: Spatial distribution of person miles traveled (miles/day) .................................. 138

Figure 5.5: Spatial distribution of travel costs ($/day) ......................................................... 140

Figure 5.6: Spatial distribution of the travel cost to income ratio (%) .................................. 140

Figure 5.7: Subgroup inequality index versus outcome levels for job accessibility and school accessibility ............................................................................................................ 148

Figure 5.8: Subgroup inequality index versus outcome levels for social activity accessibility and person miles traveled .......................................................................................... 149

Figure 5.9: Subgroup inequality index versus outcome levels for travel cost and travel cost to income ratio ............................................................................................................ 150

Figure B.1: An illustrative example to prove Propositions 3.1 and 3.2 ............................... 180

Figure B.2: An illustrative example to prove Proposition 3.4 ............................................... 183
Abstract

Recent advances in computing and artificial intelligence have enabled the development of various emerging vehicle technologies, e.g., autonomous vehicles (AV) and modular autonomous vehicles (MAV). These technologies bring new scientific and engineering problems challenging transportation researchers and practitioners. This dissertation aims to develop a suite of scalable computational and analytical tools for designing and analyzing next-generation transportation systems with the MAV technologies. Also, we intend to empirically study the system impacts in terms of the quality of service, energy implications, and inequality impacts.

For MAV system design, we develop a methodological framework centering at theoretical properties of the optimal system design under two system settings: i) shuttle systems consisting of one origin and one destination, ii) corridor systems where vehicles travel along a set of stations. We mathematically prove a series of elegant properties of system design variables when the system optimum is reached. These properties are then used to decompose the spatiotemporal correlation between the system design variables, which allows us to formulate the system design into separable continuum approximation models. Also, these properties are used to derive valid inequalities that can dramatically reduce the solution space of the system design problem. As a result, they can be applied to solve discrete formulations of the problem to expedite the search for the exact optimal design. Extensive numerical studies are conducted to evaluate the computation performance of the proposed solution methods. Results indicate that discrete models expedited by the theoretical properties find optimal system design more quickly. The continuous models, instead, offer highly
accurate near-optimal design in less than one second. These two methods complement each other in terms of the solution accuracy and computation time.

With the methodological framework, we also conduct extensive experiments to assess the impacts of the MAV technologies on system performance. We compare the energy cost for vehicle operations, the passenger waiting cost, and the total system cost between the proposed MAV transportation system and a benchmark system where fixed capacity vehicles are operated. Results reveal that by dynamically adjusting the vehicle capacity to accommodate the passenger demand, MAVs consistently decrease the energy cost of shuttle systems and corridor systems in a range of parameter settings. Meanwhile, the passenger waiting cost are reduced or remain at least the same as the fixed capacity operation in most cases. In a few cases the passenger waiting cost is slightly increased. However, the slight increase in the passenger waiting cost is compensated by the large reduction in the energy cost, resulting in a decrease in the total system cost. Thus, MAVs increase the energy efficiency and quality of services in the majority of system settings.

Finally, as the first step to investigate the potential equity impacts of the MAV technologies, we develop an analysis framework for quantifying the inequality impacts of AVs. The framework is built on a state-of-the-art multi-agent transportation simulator, MATSim, by incorporating the influence of individual demographics on travel decisions and private AVs. The simulation model generates individual-level daily travel itineraries that can be used to analyze benefits / costs of AVs. We apply the framework to analyze the inequality impacts of AV systems in the Tampa Bay Region in the 2040 planning scenario. Results reveal the capability of the proposed methodological framework in analyzing the inequality impacts of AV systems. Further, an AV system with private AVs and low market penetration rates may not improve the performance of a transportation system as expected due to the additional empty vehicle trips induced by AV operations. However, it leads
to a more even outcome distribution between homogeneous individuals and among the geographic space in the Tampa Bay Region. It may not change the disparity direction of the transportation outcome distributions between different population subgroups but will affect the magnitude of the disparity. Whether the disparity between groups will be widened or bridged depends on the specific outcomes and the groups being analyzed.

Overall, this dissertation offers scalable numerical and analytical tools for designing and analyzing MAV transportation systems. These models and algorithms can be used as benchmarks for researchers to evaluate the performance of their methods in future studies. They can also be used by transportation practitioners to design and analyze MAV transportation systems when the technology is mature. Further, findings from the extensive empirical case studies add to the body of knowledge of system properties of MAV transportation systems and their impacts to society including the level of service, energy implications, and inequality impacts. These results will offer managerial insights for future transportation planners and operators.
Chapter 1: Introduction

1.1 Demand Supply Asymmetry in Transportation Systems

The “asymmetry” between spatiotemporally varying travel demand and fixed capacity infrastructure has been a persistent problem in transportation systems. While most transportation facilities (e.g., transit stations, road segments) provide fixed capacity, the hourly travel demand for an infrastructure may vary over tenfold across a day (McNally, 2000). While most transportation activities aggregate on sparsely spaced highway corridors with low accessibility (FHWA, 1998; 2015) or infrequent transit services with limited spatial coverage (Lyons et al., 2017), travel demands are widely distributed across small parcels across the space and emerge continuously throughout the day (Wang, 2012). This asymmetry contributes to most notable transportation issues that compromise life quality, societal prosperity and sustainability, including energy wastes, excessive long commute time, traffic congestion, and social inequity.

To resolve relevant issues caused by this asymmetry, transportation professionals have created various innovative solutions in the past decades. A typical solution is to adjust the service frequency (e.g., transit timetabling, ride-sharing repositioning) so that the transportation capacity can be better aligned with the time-varying passenger demand (Caimi et al., 2017). However, time-varying service frequency alone is not a cure to the substantial gap between passenger demand and transportation capacity. Consequently, excessive queues at central business areas during peak hours (since demand is much greater than supply) and empty facilities in suburban and rural areas during off-peak hours (since supply is far greater than travel demand) are still commonly observed in transportation systems all over the world (Shi et al., 2018). Therefore, the problems associated
the demand supply asymmetry are still of great concerns to transportation planners and system operators. Innovative solutions are necessary to further address these issues.

1.2 Embracing a Modular Autonomous Vehicle Era

Recent advances in computing and artificial intelligence have enabled the development of various emerging vehicle technologies. Probably the most well-known is the autonomous vehicles (AV) - vehicles capable of driving with little or no human interference. AVs are expected to bring substantial mobility, environmental and public health benefits to society (Fagnant and Kockelman, 2014; Fagnant and Kockelman, 2015). Modular autonomous vehicles (MAV) integrate modularity operations with AV technologies to allow modular pods (MP) to be flexibly docked and undocked into MAVs of different sizes during operations. Additionally, the AV control technologies enable automated en-route docking (and undocking) operations at a relatively high running speed (that would otherwise be impossible with human drivers). These operations enable passengers make transfers en-route with negligible transfer impedance.

As shown in Figure 1.1, MAV is not only a concept, but also has been tested in field experiments by industry. Pilot experiments on autonomous rail rapid transit with adjustable train lengths were carried out in Zhuzhou, China (Lambert, 2017) and flexible grouped electric multiple units have been tested in China in 2017 (Rail, 2017). In addition to railroad transit, the modular vehicle concept has also been tested in highway transportation. For example, MAVs featured with elastic vehicle capacity designed by the Next Future Transportation Inc. have been tested in Dubai (Tarek, 2018) and Singapore (Ackerman, 2016). In Singapore, TUMCREATE, a research platform for Singapore’s public transportation, has initiated the investigation of dynamic autonomous road transit (Marinkovic and Tan, 2020). In addition, in 2019, Ohmio revealed its first Ohmio LIFT vehicle, an autonomous modular shuttle that can be flexibly adopted for different purposes (Ohmio,
Further, a simulation study (Caros, 2019) of providing MAV services under door-to-door operations and feeder-trunk service connecting Dubai and Sharjah showed that a profit increase by 20% over non-modular operations was possible.

Figure 1.1: MAV prototypes in industry. (a) MAV prototypes designed by NEXT (Source: http://www.next-future-mobility.com/), (b) Autonomous rail rapid transits in China (Source: https://designmuseum.org/), (c) TUMCREAT’s DART module prototype (Source: https://www.tum-create.edu.sg/research/design-autonomous-mobility), (d) Ohmio LIFT (Source: https://ohmio.com/news/the-first-ohmio-lift-is-revealed-by-ohmio).

1.3 How Will MAVs Change Transportation Systems?

The MAV technology offers a unique solution to mitigate the demand supply asymmetry in transportation systems. With the MAV’s modularity design, vehicles can flexibly change their sizes to “cheat through” fixed infrastructure capacity. Specifically, vehicle capacity can be flexibly adjusted along with the service frequency according to the passenger demand level; long vehicles are used in central business areas during peak hours while short vehicles are dispatched in suburban and rural areas during off-peak hours. This way, the excessive passenger waiting (when demand
surpasses capacity) and the unnecessary energy wasted on moving vehicles with low occupancies (when capacity exceeds demand) can both be decreased.

In light of these great potentials, as opposed to an existing transportation system with fixed capacity vehicles, this dissertation envisions a next-generation transportation system that features an elastic vehicle capacity operation. MAV-based elastic capacity transportation systems allow MPs to be properly combined and separated to bridge the gap between the infrastructure capacity and passenger demand. As a result, this next-generation transportation system transforms vehicles into an “elastic” medium to reconcile the asymmetry between travel demand and transportation infrastructure. This medium dampens highly fluctuating travel demand over space and time by consolidating them into vehicles of dynamic sizes and occupancies with minimum access and transfer impedance. Such a simple innovation may substantially increase the quality of services (by reducing passenger waiting time) and energy efficiency (by increasing vehicle ridership) of transportation systems.

However, as other emerging transportation technologies and services (e.g., bike sharing), outcomes from the MAV (and AV) technologies may not necessarily be equally distributed among society. It is possible that these outcomes are disproportionally received by certain population groups, therefore worsening the status quo for traditionally transportation-disadvantaged groups. The equity/inequality impacts of these technologies would be a valid concern for transportation planners and policy makers as our society steps into the modular and autonomous vehicle era.

1.4 Research Goal and Scientific Questions

Despite the promising outlook, MAV (and AV) is not yet operational, and thus it is open questions how to design MAV transportation systems and what their impacts on transportation systems are. Answering these fundamental research questions are critical for realizing the vision
of the MAV transportation system. Thus, the overarching goal of this dissertation is to develop methodologies for designing and analyzing the next-generation transportation systems with the emerging MAV technologies as well as to understand the impacts of these systems on society. Key scientific questions of interests include:

- Given the heterogeneity and complexity of transportation systems, what are the tools and techniques suitable for optimally designing MAV transportation systems and understanding the capacities of these systems?
- Are there shared properties of the optimal system design for MAV transportation systems? If yes, what are they? How can these properties offer managerial insights to system operators?
- What are the capacities of MAV transportation systems? How does a transportation system embracing the elastic capacity operations of MAVs behave differently from existing fixed capacity transportation systems?
- What are the potential impacts of the MAV transportation systems in terms of key system performance metrics such as quality of service, energy cost (or more broadly vehicle operational cost), and resource utilization rate?
- How do system design and the performance of MAV transportation systems interact with various input parameters such as the passenger demand level, cost parameters, and resource restrictions?
- How will the potential impacts of transportation systems with emerging vehicle technologies (e.g., AV, MAV) be distributed among different user groups in society?

To answer these questions, this dissertation investigates the design of two transportation system structures with the MAV technologies. We start with the simplest travel demand structure
– a shuttle system consisting of one origin and one destination. Then, a more complex corridor system with a set of physical stations representing multiple origin-destination pairs is investigated. Additionally, we develop an agent-based simulator to capture the operational characteristics of AVs in a metropolitan area. This can be extended to study the inequality impacts of MAV transportation systems. This dissertation is the first step in achieving the overarching goal. Many additional research questions related to MAV transportation systems are also discussed at the end of this dissertation.

1.5 Dissertation Overview and Contributions

This dissertation introduces and develops scalable computational and analytical models to address concrete challenges in the proposed next-generation transportation systems, as shown in Figure 1.2. We will start with a literature review in Chapter 2, including reviews of studies on the operational design paradigms of transportation systems, the discrete optimization methods, the continuous approximation methods, and the equity impacts of emerging vehicle technologies. Following the literature review are the outcomes from three specific research topics.

Figure 1.2: Summary of dissertation contents.
Chapter 3 first demonstrates the MAV concept in the simplest travel demand structure – a shuttle system consisting of one origin and one destination. This work demonstrates that a MAV-based transportation system yields vastly reduced system costs including vehicle dispatch cost and passenger waiting cost, as compared to existing systems. More importantly, it establishes a general methodological framework for tackling the operational design problem for more complex elastic capacity mobility systems.

Chapter 4 investigates a more common and complex mobility system setting – a corridor system with a set of physical stations representing multiple origin-destination pairs. The innovation of this system setting is that MAVs can dock and undock at any station along the corridor, which we call station-wise docking in this dissertation. By casting the system design problem into a similar methodological framework proposed in Chapter 3 but with different mathematical efforts, we propose a discrete model and a continuous model for MAV corridor system design. Also, we demonstrate the potential benefits for station-wise docking via extensive numerical experiments.

Chapter 5 designs a methodological framework to study the inequality impacts of the emerging AV technologies. We develop a sophisticated agent-based simulation model based on a state-of-the-art multi-agent transport simulator, MATSim, by considering key operational characteristics of AVs. This simulation model is used to analyze the inequality impacts of AV technologies on system users in the Tampa Bay Region in the 2040 planning scenario. The analysis in this chapter sets up a general methodological framework for investigating the equity impacts of emerging AV-related technologies and thus can be extended for MAV equity studies.

Finally, Chapter 6 closes this dissertation with a discussion on other questions that need to be answered before deploying the MAV transportation systems. We summarize directions of future research for MAV transportation systems.
Chapter 2: Literature Review\(^1\)

This chapter provides a comprehensive review of the relevant literature. First, a review of operational design paradigms for urban mass transportation (UMT) systems is provided in Section 2.1. Then, Section 2.2 reviews existing studies on the continuous approximation methods for UMT operational design. Next, Section 2.3 presents a review of discrete optimization methods for UMT operational design. Finally, Section 2.4 reviews equity studies on emerging vehicle technologies.

2.1 Operational Design Paradigm

This dissertation proposes to use vehicles as a medium linking transportation infrastructure and demand. In the independent operational design of a specific mode of transportation, multitudes of efforts have been made to close the demand-supply gap via various strategies, e.g., timetabling in UMT systems (Yin et al., 2017), driver-rider matching in ride-sharing systems (Agatz et al., 2011), and vehicle repositioning in shared mobility systems (Pal and Zhang, 2017). This section reviews different operational design paradigms for UMT timetabling, the strategy of interest in this dissertation.

Timetabling is an important topic in urban mass transportation studies on which extensive research efforts have been made in the past decades. Due to the lack of detailed passenger demand information, a constant headway paradigm is first adopted (Ghoneim and Wirasinghe, 1986; Ceder, 2001; Yu et al., 2011; Szeto and Jiang, 2014), with the assumption that the passenger arrival demand remains constant during the operational horizon. This simple strategy allows passengers

\(^1\) Portions of this chapter has been previously published in Chen et al. (2019a) and Chen et al. (2020). Permission is included in Appendix A.
to memorize the departure times but does not match realistic demand patterns that demonstrate strong temporal fluctuations (Niu and Zhou, 2013; Niu et al., 2015). Therefore, Ceder (1984; 1987; 2001) highlighted the significance of passenger demand dynamics and put forward demand-driven timetabling, which can be further divided into periodic timetabling and non-periodic timetabling. In the periodic timetabling problem, the operational horizon is split into several time blocks (e.g., peak hours and off-peak hours) where a timetable is repeated (Banks, 1990; Liebchen, 2007; Liebchen et al., 2008; Li and Lo, 2014; Yang et al., 2016; Guo et al., 2017a). This method, in essence, assumes that the passenger arrival demand is stationary in each time block, yet cannot suit highly time-varying demand within each time block in the real world (Sun et al., 2014). As a result, recent studies have demonstrated increasing interest in non-periodic timetabling that adjusts the transit service frequency dynamically to better fit the temporal demand heterogeneity (Niu and Zhou, 2013; Barrena et al., 2014a; Barrena et al., 2014b; Niu et al., 2015; Gao et al., 2016; Yin et al., 2016; Yin et al., 2017; Shi et al., 2018).

Regardless of the abundant studies on transit timetabling that provide valuable insights into the operations of UMT systems with evident spatiotemporal demand fluctuations, only a handful of them has considered the option of using vehicles as an elastic medium linking infrastructure and demand. This consideration requires incorporating the vehicle capacity into the decision space of the operational design models so that it can be designed with the dispatch headway (or equivalently, frequency) simultaneously to improve system performance. For example, Albrecht (2009) adopts a two-step model to determine the train capacity and dispatch headway sequentially. However, this sequential approach may lose certain optimality due to a lack of consideration of the interactions between capacity and headway. In a series of seminal works by Ceder and his colleagues (e.g. Ceder, 2011; Hassold and Ceder (2014)), they propose heuristic methods to create transit
timetables featured with even headways and even loads by considering multiple vehicle types, but exact solution methods have not been discussed. Recently, Guo et al. (2017b) propose an analytical stochastic and dynamic model for transit switching in an autonomous vehicle context but only two types of vehicles have been considered. Moreover, the vehicle operational cost, particularly energy consumption, has not been investigated in their paper. Therefore, the joint design of dispatch headway and vehicle capacity with an elastic capacity paradigm is still a new topic. Fundamental methodologies addressing this topic have yet to be investigated.

Overall, an elastic capacity operation paradigm in transportation systems that will likely be realized with the emerging MAV technology has not been well investigated in the literature. This new paradigm may further balance the tradeoff between the generalized operational cost (including congestion effects, fuel consumption and environmental impacts) and service quality. Proposing a transformative operational framework for enabling the MAV service and harvesting their expected benefits is worth investigating for the benefit of future transportation systems. Further, without proper methodologies to solve the operational design for MAV transportation systems, it is hard to study whether the MAV technology will benefit society as expected. Thus, the development of such methodologies is also important for answering this science question via empirical studies.

2.2 Continuous Approximation Method

A popular approach for UMT operational design is the classical continuous or continuum approximation (CA) model proposed by Newell (1971) for the first time to design the optimal dispatch headways for a transportation route with unsaturated traffic. This section reviews related literature of the CA method.

The CA method approximates discrete variables with continuous smooth functions. Due to the relatively homogeneous setting of the variables in a local neighborhood (a time interval for a
time horizon, a sub-region for a two-dimensional space), the decisions in each neighborhood are mainly affected by settings in nearby neighborhoods rather than those in distant neighborhoods. Thus, the original problem can be decomposed into a finite number of separable sub-problems. The resultant sub-problems can usually be formulated with simple multivariate functions that can be analytically solved independently, rendering the CA approach promising in tackling large-scale instances. Indeed, we will see in the remainder of this dissertation that the computational time of a CA model only linearly increases with the number of sub-problems, which is satisfactory for engineering applications. Besides, the analytical solution to a CA model explicitly quantifies the tradeoff between critical system cost components and their relationship with the input parameters. Such a simple relationship reveals important managerial insights for UMT operators to understand the problem structures and even design operational plans manually when advanced computational resources are not available. For example, Newell (1971) shows that the dispatch headway of a transportation route should be proportional to the square root of the passenger arrival demand. Thus, UMT operators can then roughly estimate the operational headway based on the historical passenger arrival counts rather than solving a complicated optimization model.

Due to these advantages, the CA approach has been extended to develop operational plan for UMT system with more realistic constraints, e.g. round-trip constraints and fleet size (Salzborn, 1972; Hurdle, 1973a; Hurdle, 1973b), capacitated vehicles (Sheffi and Sugiyama, 1982), bus bunching (Barnett and Kleitman, 1973; Newell, 1974; Daganzo, 2009), multiple periods (Chang and Schonfeld, 1991), transit line spacing (Hurdle, 1973c), interchange (Salzborn, 1980), many-to-many time-varying demand (Wirasinghe, 1990), skip-stop operations (Freyss et al., 2013), etc. Recent studies have applied the CA method to more complicated transit network design problems where the dispatch headway serves as a decision variable (Daganzo, 2010; Estrada et al., 2011;
Ouyang et al., 2014; Chen et al., 2015; Fan et al., 2018; Luo and Nie, 2020). Besides, the CA method has been extensively applied in facility location problems (Li et al., 2016) and logistics distribution studies (Daganzo, 1987). Interested readers are referred to Anasari et al. (2017) for a recent review of the CA methods.

Despite the abundant advancements in the CA method, few of them can be directly applied to the MAV-based transportation systems investigated in this dissertation. A fundamental premise of the CA framework to be applicable is the “local-impact property”; i.e., decisions in a local neighborhood are mainly affected by settings in this and nearby neighborhoods but not much by distant ones. This “local-impact” property, yet, is not satisfied in the MAV systems investigated in this research owing to various factors, e.g., the oversaturated traffic in the shuttle system, the correlations of the passenger demands with different origins and destinations across the multiple stations in the corridor system. Take the simplest shuttle system under oversaturated traffic as an example. Under oversaturated traffic, passengers may not be able to board the first vehicle reaching their station at its departure time, and thus the passenger queue is built up when this happens. The queuing mechanism actually couples decisions in relatively distant (time) neighborhoods since the current dispatch decision may significantly affect the queuing state and the corresponding dispatch decision in a relatively distant future. Additionally, the consideration of time-varying (or elastic) vehicle capacity poses another challenge since a vehicle can be assembled with different numbers of MPs. This consideration will make the vehicle dispatch cost a variable related to the capacities of vehicle dispatched. However, no CA study has considered such elastic vehicle capacity and the associated dispatch cost implications in the operational design context. Thus, there still lacks a CA method that efficiently solves the operational design for an elastic capacity transportation system in real-world settings and offer simple analytical insights into the fundamental problem structure.
To sum up, CA has been a powerful analytical tool for addressing large-scale operational design problems for UMT systems. However, existing CA methods can hardly be directly applied to the proposed next-generation transportation system with MAVs due to the invalidity of the local impact property. Fundamental research efforts on formulating the operational design problem for the MAV-based elastic capacity transportation system under the CA framework are needed.

2.3 Discrete Optimization Method

Apart from the CA method, another popular modeling approach is discrete optimization, whose application in UMT design can date back to the seminal work by Bisbee et al. (1968). This section reviews recent advances of the discrete optimization method in UMT operational design.

By discretizing the operational horizon into a finite number of time intervals, the discrete optimization method formulates the problem with a set of discrete and/or continuous variables. With this, microscopic operational details in the system such as the minimum dispatch headway, passenger boarding and alighting, vehicle circulation, fleet management can be explicitly modelled (Niu and Zhou, 2013). With a properly defined objective function, the optimal operational design can then be numerically solved (Lin and Kwan, 2016; Niu and Zhou, 2013; Niu et al., 2015; Wang et al., 2015; Xu et al., 2017). However, due to the combinatorial nature of the problem and the large problem scale in the real world, UMT operational design has been well acknowledged as a NP-hard problem (Caprara et al., 2006; Xu et al., 2017; Sun et al., 2014). As a result, extensive efforts have been seen to develop efficient solution approaches to solve the discrete models.

Generally speaking, existing studies adopting the discrete optimization methods for UMT operational design can be divided into three categories. The first category follows the traditional mathematical programming method by formulating the problem into integer programming models and solving the model via commercial solvers (e.g. CPLEX (Sun et al., 2014), GAMS (e.g., Niu
et al., 2015)) or customized algorithms (e.g. branch and bound (Albrecht, 2009), branch and cut (Barrena et al., 2014a), branch and price (Lin and Kwan, 2016), Lagrangian decomposition (Zhou and Teng, 2016)). These mathematical programming methods are still very popular because of the solid theoretical foundation and the guarantee in solution optimality or convergence performance (Zhou and Teng, 2016). However, these programming methods require expensive computational resources to reach the exact optimal solution for large-scale real-world cases. In many cases, the exact optimal solution cannot be found with the limited computational resources. To address this issue, the second category develops efficient heuristic approaches to solve large-scale instances, e.g., Yin et al., (2016), Niu and Zhou, (2013), Li and Lo, (2014), Szeto and Jiang, (2014), Gao et al., (2016), Guo et al., (2017b), Shi et al., (2018), Ceder (2011), Hassold and Ceder (2014), etc. Yet, these heuristics typically end up being stuck in a local optimum, without a guarantee on the solution quality theoretically. For example, Niu and Zhou (2013) proposed a genetic algorithm for timetabling design in an oversaturated metro corridor. However, there is no guarantee that the gap between the solution from the genetic algorithm will be close to the exact optimal solution to the problem. Additionally, due to the difficulty in formulating some operational constraints as rigorous mathematical formulas required by mathematical programming models and the desire to improve the computational efficiency, some studies use the simulation techniques for UMT timetabling (e.g. Adamski and Turnau (1988); Yang et al., (2016); Ma et al., (2019)).

Regardless of these advancements in the discrete optimization methods, the operational design in an MAV-based elastic capacity transportation system is still computationally challenging. This is because in a MAV-based transportation system, we have to consider the option of multiple vehicle capacities under specific UMT system structures, especially in real-world instances. Take the simplest MAV-based shuttle system where oversaturated traffic is present as an example. On
one hand, when the demand is oversaturated, a long queue waiting for boarding will form, which will much increase the solution space of a traditional exact discrete model that needs to incorporate all possible queue lengths growing anywhere from 0 to the total cumulative demand. Thus, the solution space of queuing variables is actually unbounded as time elapses, which makes the search for an exact optimal solution in real-world cases, if not impossible, extremely difficult. On the other hand, the consideration of flexible capacities further complicates the dynamics of the queue variables: one dispatch may result in several different reductions of the queue length due to different vehicle capacities. These two factors would result in an intimidating variable space for the investigated problem. It is expected that the problem would be even more challenging in other UMT systems with more complicates settings, e.g., a MAV UMT corridor. Hence, there still lacks efficient discrete optimization models that can solve the exact solution(s) to the operational design in an elastic capacity transportation system in real-world contexts.

In summary, numerous plausible efforts have pushed the research frontiers on discrete optimization for the operational design of existing transportation systems with fixed capacity infrastructures. However, incorporating vehicle capacity as decision variables in the operational design model for a MAV-based transportation system challenges the direct application of existing methods. Studies on discrete optimization methods for MAV transportation system are therefore necessary. Such studies will be important for UMT operators to design future UMT systems and also for researches who need a benchmark method for assessing their research efforts in solving near-optimal operational design for MAV transportation systems.

2.4 Equity Studies on Emerging Vehicle Technologies

In the transportation literature, equity is broadly divided into horizontal equity and vertical equity (Chen et al., 2019b). Horizontal equity is concerned with the distribution of outcomes
among homogeneous individuals or groups spread across space; a distribution is horizontally equitable if everyone within a population receives the same benefits or pays the same costs per unit of benefit received. Vertical equity accounts for differences in individuals’ sociodemographic characteristics, and/or capabilities; a distribution is vertically equitable if disadvantaged groups receive more benefits or pay fewer costs to bring them to the same level as other groups.

Numerous studies have been conducted to investigate the equity performance of different transportation modes. These include examinations of private cars (Guzman et al., 2017), public transportation (Delbosc and Currie, 2011; Guzman et al., 2017), bike sharing (Chen et al., 2019b; Qian and Niemeier, 2019), car sharing (Shellooe, 2013), ride-hailing (Hughes and MacKenzie, 2016), bicycling (Hamidi et al., 2019; Pritchard et al., 2019) and AVs (Cohn et al., 2019). Among the various modes of transportation, equity impacts of traditional systems (e.g., private cars, public transits) are relatively well studied, while research on equity impacts on emerging transportation technologies and services (e.g., TNC, AVs, MAVs) is scarce and mostly focused on bike sharing.

There are few published studies on the equity impacts of emerging vehicle technologies. We have found only one study investigating the equity dimension of AVs to date. That is, Cohn et al. (2019) adapted the four-step Metropolitan Washington DC MPO regional travel demand model to assess the regional changes in job accessibility, trip duration, and trip distance in AV systems across eight scenarios. Other studies adapted four-step models to investigate other impacts of AVs or shared automated vehicles (SAVs) (Zhao and Kockelman 2018; Dias et al. 2019). For example, Zhao and Kockelman (2018) modified the travel demand model data from the Capital Area Metropolitan Planning Organization to study the potential impacts of AVs in Texas, Austin. In these studies, model representations of AVs usually assume AV adoption rates and changes caused
by AV adoption to roadway capacity, value of time, vehicle ownership and policy, parking and operating cost (Cohn et al., 2019).

Recent studies on transportation equity have demonstrated the benefits of incorporating high resolution disaggregated data. This is possible thanks to advancements in travel demand modeling (Bills and Walker, 2017; Chen et al., 2019b; Gurram et al., 2019). Microsimulation-based travel demand models such as the CT-RAMP2 platform and MATSim (Horni et al., 2016) explicitly track individual daily travel itinerary and sociodemographic data, which permits fine-grained analysis of the equity impacts of transportation systems. In terms of emerging vehicle technologies, disaggregated approaches using agent-based travel demand models have been used in modeling AVs. A series of seminal papers by Dr. Kara Kockelman and her team utilized microsimulation-based modeling to analyze AVs’ impacts on society in a variety of contexts (e.g., Fagnant and Kockelman, 2014; Fagnant and Kockelman, 2015; Perrine et al., 2012; Loeb and Kockelman, 2019; Liu et al., 2017). As follows, it would be beneficial to incorporate activity- and microsimulation-based modeling to investigate the equity impacts of AV/MAV. However, as far as the author’s knowledge, there is no study using disaggregated models to investigate the equity dimension of AV/MAV transportation systems.

In conclusion, compared with traditional transportation systems, emerging transportation systems using innovative vehicle technologies have not been well investigated in terms of their equity impacts on society. The few available studies have adopted aggregated models for analysis while the more sophisticated and increasingly popular disaggregated modeling approach has not been considered. It is of urgent need to develop disaggregated analytical tools to facilitate the understanding the equity impacts of such systems on system users.
Chapter 3: Shuttle Systems

3.1 Overview

This chapter proposes a MAV shuttle system that serves passengers departing from one origin to one destination with possible presence of oversaturated traffic. Different from existing shuttle systems served by fixed capacity vehicles, the proposed MAV shuttle system jointly designs vehicles’ dispatch headway and formation in response to the time-varying passenger demand. The idea is that MAVs can be docked and undocked flexibly, so that we can dynamically change the vehicles’ capacity according to the passenger demand level. This way, not only the service quality can be improved (frequent dispatches with small vehicles reduce passengers’ waiting costs) but also the energy efficiency can be increased (higher vehicle occupancies with appropriate capacities save unnecessary energy consumption).

The following motivating example illustrates the concept of the MAV-based shuttle system and its potential in improving system performance. Consider a system with cumulative passenger demand as the black curve in Figure 3.1 shows and modular vehicles are dispatched to serve the passengers. Assume the capacity of 1 vehicle unit is 5 passengers and the energy cost of dispatching a vehicle consisting of 1 and 2 modular units are 5 and 9, respectively. The passenger waiting cost is 1 per passenger per unit time. If traditional fixed capacity vehicles are used; i.e., only vehicles consisting of 2 units can be dispatched, then the optimal solution is to dispatch a vehicle with 2 units at times 2, 3, 4, respectively, leading to a cumulative passenger departure curve as the blue

---

2 Portions of this chapter has been previously published in Chen et al. (2019) and Chen et al. (2020). Permission is included in Appendix A.
curve in Figure 3.1. This design results in a total waiting cost of 20.5 and a total energy cost of 27. Note that the passenger waiting time is measured as the area between the cumulative arrival and departure curve. However, if MAVs that allow dynamic capacity design are used, the optimal solution is to dispatch vehicles consisting of 1 modular unit at times 1, 2, 4 and a vehicle consisting of 2 units at time 3. This solution is represented as the dashed red line in Figure 3.1. This strategy results in a total waiting cost of 15.5 and a total energy cost of 24. Thus, the MAV-based shuttle system is expected to reduce both the total energy cost and the total waiting cost compared with traditional fixed capacity shuttle systems.

![Figure 3.1](image)

Figure 3.1: A motivating example for MAV shuttle systems.

This chapter studies the operational design (i.e., how to jointly design the vehicle dispatch headway and formation) in an optimal manner for the proposed MAV shuttle system. This system considers oversaturated traffic where passengers may have to wait for multiple vehicles to board, which is a frequently observed phenomenon in UMT systems (Niu and Zhou, 2013). Also, by comparing the performance of the optimal design with existing fixed capacity vehicles and that with the proposed varying capacity MAV operation, we aim to quantitatively analyse the impacts
of MAV technologies on the performance of shuttle transportation systems. The contributions of this chapter are summarized as follows.

(i) We formulate the operational design problem into a continuous approximation model and a discrete mixed integer linear programming model. The CA model presents a macroscopic view of the system and yields simple analytical rules into the optimal design. These analytical results will enable efficient solution methods to searching for near-optimal solutions for relevant large-scale transportation problems and offer managerial insights to system operators. The discrete model, instead, takes microscopic discrete inputs and aims to solve the exact optimal design. It will enable UMT operators to solve the true optimum for small- and medium-size problem instances and offer a benchmark to evaluate the continuous model.

(ii) We prove elegant theoretical properties of the investigated problem, which show that the original operational design problem can actually be solved by just solving its corresponding revised unsaturated problem. This property offers a theoretical foundation for the CA model and allows us to decompose the original problem into a set of independent unit-time revised unsaturated problems that can be analytically solved in each neighborhood. This innovation extends the CA methodology literature from traditional transit dispatching problem with unsaturated traffic to the joint design of dispatch headway and vehicle capacity considering oversaturated traffic, adjustable vehicle capacities and other factors (e.g., minimum dispatch headway). These theoretical properties are also used to mathematically formulate the relationship between the queue length and unit vehicle capacity, which result in a set of valid inequalities. With these valid inequalities, we develop an efficient customized dynamic programming (DP) algorithm to solve the discrete model. These valid inequalities provide a tight upper bound to the originally unboundedly increasing state space and thus can largely reduce the solution space of the DP algorithm.
(iii) We conduct two sets of numerical experiments with realistic traffic data from the Beijing Subway System and the Tampa Bay Area to test the proposed model and algorithm. The computation speed of the customized DP algorithm is shown to be much faster than a state-of-the-art integer programming commercial solver, Gurobi. Meanwhile, the CA model can produce highly accurate near-optimal solutions in almost no time compared with the exact solutions. It is also shown that incorporating time-varying capacity adjustment via MAVs into transit scheduling can improve the vehicle utilization rates and reduce the total system cost (i.e. the total of energy cost and passenger waiting cost). Further, sensitivity analyses reveal that the input parameters affect the effectiveness of the proposed MAV shuttle system under oversaturated and unsaturated traffic in a similar way while the unsaturated case shows more robustness.

The remainder of this chapter is organized as follows. Section 3.2 introduces the studied operational design problem for shuttle systems. Section 3.3 investigates theoretical properties of the problem. Based on the theoretical properties, Section 3.4 presents the CA model and the analytical solution approach. Section 3.5 presents the discrete model and the DP algorithm. Section 3.6 presents the results from the numerical experiments. Finally, Section 3.7 concludes this chapter.

3.2 Problem Statement

This chapter considers a UMT shuttle system (e.g. subways, bus rapid transits, autonomous rail transits, modular autonomous vehicles) with one origin and one destination over an operational time horizon \([0,T]\). The time-varying daily passenger demand at the origin is described as a cumulative arrival demand curve \(A(t), \forall t \in [0,T]\) as shown in Figure 3.2, and thus the arrival demand rate at \(t \in [0,T]\) is \(a(t) := A'(t)\). A group of MAVs in different formations \(J := [1,2,\cdots,I]\), indexed by \(i \in J\), can be dispatched at the terminal. Each vehicle formation \(i \in J\) has \(i\) identical MPs and thus a capacity of \(ic\), where \(c\) is the capacity of a MP. To serve the passenger
demand, a number of $K$ vehicles are dispatched at the origin and head to the destination over the operational horizon, and we index these dispatches as $\mathcal{K} := [1,2,\cdots,K]$ where the index increases with the dispatch time. Following Newell (1971), we assume that the availability of vehicles at the origin is independent of when previous vehicles have been dispatched. Therefore, the fleet size at the depot is always sufficient so that there are always some vehicles available in each formation for dispatching at the origin. Please note that the fleet size (i.e., the number of vehicle units) can be determined after solving the investigated operational problem. Therefore, in the planning stage, we can just procure vehicles according to this fleet size (which could be further multiplied by a certain factor for reliability). Since the fleet planning problem is a separate problem in this investigated context, we will focus on vehicle operations assuming that a sufficient number of vehicles are provided. Due to limited resources and operational safety, we assume that the minimum headway between every two consecutive dispatches is $h$. These dispatches result in a cumulative departure curve from the origin, denoted by $D(t), \forall t \in [0,T]$. The vehicle formation and dispatch time of each dispatch $k \in \mathcal{K}$ are denoted by $i_k$ and $t_k$, respectively. For the convenience of notation, define $t_0 := 0$ and $t_{K+1} = T$.

Figure 3.2: Cumulative passenger counts in MAV shuttle systems
Since vehicles are only dispatched at discrete time points separated by a minimum interval of $h$ while passengers arrive continuously, passengers need to wait before boarding a vehicle. Following previous studies (Niu and Zhou, 2013; Yin et al., 2017), we adopt the total passenger waiting cost to measure the service quality of a UMT shuttle system. Note that in Figure 3.2, the time separation between $A(t)$ and $D(t)$ for the same passenger cumulative count denotes the waiting time for this passenger, and thus the total passenger waiting time is just the shaded area between $A(t)$ and $D(t)$. We assume that each passenger has an identical unit-time waiting cost $w$, and thus the passenger waiting cost is the product of the shaded area in Figure 3.2 and $w$.

Besides the customer waiting cost, the other cost component we consider is the energy cost (including the traction energy and auxiliary energy) (Huang et al., 2017) consumed by each vehicle from the origin to the destination. The investigated problem only considers the portion of energy associated with a vehicle’s mass but not that associated with the number of passengers on board. This is because, first, the energy consumption is actually mainly determined by vehicle mass but not much affected by passengers on board (Zhao et al., 2017). Second, the total transported passengers (i.e., $A(T)$) in this problem is independent of the vehicle dispatching schedule, and thus the associated energy consumption shall not be affected by the dispatch decisions and can be removed from the optimization. Therefore, since the vehicle mass is determined by the vehicle capacity (or the number of vehicle units), we denote the energy cost of a vehicle in formation $i$ as $f_i > 0$ and assume it is concave over $i$ to account for the economies of scale (Cohen and Moon (1991); Holmberg and Tuy (1999)), i.e.,

$$\lambda f_i + (1 - \lambda)f_j \leq f_k, \forall i, j, k := \lambda i + (1 - \lambda)j \in J \cup \{0\}, \lambda \in [0,1],$$

(3.1)

where $f_0 := 0$. One common example is that $f_i = C^F + C^V (i)^{\alpha}, \forall i \in J$. In this function, the first term $C^F$ is the fixed energy cost regardless of the vehicle capacity (or the number of vehicle units),

\[f_i = C^F + C^V (i)^{\alpha}, \forall i \in J.\]
which accounts for, e.g., the locomotive cost and some auxiliary energy consumption. The second term refers to the variable energy cost attributable to the number of vehicle units. We just use a simple function $C^V(i)^\alpha$ for the variable energy cost, with $C^V$ being a positive coefficient (to account for a single unit cost) and power index $\alpha \leq 1$. Note that this study primarily focuses on an operational problem in UMT systems while fleet size design is more of a planning level problem. Further, as long as a sufficient number of vehicles are available for operations, the fleet cost, albeit considerable, does not much affect operational decisions. As pointed out in Hurdle (1973a), in a unidirectional shuttle system where vehicles make only one-way trips, it is legitimate to use a model with only two costs, for passenger waiting time and for vehicle operation. Thus, we do not consider the cost of owning the fleet. Also, this assumption is not restrictive since other cost components related to vehicle dispatch, e.g., capital costs for vehicles, driver costs, crew salaries and vehicle reconfiguration costs, can be easily incorporated into the general cost structure (3.1) without changing the problem structure. Therefore, for the simplicity of modeling, the investigated problem only considers the energy consumption.

The objective of the operational design problem, then, is to find an optimal arrangement on both the formations of vehicles dispatched, i.e. $i_k$, and the corresponding dispatch times, i.e. $t_k$, $\forall k \in \mathcal{K}$ during $[0, T]$. Note that along with $i_k$ and $t_k$, $\forall k \in \mathcal{K}$, the number of dispatches $K$ and cumulative departure curve $D(t), \forall t \in [0, T]$ , which are also decision variables, will be determined. This optimal arrangement aims to achieve the best trade-off between the vehicle energy cost and the passenger waiting cost. The significance of such a trade-off has been well explained in Yin et al. (2017). Moreover, to capture the time-variant passenger process, we denote the vehicle load, i.e., the number of passengers boarding at dispatch $k$ as $d_k, \forall k \in \mathcal{K}$. With these decision variables, we can now formulate the objective function as
\[
\min_{K, [t_k, i_k, d_k], \forall k \in K} \sum_{k \in K} f_{i_k} + w \int_0^T (A(t) - D(t)) dt.
\]  

This objective function aims to search for the optimal formation of vehicle and time for each dispatch such that the sum of the energy cost and the waiting cost across the operational horizon can be minimized. In addition, to describe the dynamic operation of a UMTS, we consider four groups of constraints as follows.

(i) Minimum headway requirement. These constraints are imposed to ensure the least time separation between two consecutive vehicles due to the safety consideration.

\[
t_k - t_{k-1} \geq h, \forall k \in K \backslash \{1\}.
\] (3.3)

(ii) Determination of the vehicle load \(d_k, \forall k \in K\), which is the minimum among the difference between the cumulative arrival demand at \(k\) and cumulative departure at dispatch \(k - 1\) and the capacity of the vehicle dispatched at \(k\).

\[
d_k = \min\{A(t_k) - D(t_{k-1}), i_k c\}, \forall k \in K.
\] (3.4)

(iii) Departure curve conservation. These constraints are imposed to describe the passenger departure dynamics. Constraint (3.5) is the initialization condition while Constraints (3.6) describe the departure conservation process. Constraint (3.7) ensures that all passengers are transported at the end of the operational horizon. Note that based on Constraint (3.7), to ensure the feasibility of the investigated problem, the arrival demand pattern in the shuttle system must satisfy \(\frac{A(T)}{T} < \frac{cI}{h}\). Otherwise passengers cannot be cleared at \(T\) even if the maximum transportation capacity is used.

\[
D(0) = 0,
\] (3.5)

\[
D(t) = \begin{cases} D(t_{k-1}) + d_k, \forall t = t_k, k \in K; \\ D(t_k), \forall t \in (t_k, t_{k+1}), k \in K \cup \{0\}, \end{cases}
\] (3.6)

\[
D(t_K) = A(T).
\] (3.7)

(iv) Feasible regions. These constraints define the domains of \(K, t_k\) and \(i_k\), respectively.
3.3 Theoretical Property Analysis

This section investigates some theoretical properties of the optimal solution to the above-defined problem, which will serve as the cornerstone for developing efficient algorithms in the following sections. The proofs of the following lemmas and propositions are available in Appendix B.

Lemma 3.1. For \( f_i \) satisfying concave property (3.1), we obtain
\[
f_{i_1} + f_{i_2} + \cdots + f_{i_n} \geq f_{i_1 + i_2 + \cdots + i_n}, \forall i_1, i_2, \cdots, i_n, i_1 + i_2 + \cdots + i_n \in \mathcal{I}, n \in \mathbb{Z}^+.
\]

Lemma 3.2. For \( f_i \) satisfying concave property (3.1), we obtain
\[
f_{i_1} + f_{i_4} \leq f_{i_2} + f_{i_3}, \forall i_1 \leq i_2 \leq i_3 \leq i_4 \in \mathcal{I} \text{ and } i_2 + i_3 = i_1 + i_4.
\]

With this, we will specify some optimal solution properties in the following propositions.

Proposition 3.1. In an optimal dispatch solution \( \{ t_k, i_k \}_{k \in \mathcal{K}} \) to problem (3.2) ~ (3.8) with arrival curve \( A(t) \) and departure curve \( D(t) \), if \( A(t_k) - D(t_{k-1}) \geq Ic \), then \( i_k = I, \forall k \in \mathcal{K} \).

Proposition 3.2. In an optimal dispatch solution \( \{ t_k, i_k \}_{k \in \mathcal{K}} \) to problem (3.2) ~ (3.8) with arrival curve \( A(t) \) and departure curve \( D(t) \), if \( A(t_k) - D(t_{k-1}) > Ic \), then \( t_k - t_{k-1} = h, \forall k \in \mathcal{K} \).

Propositions 3.1 and 3.2 indicate that in an optimal solution, the best policy is always to dispatch vehicles in formation \( I \) with \( h \) when the number of passengers who want to board is greater than the longest vehicle’s capacity. Note that these findings are well aligned with those in Newell (1982) and Daganzo (1997), which claim that the maximum service rate shall be applied to serve a system under oversaturation. To more rigorously analyze this, for any given arrival curve \( A(t) \), we define the oversaturated and unsaturated periods as follows.

Definition 3.1. For a given arrival curve \( A(t) \), the oversaturated period set denoted as \( T^S := \{ t | t \in [\mathcal{O}(z), \mathcal{E}(z)), \forall z \in [1, |\mathcal{O}|] \} \) and the unsaturated period set as \( T^U := [0, T] \setminus T^S \), where
$O(z)$ is the starting moment and $E(z)$ the ending moment of the $z$-th oversaturated period, respectively. $O$ and $E$ are obtained with Algorithm 3.1.

---

**Algorithm 3.1. Oversaturation Detection with Continuous Inputs**

**Input:** $T; I; c; h; a(t), A(t), \forall t \in [0, T]$

1. $O \leftarrow \emptyset$
2. $E \leftarrow \emptyset$
3. $SSP \leftarrow 0$
4. $z \leftarrow 0$
5. while $z = |O|$
   6. $z \leftarrow z + 1$
   7. $O(z) \leftarrow \operatorname{arginf}_{t \in (SSP, T]} \left( \int_{\max(t-h, 0)}^{t} a(t) \, dt > Ic \right)$
   8. $E(z) \leftarrow \operatorname{arginf}_{t \in (O(z), T]} \left( \frac{A(t) - A(O(z))}{t - O(z)} \leq \frac{Ic}{h} \right)$
   9. $SSP \leftarrow E(z)$
10. end while

**Output:** $O; E$

---

Note that $T \in T^U$ since no feasible solution will otherwise exist satisfying constraints (3.7).

Further, since unsaturated and oversaturated periods appear in an alternating pattern, for simplicity, we list the unsaturated periods as $T^U = \{ [O(0), O(1)), [E(1), O(2)), \ldots, [E(|O|), O(|O| + 1)] \}$ where $E(0) = 0$ and $O(|O| + 1) = T$. With this definition, we see that the starting moment of an oversaturated period, $O(z)$, is the first $t \in [0, T]$ during an unsaturated period when the cumulative arrival demand from $\max \{ t - h, 0 \}$ to $t$ is identical to $Ic$. Further, the ending moment of an oversaturated period, $E(z)$, is characterized as the first $t \in [0, T]$ when the queue that starts at $O(z)$ vanishes. With this definition, Propositions 3.1 and 3.2 imply that for any dispatch within an oversaturated period, a vehicle in formation $I$ is dispatched with headway $h$. Following the way Newell (1982) and Daganzo (1997) decomposed the oversaturated delay from unsaturated delay.
by drawing a straight line with a constant average service rate, we can revise arrival curve $A(t)$ in the following way to simplify the problem.

**Definition 3.2.** For a given arrival curve $A(t)$ at the origin station, the corresponding preferred virtual arrival demand curve is defined as

$$B(t) := \begin{cases} A(t) & t \in T^U \\ \frac{Ic}{h}(t - O(z)) + A(O(z)) & t \in [O(z), E(z)), \forall z \in [1, |O|] \end{cases}$$

and thus the preferred virtual arrival demand rate $B'(t) \leq \frac{Ic}{h}, \forall t \in [0, T]$ (the right derivative is used at points that are not differentiable). Obviously, $A(t) \geq B(t), \forall t \in [0, T]$. This is illustrated as the red dashed curve in Figure 3.3.

![Figure 3.3: Oversaturated and unsaturated periods.](image)

**Lemma 3.3.** Denote problem (3.2) ~ (3.8) as the original problem (OP), and the same problem where the original arrival curve $A(t)$ is replaced with preferred virtual arrival curve $B(t)$ as the revised unsaturated problem (RUP). Then the feasible regions of the OP and RUP are the same.
For any feasible solution \( s := \{t_k, i_k\}_{k \in \mathcal{K}}, \) OP and RUP have the same dispatch curve, and the objective values of OP and RUP, respectively denoted as \( \text{OP}(s) \) and \( \text{RUP}(s) \), are always separated by a constant difference:

\[
\text{OP}(s) - \text{RUP}(s) = W(A) := w \int_{0}^{T} (A(t) - B(t)) \, dt.
\]

Lemma 3.3 directly leads to the following result.

**Proposition 3.3.** Problems OP and RUP have the same optimal solution(s). For an optimal solution \( s \) to both problems, \( \text{OP}(s) - \text{RUP}(s) = W(A) \).

Proposition 3.3 implies that instead of solving OP, we can just solve RUP with preferred virtual arrival curve \( B(t) \). Note that RUP has constantly unsaturated arrival demands since 
\[
B'(t) \leq \frac{Ic}{h}, \forall t \in [0,T],
\]
which can dramatically reduce the complexity of the investigated problem.

Further, note that oversaturated arrival demand \( A(t) \) always leads to oversaturated waiting cost \( W(A) \). This also sheds light on the demand management side, e.g., managing the arrival demand to approach the unsaturated pattern \( B(t) \) (which explains why we call \( B(t) \) the preferred virtual arrival curve), which however is out of this paper’s scope. Actually, since \( B(t) \) is unsaturated throughout the time horizon, the theorem below shows that the corresponding optimal departure curve \( D(t) \) is not far separated from \( B(t) \).

**Proposition 3.4.** An optimal dispatch solution \( \{t_k, i_k\}_{k \in \mathcal{K}} \) to RUP satisfies: (i) if \( t_k - t_{k-1} > \frac{h}{c} \), then \( B(t_k) - D(t_k) = 0 \), \( \forall k \in \mathcal{K} \); and (ii) if \( t_k - t_{k-1} = \frac{h}{c} \), then \( B(t_k) - D(t_k) \in [0, c) \), \( \forall k \in \mathcal{K} \), where \( D(t) \) is the corresponding departure curve.

Interestingly, the property of having minimum queue at every dispatch (which approaches to 0 as \( c \to 0 \)) echoes the finding in Newell (2002) that the queue in the major approach needs to be always cleared before changing the signal. This alludes to certain structural connection between
these two seemingly unrelated problems that is worth future investigation. These elegant properties simplify the problem structure and largely reduce the domain where an optimal solution may appear for OP. They will be used to decompose the problem in the CA model and derive valid inequalities for the discrete model in the following sections.

3.4 Continuous Model

This section presents a CA model that aims to shed macroscopic analytical insights and tackle large-scale problems. A CA approach can approximate a local neighborhood in the searching space with a continuous function whose parameters are homogenous when properties of the searching space change slowly (Li et al., 2016). If the unit-time or unit-area cost around this neighborhood is mainly determined by the settings in its vicinity but less dependent on distant neighborhoods, then the optimal solution to this separable continuous function can well approximate the optimal solution of this neighborhood. In our problem, the total cost in a neighborhood is largely determined by the vehicle capacity and the dispatch headway around this neighborhood and \( A'(t) \) varies relatively smoothly in each neighborhood (e.g., with a time period comparable to the optimal dispatch headway). Thus, we can well approximate the original model with CA.

3.4.1 Model Formulation

Proposition 3.3 shows that instead of solving OP, we can just solve RUP with preferred virtual arrival curve \( B(t) \) where only unsaturated traffic is present. Hence, we can formulate the CA model based on the RUP.

Given an optimal solution \( s := \{t_k, i_k\}_{k \in \mathcal{K}} \) to RUP with preferred virtual cumulative arrival \( B(t) \) and departure \( D(t) \), we define the dispatch headway at time \( t \in [0, T] \) as \( \tilde{h}(t) := t_k - t_{k-1}, \text{s.t. } t \in [t_{k-1}, t_k), \exists k \in \mathcal{K} \) and the vehicle formation at time \( t \in [0, T] \) as \( i(t) := \)
$i_k, \text{s.t. } t \in [t_{k-1}, t_k), \exists k \in \mathcal{K}$. For the convenience of the notation, define $\hat{h}(t_k) := t_k - t_{k-1}$ and $i(t_K) := i_K$. Then, the optimal objective value of RUP can be formulated as

$$RUP(s) = \sum_{k \in \mathcal{K}} f_i + w \int_0^T (B(t) - D(t)) dt = \int_0^T \left( f_i(t) + \frac{w(B(t) - D(t))}{\hat{h}(t)} \right) dt,$$

s.t. (3.2) ~ (3.8) (where $A(t)$ is replaced with $B(t)$).

Figure 3.4: Revised preferred virtual arrival demand curve in MAV shuttle systems.

In order to approximate the OP, we revise the preferred virtual arrival demand curve as $\bar{B}(t) := B(t) - (B(t_{k-1}) - D(t_{k-1})), \forall t \in [t_{k-1}, t_k), k \in \mathcal{K}$ (see the dotted line in Figure 3.4). Then we reformulate (3.9) as

$$RUP(s) = \int_0^T \left( \frac{f_i(t)}{\hat{h}(t)} + w(\bar{B}(t) - D(t)) \right) dt + w \int_0^T (B(t) - \bar{B}(t)) dt \quad (3.10)$$

$$= \sum_{k \in \mathcal{K}} \int_{t_{k-1}}^{t_k} \left( \frac{f_i(t)}{\hat{h}(t)} + w(\bar{B}(t) - D(t)) \right) dt + w \int_0^T (B(t) - \bar{B}(t)) dt.$$
From Proposition 3.4, we know \( B(t) - \bar{B}(t) < c, \forall t \in [0, T] \) and thus \( \int_0^T (B(t) - \bar{B}(t)) dt < cT \).

Note that \( cT \) is a relatively small value compared with \( RUP(s) \). Further, numerical experiments with exact solution methods show that \( B(t) - \bar{B}(t) \) is 0 at most of the time. Therefore, we can drop \( w \int_0^T (B(t) - \bar{B}(t)) dt \) from (3.10) and obtain

\[
RUP(s) \approx \sum_{k \in \mathcal{K}} \left[ \int_{t_{k-1}}^{t_k} \frac{f_i(t)}{\hat{h}(t)} dt + w \int_{t_{k-1}}^{t_k} \left( \bar{B}(t) - D(t) \right) dt \right].
\]  

(3.11)

Note that in a specific neighborhood \([t_{k-1}, t_k], k \in \mathcal{K}, \bar{B}(t) \) is a continuous function. With

\[
B'(t) = \frac{\bar{B}(t) - \bar{B}(t_{k-1})}{t - t_{k-1}}, \forall t \in [t_{k-1}, t_k],
\]

we have

\[
\int_{t_{k-1}}^{t_k} \left( \bar{B}(t) - D(t) \right) dt = \int_{t_{k-1}}^{t_k} \left( \bar{B}(t) - \bar{B}(t_{k-1}) \right) dt = \int_{t_{k-1}}^{t_k} (t - t_{k-1}) B'(t) dt,
\]

which yields

\[
RUP(s) \approx \sum_{k \in \mathcal{K}} \left[ \int_{t_{k-1}}^{t_k} \frac{f_i(t)}{\hat{h}(t)} dt + wB'(t) \int_{t_{k-1}}^{t_k} (t - t_{k-1}) dt \right].
\]

(3.12)

\[
\approx \sum_{k \in \mathcal{K}} \left[ \int_{t_{k-1}}^{t_k} \frac{f_i(t)}{\hat{h}(t)} dt + wB'(t) \frac{(t_k - t_{k-1})^2}{2} \right]
\]

\[
\approx \sum_{k \in \mathcal{K}} \left[ \int_{t_{k-1}}^{t_k} \frac{f_i(t)}{\hat{h}(t)} dt + wB'(t) \int_{t_{k-1}}^{t_k} \frac{(t_k - t_{k-1})}{2} dt \right]
\]

\[
\approx \int_0^T \left( \frac{f_i(t)}{\hat{h}(t)} + \frac{wB'(t)\hat{h}(t)}{2} \right) dt.
\]

Note that in Eq. (3.12), \( B'(t) \) can be pulled out of the integral since \( B'(t) \) varies relatively slowly in each local time neighborhood so that we can regard it approximately as a constant. Also note that \( \hat{h}(t) \) is essentially a step function, to solve which is akin to determining \( t_k, \forall k \in \mathcal{K} \) themselves. To simplify this, we approximate \( \hat{h}(t) \) with a smooth function \( h(t) \) and then the RUP can be decomposed across the operational horizon as the following unit-time RUP
\[
\min_{h(t), i(t) \in \mathcal{J}} c_t(i(t), h(t)) := \frac{f_{i(t)}}{h(t)} + \frac{wB'(t)h(t)}{2}, \forall t \in [0, T].
\] (3.13)

Further, based on Proposition 3.4 and the fact that \(B(t) - \bar{B}(t)\) is 0 at most of the time, we can view that \(B(t) - \bar{B}(t) = 0\) approximately holds \(\forall t \in [0, T]\) in the optimal solution(s). That is to say, the queue length reduces to 0 for almost all dispatches across the operational horizon in the optimal solution(s). With this, \(h(t) < \frac{i(t)c}{B'(t)}\) must approximately hold in the optimal solution(s) (and otherwise \(B(t) - \bar{B}(t) = 0, \forall t \in [0, T]\) would be violated; i.e., the queue lengths after some dispatches are not 0), which together with (3.3) yields the following constraints

\[
h \leq h(t) < \frac{i(t)c}{B'(t)}, \forall i(t) \in \mathcal{J}, t \in [0, T].
\] (3.14)

\[
h < \frac{i(t)c}{B'(t)}, \forall i(t) \in \mathcal{J}, t \in [0, T].
\] (3.15)

Note that the above unit-time RUP at each time point \(t\), i.e., (3.13), (3.14), and (3.15), has only two decision variables and three constraints so it is much simpler than both the original formulation and the discrete model proposed in the next section. By solving the unit-time RUP \(\forall t \in [0, T]\), we can apply the optimal solution \(c^*(t) = \min_{h(t), i(t) \in \mathcal{J}} c_t(i(t), h(t))\) to RUP to obtain the objective value of OP as

\[
OP(s) \approx RUP(s) + W(A) \approx \int_0^T c^*(t) dt + W(A).
\] (3.16)

3.4.2 Analytical Solution

This section presents a closed-form analytical optimal solution to the proposed CA model. As mentioned in the previous section, the OP can be solved by solving the unit-time RUP across the operational horizon. Thus, the following analysis primarily focuses on the unit-time RUP.
This solution approach for Problem (3.13) at each time $t \in [0, T]$ has two steps. The first step is to solve $h(t)$ for a given feasible $i(t) \in \mathcal{J}(t)$ where \( \mathcal{J}(t) \equiv \{ i | i \in \mathcal{J}, ic > hB'(t) \} \) is the set of vehicle formations at time $t \in [0, T]$ feasible to Constraints (3.15). The second step solves the full problem by jointly optimizing $h(t)$ and $i(t)$.

Now we describe the first step for a given $i(t) \in \mathcal{J}(t)$. Without considering Constraints (3.14), note that since 
\[
\frac{\partial c_t(i(t), h(t))}{\partial h(t)} = -\frac{f_i(t)}{(h(t))^2} + \frac{wB'(t)}{2},
\]
function $c_t(i(t), h(t))$ for a fixed $i(t)$ strictly decreases with $h(t) \in \left(0, \frac{2f_i(t)}{\sqrt{wB'(t)}}\right)$ and strictly increases with $h(t) \in \left(\frac{2f_i(t)}{\sqrt{wB'(t)}}, \infty\right)$. With this property, we will solve the optimal $h(t)$ for given $i(t)$, denoted by $h_{i(t)}^*$, after adding back Constraints (3.14). Solving the left hand side and right hand side of Constraints (3.14) yield $a(t) \leq \frac{2f_i(t)}{w\h^2}$ and $a(t) \leq \frac{w[i(t)c]^2}{2f_i(t)}$, respectively. Depending on the relationship between $\frac{2f_i(t)}{w\h^2}$ and $\frac{w[i(t)c]^2}{2f_i(t)}$, the analytical solution to the unit-time RUP can be divided into the following two cases.

(i) When $0 < \h \leq \frac{2f_i(t)}{w[i(t)c]}$, i.e., $\frac{w[i(t)c]^2}{2f_i(t)} \leq \frac{2f_i(t)}{w\h^2}$.

\[
h_{i(t)}^*(t) = \begin{cases} \frac{2f_i(t)}{wB'(t)}, & \text{if } 0 \leq B'(t) \leq \frac{w[i(t)c]^2}{2f_i(t)}, \\ \frac{i(t)c}{B'(t)}, & \text{if } B'(t) > \frac{w[i(t)c]^2}{2f_i(t)} \end{cases} \forall i(t) \in \mathcal{J}(t), t \in [0, T]. \tag{3.17}
\]

(ii) When $\h > \frac{2f_i(t)}{w[i(t)c]}$, i.e., $\frac{w[i(t)c]^2}{2f_i(t)} > \frac{2f_i(t)}{w\h^2}$.

\[
h_{i(t)}^*(t) = \begin{cases} \frac{2f_i(t)}{wB'(t)}, & \text{if } 0 \leq B'(t) \leq \frac{2f_i(t)}{w\h^2}, \\ \h, & \text{if } B'(t) > \frac{2f_i(t)}{w\h^2} \end{cases} \forall i(t) \in \mathcal{J}(t), t \in [0, T]. \tag{3.18}
\]
With these results, we are ready to move to the next step on the joint optimization. Because the cardinality of $J(t)$ is limited in real-world cases, we can plug (3.17), (3.18) into the unit-time RUP and then solve it simply by exhaustive enumeration, i.e.,

$$i^*(t), h^*(t) = \arg\min_{i(t) \in J(t)} \left\{ c_t \left( i(t), h_{i(t)}^*(t) \right) \right\}, \forall t \in [0, T], \tag{3.19}$$

which directly leads to the optimal unit-time cost $c^*(t) = c_t(i^*(t), h^*(t)), \forall t \in [0, T]$. Finally, after solving the RUP’s for all neighborhoods $t \in [0, T]$, we can plug $c^*(t), \forall t \in [0, T]$ into (3.16) to obtain an estimated value of $OP(s)$.

### 3.4.3 Discretization Method

The analytical solution provides us with a closed-form expression of the optimal solution to the original problem. Yet the continuous solutions $h^*(t), i^*(t)$ cannot be directly applied to the schedule vehicles directly. Therefore, a discretization approach is needed to convert $h^*(t)$ into discrete time points for each dispatch $t_k^*, \forall k \in K$ and $i^*(t)$ into the discrete vehicle capacity $i_k^*, \forall k \in K$. Daganzo (2005) proposed a systematic method that can find an approximate step function to the continuous headway function generated by the CA model efficiently. The idea of Daganzo’s method is to find a fixed headway such that area under the $h^*(t)$ curve and that above the curve is the same. Although this method is already parsimonious, it still requires non-trivial computation burden involving numerical integral and trial-and-error (e.g. in evaluating the areas below and above and finding points on the vertical axis to draw horizontal segments). Just to further simply this process, we propose a method to discretize $h^*(t)$ by using the definition of headway, i.e. $h^*(t_{k-1}^*) = t_k^* - t_{k-1}^*$, iteratively for $k = K, K - 1, \ldots, 1$, as shown in Figure 3.5.

Specifically, given $t_k^*$, we need to determine $h^*(t_{k-1}^*)$ and $t_{k-1}^*$ such that $h^*(t_{k-1}^*) = t_k^* - t_{k-1}^*$. This can obviously be realized by drawing a $45^\circ$ line from $(t = t_k^*, h^*(t) = 0)$ backward in time
and finding its intersection with $h^*(t)$, which only requires at maximum one scans through the data points on the $h^*(t)$ curve without numerical integral or further iterations. Thus, we can discretize $h^*(t)$ by drawing $45^\circ$ lines recursively for $k = K,K-1,\cdots,1$. With the discretized $h^*(t)$, $i^*(t)$ can then be discretized with a weighted average method. The basic framework of this approach is shown as the pseudocode in Algorithm 3.2.

Figure 3.5: Discretization of $h^*(t)$.

Since the investigated problem requires that all passengers are transported at the end of the operational horizon, this algorithm starts from the last time point $T$, i.e. $t_K^* = T$. From this point ($t = T, h^*(t) = 0$), we draw a $45^\circ$ line backward in time and find its intersection with $h^*(t)$, the abscissa of which is the time point of the previous dispatch, $t_{K-1}^*$. Then we compute the weighted average of $i^*(t), \forall t \in [t_{K-1},t_K]$ and round it to an integer, which results in the vehicle formation for the $K^{th}$ dispatch, i.e. $i_K^*$. This step locates $t_{K-1}^*$ given $t_K^*$ and determines $i_K^*$. We repeat this procedure from $t_{K-1}^*$ to locate the previous dispatch times and formations until reaching a point whose distance to the origin is no greater than the minimum headway. Note that when implementing this algorithm, the index starts from $0$ to $K$, so at the end we need to reverse the indexes to obtain the correct solution. Once the discrete time point $t_k^*, \forall k \in \mathcal{K}$ and vehicle formation $i_k^*, \forall k \in \mathcal{K}$ are obtained, we can plug them into objective function (3.2) to obtain the
total system cost with respect to the discrete design. Note that it is possible that this algorithm ends up with a last interval that is slightly greater than the optimal one. One can adjust the solution by adding an extra interval and squeezing all the other ones a bit, or keep the same number of intervals and distributing the interval slack among all intervals. However, we do not implement this operation since so many intervals would be considered for discretization and this one interval does not show significant impacts, which can be verified from the numerical examples.

**Algorithm 3.2. Discretization**

**Input:** $h^*(t), i^*(t), \forall t \in [0, T], h$

1. $K \leftarrow 0$
2. $t^*_K \leftarrow T$
3. while $t^*_K \geq h$
4. \hspace{0.5cm} $K \leftarrow K + 1$
5. \hspace{0.5cm} $t^*_K \leftarrow \text{argsup}_{t \in [0, t_{K-1}]} (h^*(t) = t^*_K - t)$
6. \hspace{0.5cm} $i^*_{K-1} \leftarrow \text{Round} \left( \frac{\int_{i^*_K}^{i^*_{K-1}} i^*(t) dt}{t^*_K - t^*_{K-1}} \right)$
7. end while
8. $i^*_K \leftarrow \int_0^{i^*_K} i^*(t) dt / t^*_K$
9. Reverse $(i^*_k, t^*_k) = (i^*_K-k, t^*_K-k), \forall k \in \mathcal{K} := \{0, 1, \ldots, K\}$

**Output:** $t^*_k, i^*_k, \forall k \in \mathcal{K}$

### 3.5 Discrete Model

This section presents a discrete model for the investigated problem to solve the exact optimal solution from a microscopic perspective. The discrete model is essentially a mixed integer linear programming model with a discretized time representation, which can be solved by commercial solvers (e.g., Gurobi, CPLEX) if the instance size is not too large. However, the computation time of commercial solvers still increases rapidly with the instance size. To tackle
this problem more efficiently, we design a customized dynamic programming (DP) algorithm where the state space is largely reduced with a set of valid inequalities derived from some theoretical properties of the investigated problem.

Figure 3.6: Cumulative passenger counts in MAV shuttle systems in a discrete setting.

3.5.1 Model Formulation

In order to represent the system dynamics in a discrete setting, we discretize the investigated operational horizon into $J$ intervals with a set of discrete time points $J = [0, 1, 2, \ldots, J]$, indexed by $j \in J$. The discretization interval is $\delta$ such that $\delta J = T$. Please note that to minimize the numerical error, $\delta$ can be set as a value such that $h/\delta$ is an integer, where $h$ is the minimum dispatch headway. With this, the number of arrival demand during interval $[j-1, j]$ can be defined as $a_j := \int_{t=(j-1)\delta}^{j\delta} a(t) dt, \forall j \in J \setminus \{0\}$. The cumulative arrival demand up to time point $j$ can then be defined as $A_j := \sum_{j'=0}^{j} a_j, \forall j \in J$, as shown in Figure 3.6. Further, in this discrete setting, the decision problem turns out to be determining whether we need to dispatch a vehicle in formation $i \in I$ at any discrete time point $j \in J$. To formulate this decision, we introduce a binary variable $x_{ij}$ that equals 1 if a vehicle in formation $i \in I$ is dispatched at time $j \in J$ and otherwise
These decisions result in a cumulative departure curve from the origin, denoted by $D_j, \forall j \in \mathcal{J}$. Further, in order to formulate the system dynamics, we introduce a continuous decision variable $q_j := A_j - D_j, \forall j \in \mathcal{J}$, which actually represents the number of passengers that are waiting for boarding right after time $j, \forall j \in \mathcal{J}$.

We first reformulate the physical constraints of the shuttle system as follows.

(i) Minimum headway requirement. Since $h$ is a continuous value in the real world, we define the minimum headway interval as $\bar{\Delta} = \lceil \frac{h}{\delta} \rceil$, where $\lceil \cdot \rceil$ is the ceiling operator. Then at each time $j \in [\bar{\Delta} - 1, \bar{\Delta}]$, the minimum headway requirement actually requires that no more than one vehicle can be dispatched within the minimum headway interval, i.e.,

$$\sum_{i \in \mathcal{I}} \sum_{j'=j-\bar{\Delta}+1}^j x_{ij'} \leq 1, \forall j - 1 \leq j \leq \bar{\Delta}, \quad (3.20)$$

(ii) Queuing conservation process. These constraints reformulates constraints (3.5) ~ (3.7). Constraints (3.21) indicate that the number of passengers that are waiting right after time $j$ should be no less than the number of passengers that are waiting for boarding right after $j - 1$ plus the arrival passengers at $j$ and then subtract the vehicle capacity. Constraints (3.22) are imposed due to the fact that $A_j \geq D_j, \forall j \in \mathcal{J}$. Constraint (3.23) initializes the queuing process while Constraint (3.24) ensures that all passengers are transported at the end of the operational horizon.

$$q_j \geq q_{j-1} + a_j - c \sum_{i \in \mathcal{I}} i x_{ij}, \forall j \in \mathcal{J}\setminus\{0\}, \quad (3.21)$$

$$q_j \geq 0, \forall j \in \mathcal{J}, \quad (3.22)$$

$$q_0 = 0, \quad (3.23)$$

$$q_J = 0, \quad (3.24)$$

(iii) Feasible regions. These constraints define the feasible regions for $x_{ij}, \forall i \in \mathcal{I}, j \in \mathcal{J}$. 

$$q_j \geq 0, \forall j \in \mathcal{J}, \quad (3.25)$$
Next, we reformulate the objective function. Since here we are interested in formulating the problem into a linear model, we approximate the total waiting time in an interval \([j-1, j]\), \(\forall j \in J\setminus\{0\}\) with the area of a triangle (if one of \(q_{j-1}\) and \(q_j\) is 0) or a trapezium (if neither \(q_j\) nor \(q_{j-1}\) is 0). It is obvious that the approximation error goes to zero as \(\delta\) approaches zero. With this, the objective function can be formulated as

\[
\min_{x_{ij}, q_j, \forall i \in I, j \in J} \left( \sum_{j \in J \setminus \{0\}} \left( \sum_{i \in I} x_{ij} f_i + \left(q_{j-1} + \frac{a_j}{2}\right) \delta w \right) \right),
\]

where the first and second term represents the total energy cost and total passenger waiting cost, respectively.

3.5.2 Valid Inequalities

The theoretical properties proved in Section 3.2 can be equivalently translated into the discrete settings as the following propositions.

**Proposition 3.5.** In an optimal dispatch solution \(\{x_{ij}, q_j\}_{\forall i \in I, j \in J}\) to problem (3.20) \(\sim\) (3.26) with arrival curve \(A_j\) and departure curve \(D_j\), if \(\sum_{i \in I} x_{ij} = 1, \sum_{i \in I} x_{ij' = 1, \sum_{j' \in (j', j)} \sum_{i \in I} x_{ij''} = 0}\) and \(A_j - D_j \geq 1c\), then \(x_{ij} = 1, \forall j' \leq j - j \in J\).

**Proposition 3.6.** In an optimal dispatch solution \(\{x_{ij}, q_j\}_{\forall i \in I, j \in J}\) to problem (3.20) \(\sim\) (3.26) with arrival curve \(A_j\) and departure curve \(D_j\), if \(\sum_{i \in I} x_{ij} = 1, \sum_{i \in I} x_{ij' = 1, \sum_{j' \in (j', j)} \sum_{i \in I} x_{ij''} = 0}\) and \(A_j - D_j \geq 1c\), then \(j - j' = j, \forall j' \leq j - j \in J\).

**Definition 3.3.** In the discrete model for given arrival curve \(A_j\), the oversaturated period set is defined as \(J^S := \{j | j \in J, 0(z) \leq j < \mathcal{E}(z), \forall z \in [1, |O|]\}\) and the unsaturated period set is defined as \(J^U := J \setminus J^S\), where \(O\) and \(\mathcal{E}\) are obtained with Algorithm 3.3.
**Definition 3.4.** For a given arrival curve $A_j$, the desired arrival demand curve is defined as

$$B_j := \begin{cases} A_j & j \in \mathcal{J}^u \\ \frac{Ic}{j} (j - \mathcal{O}(z)) + A_{\mathcal{O}(z)} & j \in [\mathcal{O}(z), \mathcal{E}(z)), \forall z \in [1, |\mathcal{O}|] \end{cases}$$

and thus the desired arrival demand rate is less than $Ic/j$ across the operational horizon. Obviously, $A_j \geq B_j, \forall j \in \mathcal{J}$.

**Algorithm 3.3.** Oversaturation Detection with Discrete Inputs

**Input:** $\mathcal{J}; l; c; j; a_j, A_j, \forall j \in \mathcal{J}$

1. $\mathcal{O} \leftarrow \emptyset$
2. $\mathcal{E} \leftarrow \emptyset$
3. $SSP \leftarrow 0$
4. $z \leftarrow 0$
5. while $z = |\mathcal{O}|$
6. $z \leftarrow z + 1$
7. $\mathcal{O}(z) \leftarrow \arg\min_{j \in ([SSP], l)} \left( \sum_{j'=\max(j-l+1,0)}^{j} a_{j'} > Ic \right)$
8. $\mathcal{E}(z) \leftarrow \arg\min_{j \in (\mathcal{O}(z), l]} \left( A_j - A_{\mathcal{O}(z)} \leq \frac{Ic}{j} \right)$
9. $SSP \leftarrow \mathcal{E}(z)$
10. end while

**Output:** $\mathcal{O}; \mathcal{E}$

**Proposition 3.7.** Denote problem (3.20) ~ (3.26) as the discretized original problem (DOP), and the same problem where original arrival demand $A_j$ is replaced with desired arrival demand $B_j$ as the discretized revised unsaturated problem (DRUP). Problems DOP and DRUP have the same optimal solution(s). For an optimal solution $s$ to both problems, the objective values of DOP and DRUP, respectively denoted as $DOP(s)$ and $DRUP(s)$, are always separated by a constant difference:

$$DOP(s) - DRUP(s) = W(A) := \delta w \sum_{j \in \mathcal{J} \setminus \{0\}} (A_j - B_j).$$
Proposition 3.8. An optimal dispatch solution \( \{ x_{ij}, q_j \}_{\forall i \in J, j \in J} \) to DRUP with desired arrival curve \( B_j \) and departure curve \( D_j \) satisfies: (i) if \( \sum_{i \in J} x_{ij} = 1, \sum_{i \in J} x_{ij'} = 1, \sum_{j'' \in (j', J)} \sum_{i \in J} x_{ij''} = 0, \) and \( j \sum_{i \in J} x_{ij} - j' \sum_{i \in J} x_{ij'} > j \), then \( B_j - D_j = 0, \forall j > j' \in J \); (ii) if \( \sum_{i \in J} x_{ij} = 1, \sum_{i \in J} x_{ij'} = 1, \sum_{j'' \in (j', J)} \sum_{i \in J} x_{ij''} = 0, \) and \( j \sum_{i \in J} x_{ij} - j' \sum_{i \in J} x_{ij'} = j \), then \( B_j - D_j \in [0, c], \forall j > j' \in J \).

With Proposition 3.8, we directly obtain the following corollary to describe the queue-capacity relationship to DRUP.

**Corollary 3.1.** An optimal solution \( \{ x_{ij}, q_j \}_{\forall i \in J, j \in J} \) to DRUP satisfies

\[
q_j < \begin{cases} 
0 & \text{if } \sum_{i \in J} x_{ij} = 1, \sum_{i \in J} x_{ij'} = 1, \sum_{j'' \in (j', J)} \sum_{i \in J} x_{ij''} = 0, j \sum_{i \in J} x_{ij} - j' \sum_{i \in J} x_{ij'} > j; \\
0 & \text{if } \sum_{i \in J} x_{ij} = 0, \\
(I + 1)c & \text{if } \sum_{i \in J} x_{ij} = 0, \\
c & \text{if } \sum_{i \in J} x_{ij} = 1, \sum_{i \in J} x_{ij'} = 1, \sum_{j'' \in (j', J)} \sum_{i \in J} x_{ij''} = 0, j \sum_{i \in J} x_{ij} - j' \sum_{i \in J} x_{ij'} = j; \forall j > j' \\
(I + 1)c & \text{if } \sum_{i \in J} x_{ij} = 0,
\end{cases}
\]

\( \in J. \)

Further, based on Definition 3.4, we obtain

\[
\Delta q_j = A_j - B_j = \left\{ \sum_{j' = \mathcal{O}(z)}^j a_{j'} - \frac{j - \mathcal{O}(z)}{j} Ic, \forall j \in (\mathcal{O}(z), \mathcal{E}(z)) \subset J^S, \forall z \in [1, |\mathcal{O}|]. \right\}
\]

Thus, the queue-capacity relationship to DOP can be formulated as the following corollary.

**Corollary 3.2.** An optimal solution \( \{ x_{ij}, q_j \}_{\forall i \in J, j \in J} \) to DOP satisfies

\[
q_j < \begin{cases} 
\Delta q_j & \text{if } \sum_{i \in J} x_{ij} = 1, \sum_{i \in J} x_{ij'} = 1, \sum_{j'' \in (j', J)} \sum_{i \in J} x_{ij''} = 0, j \sum_{i \in J} x_{ij} - j' \sum_{i \in J} x_{ij'} > j; \\
\Delta q_j & \text{if } \sum_{i \in J} x_{ij} = 1, \sum_{i \in J} x_{ij'} = 1, \sum_{j'' \in (j', J)} \sum_{i \in J} x_{ij''} = 0, j \sum_{i \in J} x_{ij} - j' \sum_{i \in J} x_{ij'} = j; \forall j \\
(I + 1)c + \Delta q_j & \text{if } \sum_{i \in J} x_{ij} = 0, \\
(I + 1)c + \Delta q_j & \text{if } \sum_{i \in J} x_{ij} = 0, \\
(I + 1)c + \Delta q_j & \text{if } \sum_{i \in J} x_{ij} = 0,
\end{cases}
\]

\( > j' \in J. \)
Corollary 3.2 provides an upper bound to $q_j$, $\forall j \in J$ in an optimal solution to DOP. Further, the following lemma provides another upper bound to $q_j$ based on the feasibility condition (3.24).

**Lemma 3.3.** A feasible solution to DOP satisfies $q_j \leq \left(1 + \left\lfloor \frac{j-1}{l} \right\rfloor \right)lc$, $\forall j \in J$.

Integrating Corollary 2 and Lemma 3, we obtain the following valid inequalities to DOP.

$$q_j \leq \tilde{q}_j \left( \sum_{i \in I} x_{ij'}, j' \in [j', j] \right)$$

$$\min \left\{ \Delta q_j, \left(1 + \left\lfloor \frac{j-1}{l} \right\rfloor \right)lc \right\} \text{ if } \sum_{i \in I} x_{ij} = 1, \sum_{i \in I} x_{ij'} = 1, \sum_{j'' \in (j', j)} \sum_{i \in I} x_{ij''}$$

$$\min \left\{ c + \Delta q_j, \left(1 + \left\lfloor \frac{j-1}{l} \right\rfloor \right)lc \right\} \text{ if } \sum_{i \in I} x_{ij} = 1, \sum_{i \in I} x_{ij'} = 1, \sum_{j'' \in (j', j)} \sum_{i \in I} x_{ij''}$$

$$\min \left\{ (l+1)c + \Delta q_j, \left(1 + \left\lfloor \frac{j-1}{l} \right\rfloor \right)lc \right\},$$

$\forall j > j' \in J$.

### 3.5.3 Dynamic Programming Algorithm

This section proposes a customized dynamic programming algorithm to expedite the solution efficiency. Similar to other problems that can be solved by the DP algorithm, the challenge of “curse of dimensionality” also exists in formulating the investigated problem into a dynamic programming model, because the queue length, as a state variable, may have a large set of possible values at each stage in the original problem. To address this challenge, we derive valid inequalities from the theoretical properties to cut down the state space in the customized dynamic programming algorithm.

We consider each discrete time point $j \in J$ as a stage of the dynamic system. To construct the state space, we define auxiliary state variable $h_j := \begin{cases} \min \left\{ h_{j-1} + 1, j \right\}, & \text{if } \sum_{i \in I} x_{ij} = 0, \forall j \in \mathbb{N} \\ 1, & \text{if } \sum_{i \in I} x_{ij} = 1 \end{cases}$
\( J \setminus \{0\}, \) and \( h_0 := j. \) State variable \( h_j \) can be interpreted as the minimum of the number of time intervals from the previous dispatch to time \( j + 1 \) and minimum dispatch interval \( j. \) Note that if \( h_j < j, \) Constraints (3.20) require that no vehicles can be dispatched at \( j. \) Here we set an upper bound to \( h_j \) to restrict the state space of this variable. However, the number of time intervals from the previous dispatch to time \( j + 1, \) denoted as \( \hat{h}_j := \begin{cases} h_j - 1, & \text{if } \sum_{i \in J} x_{ij} = 0, \\ 1, & \text{if } \sum_{i \in J} x_{ij} = 1, \forall j \in J \setminus \{0\}, \end{cases} \) and \( \hat{h}_0 := -J. \) is needed to apply valid inequalities (3.27) so we also record this value during computation. With this variable, valid inequalities (3.27) can be rewritten as
\[
q_j \leq \bar{q}_j \left( \sum_{i \in J} x_{ij}, \hat{h}_j \right)
\]
\[
:= \begin{cases} \min \left\{ \Delta q_j, \left( 1 + \left\lfloor \frac{j - j'}{l} \right\rfloor \right) Ic \right\}, & \text{if } \sum_{i \in J} x_{ij} = 1, \hat{h}_j > j; \\ \min \left\{ c + \Delta q_j, \left( 1 + \left\lfloor \frac{j - j'}{l} \right\rfloor \right) Ic \right\}, & \text{if } \sum_{i \in J} x_{ij} = 1, \hat{h}_j = j; \forall j > j' \\ \min \left\{ (l + 1)c + \Delta q_j, \left( 1 + \left\lfloor \frac{j - j'}{l} \right\rfloor \right) Ic \right\}, & \text{if } \sum_{i \in J} x_{ij} = 0, \end{cases}
\]
\( \in J. \)

With these definitions, the algorithm starts at stage \( j = 0 \) with state space \( s_0 := \{(j,0)\} \) and value function \( V_0(j,0) := 0. \) It then solves the problem by recursively calling the following steps across all the following stages \( j = 1, 2, \cdots, J. \) At each stage \( j, \) the state space, denoted by \( s_j, \) is first initialized as \( \emptyset \) and then will be expanded based on the state space in the previous stage \( j - 1. \) Based on valid inequalities (3.28), Constraints (3.20) and the fact that the queue will be cleared after time \( J, \) the feasible set of vehicle formations for each state \( (h_{j-1}, q_{j-1}) \in s_{j-1} \) is
\[ j_{j-1}(h_{j-1}, q_{j-1}) \]
\[
:= \left\{ \left( \forall i \in J \cup \{0\}, q_{j-1} + a_j - ic \in [\Delta q_j, \bar{q}_j(E_{i>0}, \hat{h}_{j-1})], \text{if } \hat{h}_{j-1} \geq j \right) \right\}, \forall (h_{j-1}, q_{j-1}) \tag{3.29}
\]
\[
\in s_{j-1}, \text{if } j < J;
\]
\[
\jmath_{j-1}(h_{j-1}, q_{j-1}) := \left\{ \left( \left( \frac{q_{j-1} + a_j}{c} \leq i \in J, \text{if } h_{j-1} = \hat{j} \right) \right) \right\}, \forall (h_{j-1}, q_{j-1}) \in s_{j-1}, \text{if } j
\]
\[
= J,
\]
where \( E_{i>0} = 1 \) if condition \( i > 0 \) is true and \( E_{i>0} = 0 \) otherwise. Next, for \( \forall (h_{j-1}, q_{j-1}) \in s_{j-1}, \) based on the state transition functions, i.e., Constraints (3.21) through (3.24), we enumerate \( \forall i \in j_{j-1}(h_{j-1}, q_{j-1}) \) and add \( \left( h_j = \{ \min\{h_{j-1} + 1, \hat{j}\}, \text{if } i = 0; \quad q_j := \max\{0, q_{j-1} + a_j - ic\} \right) \) to \( s_j \). Also, \( \hat{h}_j := \{ \hat{h}_{j-1} + 1, \text{if } i = 0; \quad 1, \text{if } i = 1 \} \) is stored. Note that without the valid inequalities (3.28), \( q_j \) would increase unboundedly because cumulative demand curve \( A_j \) could increase unboundedly as \( j \) increases. We will show in the numerical experiments that the computation speed of the customized DP algorithm is significantly improved thanks to these valid inequalities. Finally, for modeling convenience, denote the set of \( (\{x_{ij}\}, h_{j-1}, q_{j-1}) \) that leads to state \( (h_j, q_j) \) at the next stage as \( s_{j-1}(h_j, q_j), \forall (h_j, q_j) \in s_j \). Then the following value function is optimized as
\[
V_j(h_j, q_j): = \min_{(\{x_{ij}\}, h_{j-1}, q_{j-1}) \in s_{j-1}(h_j, q_j)} \left\{ \sum_{i \in J_{j-1}(h_{j-1}, q_{j-1})} x_{ij} f_i \right. \\
+ \left( q_{j-1} + \frac{a_j}{2} \right) \delta w + V_{j-1}(h_{j-1}, q_{j-1}) \right\}, \forall (h_j, q_j) \in s_j. \tag{3.30}
\]
Note that if \( s_{j-1}(h_j, q_j) = \emptyset \), then \( V_j(h_j, q_j) = \infty \). This optimization can be repeated across all stage up to stage \( J \). This process can be formally stated as the pseudocode in Algorithm 3.4.

**Algorithm 3.4. Dynamic Programming Algorithm**

**Input:** \( \mathcal{J}; J; I; w; a_j, A_j, \forall j \in \mathcal{J}; C_i, f_i, \forall i \in I \)

1. \( j \leftarrow 0 \)
2. \( s_0 \leftarrow \{(j, 0)\} \)
3. \( V_0(j, 0) \leftarrow 0 \)
4. **for** \( j = 1 \) to \( J \) **do**
   5. \( s_j \leftarrow \emptyset \)
   6. Determine \( j_{j-1}(h_{j-1}, q_{j-1}), \forall (h_{j-1}, q_{j-1}) \in s_{j-1} \) using (3.29)
   7. Add \( h_j \leftarrow \begin{cases} \min\{h_{j-1} + 1, j\}, & \text{if } i = 0 \\ 1, & \text{if } i > 0 \end{cases}, q_j \leftarrow \max\{0, q_{j-1} + a_j - ic\}, \forall i \in j_{j-1}(h_{j-1}, q_{j-1}), (h_{j-1}, q_{j-1}) \in s_{j-1} \) to \( s_j \) and store the value of \( \hat{h}_j \)
   8. Solve the Bellman’s optimality equation (3.30) for \( \forall (h_j, q_j) \in s_j \). For each \( (h_j, q_j) \), record the minimizers to equation (3.30) as \( x_{ij}(h_j, q_j), \hat{h}_{j-1} (h_j, q_j), \hat{q}_{j-1} (h_j, q_j) \).
5. **end for**
6. For stage \( J \), set \( V_J^* = \min_{(h_J, 0) \in s_J} \{V_J(h_J, 0)\} \). Let \( h_J^* \) denote the minimizer, and set \( q_J^* = 0 \). Set \( x_{iJ}^* = \)
   7. \( x_{iJ}^* \leftarrow [x_{ij}(h_j^*, q_j^*)] \)
8. **for** \( j = J - 1 \) to \( 1 \) **do**
9. \( [x_{ij}^*] \leftarrow [x_{ij}(h_{j+1}^*, q_{j+1}^*)], h_{j-1}^* \leftarrow \hat{h}_{j-1} (h_j^*, q_j^*), q_{j-1}^* = \hat{q}_{j-1}(h_j^*, q_j^*) \).
10. **end for**

**Output:** Optimal solution \( x_{ij}^*, q_j^*, \forall i \in \mathcal{I}, j \in \mathcal{J}, V_j^* \)
3.6 Numerical Experiments

This section illustrates the performance of the proposed modeming method and investigates managerial insights into the MAV shuttle operation with two sets of numerical examples. In the first set of experiments, we set up a shuttle system with smart-card data from Batong line in the Beijing Subway System, China to investigate the computation performance of the proposed solution approach and how the system performance of the MAV shuttle varies with different input parameters. The second set of experiments further investigates the effects of different input parameters on the system performance in a hypothetical shuttle system with MAV information from NEXT (2018) and future travel demand data in Tampa Bay Area, USA (obtained from Gannett Fleming, Inc.). The numerical experiments are conducted on a DELL Studio PC with 3.60 GHz of Intel Core i7-7700 CPU and 16 GB RAM in a Windows environment. The algorithms using Gurobi and CA algorithms were implemented with MATLAB 2017b, and the DP algorithm was implemented with Visual C++ 2015.

3.6.1 Case Study 1: Batong Line in Beijing Subway System

We first explain how we set up the numerical experiments by transforming a corridor system in Beijing Subway to a shuttle system that our model can address. As shown in Figure 3.7, Batong line is a bi-directional line with a total length of 18.964 km and 13 stations that are numbered sequentially from 1 to 13. Since both directions show strong temporal demand fluctuations and an evident oversaturated period in each direction over a day, without loss of generality, we selected the direction with a morning oversaturated period (i.e., the direction from Stations 13 to 1) for our experiments. The maximum passenger flow section in this direction is 4-3, so we treat stations 4-13 as a virtual origin while stations 1-3 and all stations on other lines in the metro network a virtual destination. Note that all stations on other lines in the network are
considered since the destinations of a large portion of demand emanating from stations 4-13 are out of Batong line. In this way, this metro line is converted into a one-to-one shuttle system that essentially represents the bottleneck of the metro line. Then, we count the number of passengers that will cross the maximum passenger flow section (i.e. passengers whose origins are one of stations 4-13 and destinations are one of the rest of stations in the network) per 0.1 minute and obtain the time-dependent cumulative arrival demand curve and the arrival demand rate curve over the entire operational horizon as Figure 3.8. Note that in this case study we are solving an optimal schedule to the converted shuttle system (as if trips not passing through the bottleneck do not exist) rather than to the original corridor system. Such an analysis (i.e. operational design for multiple stop systems based on simply demands passing through the maximum load points) has been applied in Salzborn (1972), Daganzo (1997) and Sheffi and Sugiyama (1982). Further, although they are not optimal, solutions from the converted shuttle system suffice to provide a reasonable or even near-optimum solution to the operation of the original corridor system. Therefore, the results in this case study can still provide managerial insights to system operators.

The MAVs serving on this line are composed of 6 carriages, each with a capacity of 226 pax (i.e., passengers). Hence, we set \( J = [1,2,3,4,5,6] \) and \( c = 226 \) pax/carriage. The minimum dispatch headway is set as \( h = 3 \) minutes. We adopt the monthly average salary per capita to compute the waiting cost per passenger per unit time. Beijing Municipal Human Resources and Social Security Bureau reports that the average monthly salary per capita is $1096.19 in 2017 (http://www.bjrbj.gov.cn/). Provided that one works 22 days a month and 8 hours per day, the hourly average salary per capital is around $6.35 and thus \( w = 0.11 \) $/min. The dispatch energy cost function we adopt in this case is \( f_i = C^F + C^V(i)^\alpha, \forall i \in J \) and we obtain \( \alpha = 0.5, C^F=2.049 \) $, and \( C^V=5.56 \) $ through a calibration process that is explained in Chen et al., (2020). Since there
is no empirical data regarding the energy cost of Batong Line in the Beijing Subway System, the calibration is conducted using data from a similar system, the Shenzhen Subway System, in China.

![Diagram of Batong Line and the converted shuttle system](image)

Figure 3.7: Batong Line and the converted shuttle system.

![Graphs showing passenger demand](image)

Figure 3.8: Passenger demand in the converted shuttle system in Batong Line. (a) Cumulative arrival demand, (b) arrival demand rate.

### 3.6.1.1 Computation Performance

This subsection analyzes the computation performance of the solution approaches. To investigate how the computation performance varies with the problem scale, we designed 21
instances with different operational horizons (i.e., 7:00 a.m. to 10:00 a.m., 5:00 a.m. to 14:00 p.m. and 5:00 a.m. to 23:00 p.m.) and discretization time intervals $\delta$ (i.e., 0.1, 0.2, 0.3, 0.5, 1.0, 1.5, 3.0 minutes). In the experiments, our solution approaches are benchmarked against a state-of-the-art commercial solver, Gurobi. To explore whether the valid inequalities can improve the computation speed, we solve the instances with and without adding these valid inequalities in the DP algorithm. Thus, each instance is solved with 5 approaches, i.e., Gurobi without state reduction (G-NSR), Gurobi with state reduction (G-SR), DP without state reduction (DP-NSR), DP with state reduction (DP-SR) and the CA model. Further, we compute the objective value of the original problem by both integral and discretization after solving the problem with CA. To differentiate these two methods, we denote them as CA-I and CA-D in the following analysis. Further, we stop the program if it cannot solve the problem to optimality after 1 hour. The statistics for the solution quality and computation time are summarized in Table 3.1 and Table 3.2, respectively.

As can be seen from Table 3.1, G-NSR can only solve 12 instances to optimality within an hour. For the rest of them, G-NSR provides near-optimal solutions to the problem (i.e., the italic numbers in Table 3.1, with the optimality gap presented in the round brackets) even if it cannot solve them to optimality. However, after adding valid inequalities (3.27), the number of constraints in the mathematical model became so huge that it usually took more than an hour to construct the model. As a result, Gurobi did not start solving the model at all. Therefore, the results of G-SR are omitted here. For the customized DP algorithm, both DP-NSR and DP-SR result in the same solutions as G-NSR, which verify the correctness of these algorithms. Further, both CA approaches produce near-optimal approximations across all instances, with the largest absolute gap of 0.63% from CA-I and 1.11% from CA-D compared with the exact solutions. Further, CA-D can obtain solutions with smaller gaps than both G-NSR and G-SR in most of the cases when Gurobi cannot
solve the instances to optimality, which means that CA-D can offer us better near-optimal solutions in this situation. Note that the gap for the CA approach shows us how far the CA solutions are from the true optimum. However, the optimality gap from Gurobi is the relative difference between the best found upper and lower bound. Therefore, no comparison should be made between these two types of gaps. For instance, when the length of horizon is 3 hours with a discretization interval of 0.1 minute, we cannot say that CA-D results in better solutions than Gurobi simply because CA-D has a gap of 0 while G-NSR and G-SR yield optimality gaps of 0.18 and 0.08, respectively.

Table 3.1: Statistics of solution quality metrics. The numbers in the brackets show the corresponding optimality gaps (between the best found upper bound and lower bound) from Gurobi. “/” indicates unsolved runs within an hour.

<table>
<thead>
<tr>
<th>Length of horizon (hours)</th>
<th>δ (min)</th>
<th>Computational time (sec)</th>
<th>Gap (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Gurobi</td>
<td>DP-NSR</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>95708</td>
<td>95708</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>95708</td>
<td>95708</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>95707</td>
<td>95707</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>95707</td>
<td>95707</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>95706</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>95706</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>95706 (0.18)</td>
<td>/</td>
</tr>
<tr>
<td>9</td>
<td>3.0</td>
<td>101910</td>
<td>101910</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>101910</td>
<td>101910</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>101902</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>101950 (0.36)</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>102150 (1.26)</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>102490 (2.47)</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>102660 (5.03)</td>
<td>/</td>
</tr>
<tr>
<td>18</td>
<td>3.0</td>
<td>107108</td>
<td>107108</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>107092</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>107062</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>107110 (0.74)</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>107710 (2.97)</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>108430 (5.40)</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>108910 (14.62)</td>
<td>/</td>
</tr>
</tbody>
</table>

Table 3.2 shows that the computation time of G-NSR increases dramatically as the instance size increases and it cannot solve the model in 9 instances; i.e., when the number of discrete time
points reaches 2161, G-NSR cannot solve the problem to optimality. The computation time of DP-NSR shows a similar trend as G-NSR but the growth rate is even higher, which indicates that the standard DP algorithm is even worse than G-NSR in terms of the computation efficiency because of the state space of queue length increases exponentially over time. However, the DP algorithm combined with the valid inequalities (DP-SR) can solve all instances and the computation time of DP-SR is much shorter than Gurobi. This observation indicates that the combination of the DP algorithm and valid inequalities allows us to obtain the exact optimal design much more efficiently than commercial solvers. Nevertheless, it still takes DP-SR minutes to solve large scale instances (e.g. when the length of the operational horizon is 18 hours and the discretization interval is 0.1 minutes) and its computational time increases quickly as the problem size increases. In contrast, the CA method can solve all instances in almost no time, which is much shorter than both G-NSR and DP-SR. Further, CA’s computational time remains relatively stable despite the increase in the discretization time interval. Actually, since each neighborhood is just solved analytically in a constant time, the computational time increases linearly with the number of total neighborhoods, which depends on the discrete size of a neighborhood and the total time horizon. The linear time complexity of CA is much more computationally-friendly than discrete models whose solution time often exponentially increases with the instance size. Hence, CA completely dominates the rest of approaches in terms of the computational efficiency despite minor approximation errors. Thus, CA shows a promising potential to tackle large-scale instances in real-world contexts.

To sum up, the DP algorithm with state reduction produces exact solutions to the joint design problem given enough computation resources but CA can only produce near-optimum solution(s) and its solution quality is yet to be verified by an exact solution approach. Thus, the discrete model is better in solution quality. On the other hand, the CA approach shows a dominant
advantage on solution speed since it can solve the problem in almost no time, even for extremely large-scale instances. Thus, the two approaches are both valuable and complementary to each other. Which approach shall be selected in practice depends on the problem size and whether transit operators are in pursuit of solution quality or speed.

Table 3.2: Statistics of computational time metrics

<table>
<thead>
<tr>
<th>Length of horizon (hours)</th>
<th>δ (min)</th>
<th>Computational time (sec)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gurobi</td>
<td>DP-NSR</td>
<td>DP-SR</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>0.03</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.10</td>
<td>7.81</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.20</td>
<td>35.93</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1.13</td>
<td>771.20</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>8.91</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>146.70</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>&gt;3600</td>
<td>&gt;3600</td>
</tr>
<tr>
<td>9</td>
<td>3.0</td>
<td>0.13</td>
<td>219.40</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.65</td>
<td>3239.11</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>4.14</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>&gt;3600</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>&gt;3600</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>&gt;3600</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>&gt;3600</td>
<td>&gt;3600</td>
</tr>
<tr>
<td>18</td>
<td>3.0</td>
<td>0.14</td>
<td>1293.41</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>3.64</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>22.35</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>&gt;3600</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>&gt;3600</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>&gt;3600</td>
<td>&gt;3600</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>&gt;3600</td>
<td>&gt;3600</td>
</tr>
</tbody>
</table>

3.6.1.2 Effectiveness of MAV Shuttle Systems

In this subsection, we assess the effectiveness of dynamic vehicle capacity adjustment in the proposed MAV shuttle systems under various demand scenarios. We benchmark the system performance against that in a benchmark system where only trains with 6 MP s can be dispatched. The arrival demand profile of the 18-hour operational horizon with a discretization interval of 0.1 minutes is used as the base case. To illustrate how different demand levels impact, we multiplied the demand data in the base scenario by 0.25, 0.5, 1.0 and 1.5 to obtain 4 demand datasets, labeled
sequentially from Scenario 1 to 4 in the following analysis. We use average load percentage (ALP), total energy cost (TEC) and total waiting cost (TWC) as system performance metrics. ALP is defined as $\frac{1}{K} \sum_{k \in K} d_k c_k$ while TEC and TWC refer to the first and second terms in objective function (3.2), respectively, and TC is the sum of this two terms. Note that a larger value of ALP imply better system performance while TWC, TEC and TC act in the opposite way. Further, to quantitatively compare the system performance, we define a gap measure as $\frac{V_{TCA} - V_{BEN}}{V_{BEN}}$, where $V_{TCA}$ denotes the system performance value (ALP, TWC or TEC) with TCA while $V_{BEN}$ denotes the corresponding value of the benchmark system without TCA.

Table 3.3: System performance metrics of different demand scenarios

<table>
<thead>
<tr>
<th>Demand scenario</th>
<th>With TCA</th>
<th>Without TCA</th>
<th>Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ALP (%)</td>
<td>TWC (10^3 $)</td>
<td>TEC (10^3 $)</td>
</tr>
<tr>
<td>2</td>
<td>47.66</td>
<td>10.28</td>
<td>3.66</td>
</tr>
<tr>
<td>3</td>
<td>67.78</td>
<td>104.08</td>
<td>4.08</td>
</tr>
<tr>
<td>4</td>
<td>73.16</td>
<td>680.55</td>
<td>4.59</td>
</tr>
</tbody>
</table>

The system performance metrics of all demand scenarios are summarized in Table 3.3. We see that the ALP gaps are positive while the TEC gaps are negative across all demand scenarios, indicating that the MAV shuttle system improves the vehicle capacity utilization and reduces the total energy cost. Further, while in some instances the small positive gaps of TWC reveal that the total waiting cost is slightly increased, the total system cost still experience a reduction (i.e. the TC gaps are negative). Thus, in general, introducing MAVs into shuttle systems improve the overall system performance. Further, the improvements are more significant when the passenger demand is relatively low since the maximum ALP and minimum TC are both observed in the lowest demand scenario. The main reason is that, as proved in Section 3.2, the best policy is always
to dispatch vehicles in formation $I$ in relatively high demand scenarios so there is not much possibility to adjust vehicle capacities. In contrast, a low demand scenario provides us with sufficient flexibilities to time-varying capacity design.

![Figure 3.9: Optimal MAV system design for Batong Line. Optimal dispatch headway and MAV formation in the MAV (left) and benchmark (right) design for demand scenarios 1 to 4 (from top to bottom)](image)

To explore variations of system design under various demand levels, we plot the temporal headways and MAV formations in all demand scenarios by ranking them in a descending order from top to bottom in Figure 3.9. As can be seen from these figures, the optimal headways in the MAV system and benchmark system show similar patterns over the operational horizon. Moreover, the differences between them are not evident in most cases, which indicates that the capability to improve system performance by simply adjusting the dispatch headway is limited. In contrast, vehicle capacities show great variations with demand fluctuations in the MAV system. Indeed, it is not necessary to dispatch the longest vehicle during unsaturated periods, especially in Scenarios...
1 and 2 where oversaturated periods are not present. Since the oversaturated period lasts longer in high demand scenarios, the flexibility to adjust vehicle capacity is restricted. This observation, again, justifies that the effectiveness of TCA is stronger when the demand is relatively low.

### 3.6.1.3 Sensitivity Analysis

This subsection analyzes how the effectiveness of dynamic capacity design (DCA) in the proposed MAV shuttle system varies with different input parameters. In each experiment, we vary one input parameter and keep the other parameters the same as those in the default setting. The performance metrics of all instances are plotted in Figure 3.10.

Figure 3.10 (a) shows that as the minimum dispatch headway $h$ increases, the effectiveness of DCA will be weakened because the TWC gap (i.e., the vertical distance between TWC curves) is almost 0 while the TEC gap (i.e., the vertical distance between the TEC curves) is decreasing. This is because as the minimum dispatch headway increases, the number of passengers waiting at the station at each dispatch will increase. As a result, MAVs with higher capacities are preferred yet the maximum capacity is fixed, which restricts the capability to adjust train capacities. We can also see from Figure 3.10 (a) that two TEC curves almost overlap with each other when $h$ reaches 6 minutes, which indicates that no improvements can be obtained. This implies reducing $h$, if possible, without increasing any safety risks, will amply the effectiveness of DCA.

Figure 3.10 (b) reveals that DCA brings more significant improvements when a larger $l$ is adopted since we can save more energy from hauling MAVS consisting of many MPs with low load percentages. Further, when $l$ changes from 6 to 8, TWC declines sharply because trains with larger capacities are dispatched during the oversaturated period to make as fewer passengers wait for the next train as possible, while TEC remains relatively stable. When $l$ is greater than 8, both TWC and TEC with DCA remain stable but the TEC without DCA still increases linearly. Thus,
the overall performance is still improved by DCA. Also, we find that the current maximum number of trains in a formation (i.e., 6) in Batong line is not sufficient to accommodate the intensive demand during the oversaturated period. Thus, to improve their service quality, the operator can change the maximum number of trains in a formation to 8 if possible.

Figure 3.10 (c) and Figure 3.10 (d) imply that the waiting cost per passenger per time $w$ and the fixed energy cost $C^F$ does not impact the effectiveness of DCA significantly since both the TWC and TEC gaps remain relatively stable. This is because the costs in the system with DCA and the system without DCA both increase by the same amount as these parameters increase. From Figure 3.10 (e), we learn that the growth in the coefficient for capacity-dependent vehicle energy cost, i.e., $C^V$, leads to a linear increment in the total system cost and the improvement is more significant as $C^V$ gets larger. The total waiting cost is not evidently affected but the ALP experiences a slight growth since dispatch headways become longer after the dispatch cost receives a higher weight.

Figure 3.10 (f) shows how the economics of scale affects the effectiveness of DCA. When $\alpha \leq 0.4$, the TWC and TEC curves almost overlap with each other, indicating the improvements in system costs are very limited. When $0.4 < \alpha \leq 0.8$, a gap can be observed between the TEC curves while two TWC curves are close to each other. Hence, the overall cost is reduced by DCA in this situation due to the decrease in the energy cost. When $\alpha > 0.8$, the decreasing rate in the TEC gap gets steeper and a much greater upsurge in the TWC gap is witnessed. This is because when the energy cost of large vehicles is much higher than that of small vehicles and DCA is not allowed, the model will increase the dispatch headways to save energy costs. In this situation, the improvement is attributable to the decrease in the waiting and energy costs. Thus, it is important to evaluate the economies of scale of the energy cost before introducing DCA in a shuttle system.
Figure 3.10: Sensitivity analysis of performance metrics to different input parameters of Batong Line. System performance of the proposed MAV shuttle system is benchmarked against that of a fixed capacity system where only trains with six modular units can be used.
3.6.2 Case Study 2: Future MAV Services in Tampa

In this section, we investigate a hypothetical shuttle transportation system with MAVs serving between downtown Tampa and Palm River-Clair Mel (see Figure 3.11) in 2040. The travel demand over the designated operational horizon (i.e., 6:00 a.m. to 24:00 p.m., see Figure 3.11) is predicted with an activity-based travel demand simulator, Daysim 2.0 (Wildey, 2019). To address the temporal demand fluctuations, pods (i.e., a single AV unit) are joint and detached dynamically during operation. According to information from the official website of NEXT (2018) and pre-experiments, we set $I = \{1, 2, 3, 4, 5\}$ and $c = 6$ pax/pod. Based on the bus schedules of a public transit service provider in Tampa (HART, 2018), we adopt the minimum dispatch headway in their existing schedules for this case study and thus set $h = 12$ minutes. Since no empirical data about the energy consumption of the MAV’s are available, we estimate the energy cost function as stated in Chen et al. (2020) and obtain $\alpha = 1$, $C_F = 1.912$ $\$\$ and $C_V = 3.540$ $\$\$. Besides, based on the household income information in 2017 in Tampa (Statistical Atlas, 2017), the unit time waiting cost per passenger is set as $w = 0.8$ $\$/min.

Figure 3.11: Route information (left, source: Google Maps) and arrival demand rate (right) in the hypothetical shuttle transportation system in Tampa.
We assess the performance of the system under different demand scenarios by using system operations without DCA as a benchmark. In the benchmark system, only vehicles consisting of 5 units can be dispatched. We create eight demand scenarios by multiplying the demand profile shown in Figure 3.1 by 0.1, 0.3, 0.6, 0.9, 1.0, 2.0, 3.0, 4.0. The first 5 scenarios correspond to different market penetration rates (or modal share) of the MAV services with the current travel demand between these two zones, while the last three scenarios consider further mature markets where more trips are induced by the sophisticated MAV services. All instances are solved with the CA methods due to its fast computation speed. The optimal design of dispatch headway and vehicle capacity for different demand scenarios are plotted in Figure 3.12 from top to bottom. The variations of performance metrics are shown in Figure 3.13.

As can be seen from Figure 3.12, the demand in the studied area is so low that even when the market penetration rate of the AV service reaches 100%, only vehicles in formation 1 are dispatched most of the time when TCA is allowed. Further, different from rail rapid transit systems, the minimum dispatch headway in a bus transit system is a relatively large value, so there is not much space to improve the system performance through changing the dispatch headways. As a result, the optimal solutions of the first five scenarios without TCA are basically the same, which justifies the necessity to reduce the system cost through time-varying capacity design. When additional trips are induced, larger vehicles are needed due to the oversaturated traffic. In these scenarios, both operational strategies tend to select the minimum dispatch headway so the total waiting costs are the same. Yet by adjusting vehicle capacities based on the time-dependent demand, TCA can not only increase the vehicle utilization rate but reduce the total energy costs (Figure 3.13).
Figure 3.12: Optimal design for future MAV services in Tampa. Optimal dispatch headway and vehicle formation with (left) or without (right) MAVs for demand scenarios 1 to 8 from top to bottom.

Figure 3.13: Performance metrics under various demand scenarios.
3.7 Chapter Summary

This chapter investigates the joint design problem of dispatch headway and vehicle formation for a MAV shuttle system under oversaturated traffic. We prove a series of theoretical properties of the optimal operational design. Based on the theoretical properties, a continuous model and a discrete model are proposed to solve the near and exact optimal solution to the investigated problem, respectively. The main findings are summarized as follows.

(i) The customized DP algorithm with valid inequalities can produce exact solutions to the joint design problem with relatively expensive computational costs (but still much less than Gurobi) to solve the discrete model. In contrast, the CA model offers near-optimal solutions with acceptable errors and dramatically reduced computational time.

(ii) It is optimal to dispatch vehicles with the maximum capacity and the minimum headway during oversaturated periods. The operational design problem can actually be solved by simply solving its revised unsaturated problem, whose optimal solution ensures that the queue length right after each dispatch is less than the unit vehicle capacity.

(iii) By incorporating dynamic capacity adjustment into transit scheduling, MAV shuttle systems can improve the vehicle utilization rate and reduce the total system cost in both unsaturated and oversaturated traffic. The effectiveness of this innovative operational strategy is stronger when unsaturated traffic is present since vehicle capacities can be changed more flexibly.

(iv) The larger the maximum vehicle type and the added dispatch cost per unit capacity, the more significant the improvements will be. However, longer minimum dispatch headways will diminish the effectiveness of dynamic vehicle scheduling but improvements can still be achieved in certain ranges. Further, the unitless parameter in the dispatch cost function and the cost of changing vehicle lengths are vital parameters that should be analyzed before implementing
dynamic capacity adjustment since system performances may be degraded if these parameters are not within a certain range.

Overall, this chapter provides a methodological foundation to obtain the exact and near optimal solution to the operational design in a MAV shuttle system and reveals important managerial insights to different UMT operators, such as subways, modular autonomous vehicles, etc. Also, this study constitutes a building block to develop methodologies for solving similar problems in other system settings with more complicated demand pattern and operational details.
Chapter 4: Corridor Systems

4.1 Overview

This chapter investigates the incorporation of MAVs into corridor transportation systems. Different from the shuttle system investigated in the last chapter, a corridor system involves a more complicated many-to-many demand pattern because passengers can travel between any station pair along the corridor. Further, while MAVs only change their formations at the origin (or first) station in a shuttle system, MPs can dock or undock into vehicles of different lengths at any station along the corridor, which we name station-wise docking in this study. The station-wise docking operation may bring further system cost savings but also substantially increases the solution space of the operational design problem, thus resulting in a new challenging problem that the solution methods for the shuttle system design cannot address.

To understand how the station-wise docking operation will potentially bring substantial benefits to UMT systems, consider a UMT corridor with 3 stations as shown in Figure 4.1. MAVs consisting of MPs (each pod can accommodate 2 passengers) are adopted in the system and the operational cost arising from a MAV running between two consecutive stations equals the number of pods in it. Assume that in the existing operation, vehicles with 4MPs can be dispatched, which results in an operational cost of 8, as shown in Figure 4.1(b). If time-varying capacity is adopted but vehicles can only change their formations at the terminals, then a vehicle with 3 MPs will be dispatched to serve all passengers, resulting in an operational cost of 6, as shown in Figure 4.1(c).

---

3 Portions of this chapter has been published in Chen and Li (2020a), Chen and Li (2020b). Permission is included in Appendix A.
Nevertheless, if MAVs and station-wise docking are introduced into the system, we can drop one MP at the second station because there are only 4 passengers, as shown in Figure 4.1(d). The reduction in the number of MPs decreases the operational cost to 5. The station-wise docking operation also increases the average vehicle occupancy from 63% to 100%. Thus, allowing station-wise docking along a transportation corridor may improve the performance of UMT systems.

Figure 4.1: An illustrative example for different operational modes. (a) a corridor system with 3 stations (numbers above the arrows representing the passenger demand); (b) existing fixed-capacity operation; (c) varying-capacity operation with docking at terminals only; (d) proposed varying-capacity operation with station-wise docking.

Despite its potential and simplicity, MAV-based transportation corridors with station-wise docking has not been well investigated. This chapter aims to bridge this research gap by proposing mathematical models for designing the dispatch time and vehicle formation for UMT corridors where vehicles can flexibly adjust their capacity by docking and undocking at any station. We follow a similar design philosophy adopted in Chapter 3 with additional challenges due to the many-to-many demand pattern and station-wise docking operation addressed. Further, we intend to study whether the station-wise docking operation can indeed improve system performance via case study applications of the proposed design methods. The contributions of this chapter are summarized as follows.
(i) We investigate theoretical properties concerning the optimal vehicle formations, the relationship between the vehicle dispatch headway and passenger queue, and a dominance rule between feasible solutions. These properties break the spatiotemporal correlation between dispatch decisions, thus constituting a theoretical foundation for formulating the investigated problem under the CA framework. They also serve as a basis for deriving algorithmic techniques to expedite exact solution algorithms for solving exact solutions to the investigated problem.

(ii) We investigate the operational design of MAV corridor systems enabling station-wise docking and undocking under the CA framework. The CA model presents a macroscopic view of the system and yields a simple analytical solution approach. The analytical results enable efficient solution methods for relevant large-scale problem instances and offer managerial insights to system operators. With this, the CA methodology is advanced to the joint design of dispatch headway and vehicle capacity that considers many-to-many demand patterns, station-wise vehicle capacity adjustment and other factors (e.g., passenger boarding order, minimum dispatch headway).

(iii) A mixed integer linear programming model is proposed to solve the exact optimal solution to the operational design. The theoretical properties are used to develop an expedited branch and bound (B&B) algorithm to solve the discrete model for small- and medium- size instances. The discrete model complements the macroscopic perspective of the continuous method by explicitly specifying microscopic discrete details of UMT operations. Further, it provides a benchmark exact algorithm to verify the solution quality and computation performance for research efforts that aim to develop other solution algorithms for the investigated problem, including our CA method.

(iv) Numerical experiments verify the validity and efficiency of the proposed solution method and provide knowledge of the innovative station-wise docking operations. Results show
that the expedited B&B algorithm significantly outperforms a state-of-the-art commercial solver, Gurobi. The proposed CA method can produce near-optimal solutions to the investigated problem very efficiently in extensive parameter settings. Also, managerial insights into the effectiveness of station-wise docking and its performance under various parameter settings are provided.

The remainder of this chapter is organized as follows. Section 4.2 introduces the problem setting and general problem formulation. Section 4.3 investigates theoretical properties of the problem. Based on the theoretical properties, Section 4.4 presents the CA model along with the analytical solution approach and discretization method. Section 4.5 presents the discrete model and the expedited B&B algorithm. Section 4.6 presents the results from the numerical experiments. Finally, Section 4.7 concludes this chapter.

4.2 Problem Statement

We consider a unidirectional UMT corridor system where a set of stations indexed as \( s, u \in S := [1,2,\cdots, S] \) are placed along an urban transportation corridor, with 1 and \( S \) being the start and terminal stations, as shown in Figure 4.2. The sets of upstream and downstream stations of \( s \in S \) are denoted as \( S_s^- \) and \( S_s^+ \), respectively. Passengers arrive at each station continuously during an operational horizon \([0, T]\). We describe the passenger arrival process at station \( s \in S \) with cumulative arrival demand curves denoted as \( A_s(t), \forall t \in [0, T] \) (the solid curves in Figure 4.2). The corresponding passenger arrival rate \( \forall s \in S \) is thus defined as \( a_s(t) := A'_s(t) \). To capture the distribution of passenger destinations from origin station \( s \in S \) among downstream stations \( u \in S_s^+ \), we denote the proportion of passengers arriving at station \( s \) and heading to station \( u \) at time \( t \) (among all passengers arriving at \( s \) at \( t \)) as \( p_{su}(t), \forall s \in S, u \in S_s^+, t \in [0, T] \). Note that this study addresses a deterministic optimization problem where the cumulative arrival curves \( A_s(t) \) are assumed a priori rather than a stochastic or uncertain process. To serve the passengers, a set of \( K \)
MAVs are dispatched across the operational horizon, indexed as \( k \in \mathcal{K} := [1, 2, \ldots, K] \) where the index increases with the dispatch time. Note that in the investigated problem \( K \) is a decision variable. These dispatches result in a cumulative passenger departure curve at each \( s \in \mathcal{S} \) denoted as \( D_s(t), \forall t \in [0, T] \). The time at the start station and vehicle formation at each station \( s \) for each dispatch \( k \) are denoted as \( t_k, \forall k \in \mathcal{K} \) and \( i_{ks}, \forall k \in \mathcal{K} \), respectively. For the convenience of notation, define \( t_0 := 0 \) and \( t_{K+1} := T \). Below we illustrate the operational concept and basic assumptions of the proposed MAV-based transportation corridor systems.

(i) The dispatched vehicles move along the corridor from the start to the terminal stations. The vehicles’ dwell time at each station are relatively constant and the travel time between two consecutive stations are time-dependent. To simplify the model formulation, we define the operational offset of station \( s \) as the difference between the travel times a vehicle leaves Stations

Figure 4.2: The corridor system (top) and cumulative passenger counts (bottom).
1 and s (Sun et al., 2014). Then by subtracting the operational offset from the original time coordinate at station \( s \), \( \forall s \in S \), we unify the time coordinates of Stations 1 and s since now the departure times of any vehicle at these stations are the same. The passenger demand arrival curves are shifted along with the time coordinate.

(ii) Different from vehicles in existing UMT systems, MAVs can flexibly change their formations (i.e., length or capacity) at any station along the corridor by docking and undocking modular pods. We index vehicle formations as \( i \in I := [1,2,\ldots, I] \), where \( I \) is the maximum allowable number of pods in one MAV. Each vehicle formation \( i \in I \) consists of \( i \) identical modular pods and therefore has a capacity of \( ic \), where \( c \) is the capacity of a single pod. Again, we assume that the number of modular pods is always sufficient so that there are always some vehicles in each formation for dispatching at each station.

(iii) Following previous studies (e.g., Niu et al., 2015 and Yin et al., 2017), we further assume that oversaturated traffic is not allowed at each station. However, note that multiple vehicle formations are considered in the investigated problem. Thus, different from previous studies, this assumption incorporates more general cases where passengers waiting at the stations are not cleared after each dispatch.

(iv) At each station \( s \in S \), passengers board a vehicle following a First-In-First-Serve (FIFS) principle. In reality, the passenger boarding order is extremely difficult to determine (especially when there are multiple destinations), but in general passengers who arrive early have more chances to board a vehicle than those arriving later. Therefore, this assumption is not too distant from the reality (Niu and Zhou, 2013). With this assumption, we can decompose the cumulative arrival curve \( A_s(t) \) as \( T \) time-dependent arrival rate curves

\[
A_s(t', t) = \begin{cases} 
0 & t \in [0, t') \\
\alpha_s(t^0) & t \in [t', T] 
\end{cases}, \forall s \in S, t' \in [0, T],
\]
each of which corresponds to the demand rate at time $t' \in [0, T]$, respectively. Likewise, the cumulative departure curve at station $s$, $D_s(t)$, can be decomposed as $T$ departure rate curves, denoted as $D_s(t', t), \forall s \in S, t', t \in [0, T]$. To facilitate the illustration of these definitions, consider a simple example with $A_s(t)$ and $D_s(t)$ shown in Figure 4.3 (a). Note that here the time horizon is discretized into three time intervals for illustrative purposes. We see from Figure 4.3 (a) that 2, 2, and 1 passengers arrive at station $s$ at time 1, 2, and 3, respectively. Three MAVs are dispatched to serve the passengers. The decomposed curves are shown in Figure 4.3 (b) and the projections of these curves for different $t'$ are plotted in Figure 4.3 (c) - (e). Take passengers arriving at time 2 (i.e., $t' = 2$) as an example. The time-dependent arrival rate curve $A_s(2, t) = 0, \forall t \in [0, 2)$ and $A_s(2, t) = a_s(2) = 2, \forall t \in [2, 3]$ while the time-dependent departure rate curve $D_s(2, t) = 0, \forall t \in [0, 2), D_s(2, t) = 1, \forall t \in [2, 3)$ (1 passenger arriving at $t' = 2$ has been transported), and $D_s(2, 3) = 2$ (all passengers arriving at $t' = 2$ has been transported). The decomposed curves for other times can be illustrated in a similar way. With these definitions, we further define the number of passengers who arrive at station $s$ at time $t'$ and board dispatch $k$ as $d_{ks}(t') := D_s(t', t_k) - D_s(t', t_{k-1}), \forall k \in K, s \in S, t' \in [0, t_k]$. For dispatch $k$, all passengers arriving at station $s$ no later than $t_k$ can get onboard as long as there is still enough capacity. Thus, $\int_0^{t_k} d_{ks}(t') dt'$ represents the number of passengers boarding dispatch $k$ at station $s$. Note that the right derivative is used at points that are not differentiable (this applies throughout the paper). For example, for the second dispatch in Figure 4.3, we obtain $d_{2s}(1) = 1, d_{2s}(2) = 1, \text{and } d_{2s}(3) = 0$. Summing over all these values yields the number of passengers boarding at the second dispatch at stations, i.e., 2. Note that here summation is used since the time horizon is discretized while in a continuous time horizon an integral of $d_{ks}(t')$ with respect to $t'$ over 0 to $t_k$ is needed.
Figure 4.3: An example for cumulative arrival and departure curves decomposition. (a) The original cumulative arrival and departure curves. (b) The decomposed time-dependent arrival and departure rate curves. (c) The decomposed curves corresponding to arrival demand at time 1. (d) The decomposed curves corresponding to arrival demand at time 2. (e) The decomposed curves corresponding to arrival demand at time 3.

These operations result in two cost components in the system, i.e., the operational cost and passenger waiting cost (Yin et al., 2017). We assume that the operational cost of a vehicle in formation $i$, denoted as $f_i$, is concave over $i$ (Danganzo, 2005) as required by Eq. (3.1). With this, the total operational cost in the system can be formulated as

$$C_f(i_{ks}) := \sum_{s \in S} \sum_{k \in K} f_{i_{ks}}$$

Further, according to Newell (1971), the total passenger waiting time at station $s \in S$ can be quantified as the area between its cumulative arrival curve and cumulative departure curve, i.e., $\int_0^T \left( (A_s(t) - D_s(t)) \right) dt$. Since the passenger boarding order needs to be considered in this study, we have to formulate the passenger waiting time with the time-dependent arrival and departure rate curves. Specifically, based on the definition of $A_s(t', t)$ and $D_s(t', t)$, the term
\[ \int_{t=0}^{T}(A_s(t', t) - D_s(t', t))dt \] is the total waiting time for passengers who arrive at station \( s \) at \( t \).

Then, the integral of this term over \( t' \) from 0 to \( T \), i.e., \( \int_{t'=0}^{T} \int_{t=0}^{T}(A_s(t', t) - D_s(t', t))dt \ dt' \), is the total waiting time for all passengers arriving at station \( s \) during time horizon \([0, T]\). Thus, given a homogeneous unit-time waiting cost \( w \) for each passenger, the total passenger waiting cost can then be formulated as

\[
C_w (D_s(t', t)) := w \sum_{s \in S} \int_{t'=0}^{T} \int_{t=0}^{T} (A_s(t', t) - D_s(t', t))dt \ dt'
\] (4.2)

The objective of this study is to find an optimal schedule for corridor systems satisfying the aforementioned operational concepts such that the desired objective of the system operator can be achieved. The major decisions include the number of dispatches \( K \), dispatch time \( t_k \), vehicle formation \( i_{ks} \), and cumulative departure curve \( D_s(t) \). Along with the optimal schedule, some auxiliary decision variables will also be determined, including the time-dependent departure rate curves \( D_s(t', t) \), the number of passengers arriving at station \( s \) at time \( t' \) and boarding at the \( k \)-th dispatch \( d_{ks}(t') \), and additional variables \( e_{ks}, \forall k \in K, s \in S \) defined as the number of passengers onboard at dispatch \( k \) after passenger boarding and alighting at station \( s \). Following previous studies (e.g., Yin et al., 2017), we aim to search for an optimal trade-off between the operational cost and passenger waiting cost in the system. Thus, we formulate the objective function of the investigated problem as:

\[
\min_{K, [t_k, i_{ks}, e_{ks}, \forall k \in K, s \in S], [d_{ks}(t'), D_s(t', t), \forall s \in S, k \in K, t', t \in [0, T]]} \left[ C_f (i_{ks}) + C_w (D_s(t', t)) \right],
\] (4.3)

where the first term represents the total vehicle operational cost across all dispatches while the second term the total passenger waiting cost across the operational horizon. To reflect the general operational details of UMT systems, the following four groups of constraints are considered.
(i) Minimum headway requirement. These constraints impose a least-time separation between two consecutive dispatched vehicles due to safety considerations and limited resources.

\[ t_k - t_{k-1} \geq h_k, \forall k \in \mathcal{K}\setminus\{1\}, \quad (4.4) \]

(ii) Departure curve conservation. These constraints formulate the departure conservation process mathematically with the time-dependent departure rate curve \( \mathcal{D}(t', t) \). Constraints (4.5) define that \( \forall t \in [t', T], \mathcal{D}(t', t) \) increases by the number of boarding passengers if a vehicle is dispatched at \( t \) (i.e., \( t = t_k \)) and remains the value at the previous dispatch otherwise (i.e., \( t \in (t_k, t_{k+1}) \)). Constraints (4.6) require \( \mathcal{D}(t', t) \) to equal \( A_s(t', t) \) at \( T \), indicating that all passengers arriving at station \( s \) during the investigated time horizon are transported at the end of the operational horizon.

\[
\begin{align*}
D_s(t', t) &= \begin{cases} 
D_s(t', t_{k-1}) + d_{ks}(t') & \forall t = t_k, \forall k \in \mathcal{K}, s \in \mathcal{S}, t' \in [0, T], \\
D_s(t', t_k) & \forall t \in (t_k, t_{k+1})
\end{cases}, \\
D_s(t', t) &= A_s(t', t), \forall t = T, s \in \mathcal{S}, t' \in [0, T].
\end{align*}
\tag{4.5} \tag{4.6}
\]

(iii) Passenger boarding dynamics. These constraints model the passenger boarding process. Constraints (4.7) indicate passengers boarding at dispatch \( k \) at the start station may arrive at this station at any \( t' \) until \( t_k \). At the other stations, the number of passengers on dispatched vehicle \( k \) equals the number of passengers on vehicle \( k \) at the previous station \( s - 1 \) plus the number of passengers boarding at station \( s \) minus the number of passengers alighting at station \( s \). Constraints (4.8) are imposed due to the capacity limit of each vehicle formation. Constraints (4.9) require that the number of passengers boarding dispatch \( k \) at station \( s \) at time \( t \) should be no more than the passenger demand at station \( s \) at time \( t' \). Constraints (4.10) describe the FIFS rule for passenger boarding.
\[ e_{ks} = \begin{cases} 
\int_{t'=0}^{t_k} d_{ks}(t') dt' & \text{if } s = 1, \forall k \in \mathcal{K} \\
\sum_{u \in \mathcal{S}_s} \int_{t'=0}^{t_k} d_{ku}(t') p_{us}(t') dt' + \sum_{u \in \mathcal{S}_s} \int_{t'=0}^{t_k} d_{ku}(t') p_{us}(t') dt' & \forall s \in \mathcal{S} \setminus \{1\}, \forall k \in \mathcal{K} 
\end{cases} \] (4.7)

\[ e_{ks} \leq i_{ks} c, \forall k \in \mathcal{K}, s \in \mathcal{S} \] (4.8)

\[ d_{ks}(t') \leq A_s(t', t_k) - D_s(t', t_{k-1}), \forall k \in \mathcal{K}, s \in \mathcal{S}, t' \in [0, t_k] \] (4.9)

\[ \frac{D_s(t'', t)}{A_s(t'', t)} \leq \frac{D_s(t', t)}{A_s(t', t)} , \forall t' < t'' \in [0, T], t \in [0, T], s \in \mathcal{S}, \] (4.10)

(iv) Variable domains. These constraints define the domains of all decision variables.

\[ K \in \mathbb{Z}^+; t_k \in [0, T], \forall k \in \mathcal{K}; i_{ks} \in \mathbb{I}, e_{ks} \in \mathbb{R}^+, d_{ks}(t') \in \mathbb{R}^+, \forall k \in \mathcal{K}, s \in \mathcal{S}, t' \in [0, T]. \] (4.11)

### 4.3 Theoretical Property Analysis

This section presents the theoretical properties of the optimal solution(s) to the investigated problem. The proofs of the following lemmas and theorems are available in Appendix B.

We first rewrite Constraints (4.7) and (4.9). Integrating both sides of Constraints (4.9) over \( t \) yields

\[ \int_{t'=0}^{t_k} d_{ks}(t') dt' \leq \int_{t'=0}^{t_k} \left( A_s(t', t_k) - D_s(t', t_{k-1}) \right) dt', \forall k \in \mathcal{K}, s \in \mathcal{S} \]

Applying this inequality into Constraints (4.7) yields

\[ e_{ks} \leq \begin{cases} 
\int_{t'=0}^{t_k} \left( A_s(t', t_k) - D_s(t', t_{k-1}) \right) dt' & s = 1 \\
\sum_{u \in \mathcal{S}_s} \int_{t'=0}^{t_k} d_{ku}(t') p_{us}(t') dt' + \sum_{u \in \mathcal{S}_s} \int_{t'=0}^{t_k} d_{ku}(t') p_{us}(t') dt' & s > 1 \end{cases} \] (4.12)

Further, we define the number of passengers left at station \( s \in \mathcal{S} \) right after a dispatch \( k \) as the passenger queue at station \( s \), i.e.,
\[ q_s(t_k) := \left( \int_{t' = 0}^{t_k} (A_s(t', t_k) - D_s(t', t_k)) dt' \right), \forall s \in S, \]

and the number of passengers who are left at any upstream station of \( s \) (inclusive) and destined to any station in the downstream direction of station \( s \) as the cross-sectional passenger queue at station \( s \), i.e.,

\[ \tilde{q}_s(t_k) := \sum_{u \in S_{s+1}} \sum_{v \in S_s} \left( \int_{t' = 0}^{t_k} (A_u(t', t_k) - D_u(t', t_k)) p_{uv}(t') dt' \right), \forall s \in S. \]

Define \( \tilde{q}_s(t_0) = 0, \forall s \in S \) for the convenience of the notation. With these, we next investigate properties of queued passengers after each dispatch in the optimal solution(s) in the following three propositions.

**Proposition 4.1.** An optimal solution \( \{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\} \) to problem (4.3) ~ (4.11) with time-dependent arrival rate curves \( A_s(t', t), \forall s \in S \) and time-dependent departure rate curves \( D_s(t', t), \forall s \in S \) must satisfy one of the following conditions for each dispatch \( k \in K \): (i) \( t_k - t_{k-1} = h \); or (ii) there exists at least one \( s \in S \) such that \( q_s(t_k) = 0 \).

**Proposition 4.2.** Any dispatch \( k \) in an optimal solution \( \{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\} \) to problem (4.3) ~ (4.11) with time-dependent arrival rate curves \( A_s(t', t), \forall s \in S \) and departure rate curves \( D_s(t', t), \forall s \in S \) must satisfy \( \tilde{q}_s(t_k) < c, \forall s \in S \) if the following conditions hold:

(i) \( S(f_2 - f_1) < ch \),

(ii) \( \sum_{u \in S_{s+1}} \sum_{v \in S_s} \left( \int_{t' = t_{k-1}}^{t_k} (A_u(t', t')) p_{uv}(t') dt' \right) \leq (l - 1)c, \forall s \in S. \)

**Proposition 4.3.** Any dispatch \( k \) in an optimal solution \( \{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\} \) to problem (4.3) ~ (4.11) with time-dependent arrival rate curves \( A_s(t', t), \forall s \in S \) and departure rate curves \( D_s(t', t), \forall s \in S \) must satisfy \( q_s(t_k) < c, \forall s \in S \) if the following conditions hold:

(i) \( S(f_2 - f_1) < ch \).
(ii) $\sum_{u \in S_{s+1}} \sum_{v \in S_s^+} \left( \int_{t' = t_{k-1}}^{t_k} (A_u(t', t')) p_{uv}(t') dt' \right) \leq (l - 1)c, \forall s \in S$.

Proposition 4.1 indicates that the dispatch headway should be the minimum usable value unless the passenger queue is 0 at one or more stations. Note that no special conditions are needed for Proposition 4.1 and thus it is a general property for the investigated problem. Propositions 4.2 and 4.3 indicate that the number of passengers after each dispatch would be relatively small values (i.e., less than the capacity of a single MP) at each station. While the conditions in Propositions 4.2 and 4.3 are not universal, they enable us to draw analytical insights into the optimal solution structure that would otherwise be impossible. Further, the applicability domain of this theoretical result is quite broad in real-world cases. Condition (i) usually holds for common UMT systems since passengers waiting cost is usually of a higher magnitude compared with the associated vehicle operational cost (Yin et al., 2017; Huang et al., 2017). Condition (ii) is generally approximately true under the unsaturated demand scenario. Even for cases where the conditions are not satisfied, we found that the resultant approximation error of a continuous approximation model guided by these theoretical findings are relatively small. Indeed, the passenger queue at each station can be cleared (not just bounded by $c$) after each dispatch most of the time in the optimal solution(s). Hence, the conditions stated in these propositions, albeit non-exhaustive, provide important analytical insights into the nature of the (near-) optimal solutions and a theoretical foundation for designing efficient solution methods for the investigated problem.

With these results, we can further investigate properties of the optimal vehicle formations. Define the seat demand at time $t$ at station $s$, denoted as $\tilde{a}_s(t)$, as those who would travel through the segment between station $s$ and $s + 1$ at this time, i.e.,

$$\tilde{a}_s(t) := \sum_{u \in S_{s+1}} \sum_{v \in S_s^+} (A_u(t', t')) p_{uv}(t) = \sum_{u \in S_{s+1}} \sum_{v \in S_s^+} (a_u(t) p_{uv}(t)), \forall s \in S, t \in [0, T].$$
For dispatch $k$, the actual seat demand at station $s$ also includes the cross-sectional passenger queue of this station from the previous dispatch. Thus, the seat demand for dispatch $k$ at station $s$ can be formulated as $\tilde{a}_{ks} := \tilde{q}_s(t_{k-1}) + \int_{t_{k-1}}^{t_k} \tilde{a}_s(t) \, dt$, $\forall k \in \mathcal{K}, s \in \mathcal{S}$. With this, we define the lower-bound vehicle formation as $i_{ks} := \min \left\{ \left\lfloor \frac{\tilde{a}_{ks}}{c} \right\rfloor, I \right\}$, which represents the longest vehicle formation whose capacity is smaller than the seat demand (i.e., the shortest one that results in $\tilde{q}_s(t_k) < c$). Likewise, we define the upper-bound vehicle formation as $\tilde{i}_{ks} := \min \left\{ \left\lceil \frac{\tilde{a}_{ks}}{c} \right\rceil, I \right\}$, which represents the shortest vehicle formation whose capacity is greater than the seat demand (i.e., the shortest one that results in $\tilde{q}_s(t_k) = 0$). With these definitions, we obtain the following two propositions regarding the lower bound and upper bound to the optimal vehicle formation.

**Proposition 4.4.** Any dispatch $k$ in an optimal solution $\{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\}$ to problem (4.3) ~ (4.11) with time-dependent arrival rate curves $A_s(t', t), \forall s \in \mathcal{S}$ and departure rate curves $D_s(t', t), \forall s \in \mathcal{S}$ must satisfy $i_{ks} \leq i_{ks}, \forall s \in \mathcal{S}$ if the following conditions hold:

(i) $S(f_2 - f_1) < cwh$,

(ii) $\sum_{u \in \mathcal{S}_s} \sum_{v \in \mathcal{S}_s} \left( \int_{t' = t_{k-1}}^{t_k} \left( A_u(t', t') \right) p_{uv}(t') \, dt' \right) \leq (I - 1)c, \forall s \in \mathcal{S}$.

**Proposition 4.5.** An optimal solution $\{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\}$ to problem (4.3) ~ (4.11) must satisfy $i_{ks} \leq \tilde{i}_{ks}, \forall k \in \mathcal{K}, s \in \mathcal{S}$.

### 4.4 Continuous Model

This section presents the CA approach to solving near-optimal solutions to the investigated problem efficiently. We first present how we approximate the original problem in a relatively homogeneous setting and the resulting problem decomposition. Then, analytical solutions to the decomposed problems and discretization approach are discussed.
4.4.1 Model Formulation

With the theoretical findings in the previous subsection, we can reformulate the original problem (4.3) ~ (4.11) under the CA framework. Given an optimal solution $o := \{t_k, i_{ks}\} \forall k \in \mathcal{K}, s \in \mathcal{S}$ to problem (4.3) ~ (4.11) with time-dependent arrival rate curves $A_s(t', t)$ and time-dependent departure rate curves $D_s(t', t)$, we define the dispatch headway as $\hat{h}(t)$, and the vehicle formation at station $s$ at time $t$ as $i_s(t), \forall s \in \mathcal{S}$ in the same way as Chapter 3. With this, the objective value corresponding to optimal solution $o$ can be formulated as:

$$OP(o) := \sum_{s \in \mathcal{S}} \left( \sum_{k \in \mathcal{K}} f_{ks} + w \int_{t' = 0}^{T} \int_{t = 0}^{T} (A_s(t', t) - D_s(t', t)) dt \ dt' \right)$$

$$= \sum_{s \in \mathcal{S}} \left( \int_{t = 0}^{T} \left( f(i_s(t)) \right) dt + w \int_{t' = 0}^{T} \int_{t = 0}^{T} (A_s(t', t) - D_s(t', t)) dt \ dt' \right). \quad (4.13)$$

s.t. (4.4) ~ (4.11).

Propositions 4.1 – 4.3 and empirical experiments show that the passenger queue at each station is relatively small after each dispatch most of the time in the optimal solution(s) when only unsaturated traffic is present. This finding enables us to relax the FIFO rule when describing the passenger boarding behavior. Therefore, we can replace the time-dependent arrival and departure rate curves by the original ones to compute the passenger waiting cost. This results in a first approximation to the original objective function (4.4) as

$$OP(o) \approx \sum_{s \in \mathcal{S}} \left( \sum_{k \in \mathcal{K}} \int_{t = t_{k-1}}^{t_k} \left( \frac{f(i_s(t))}{\hat{h}(t)} + w(A_s(t) - D_s(t)) \right) dt \right). \quad (4.14)$$

Note that $A_s(t)$ is very close to $D_s(t)$ at $t_k, \forall k \in \mathcal{K}$ (Propositions 4.1 – 4.3). Hence, we have

$$a_s(t) = \frac{A_s(t) - A_s(t_{k-1})}{t - t_{k-1}} \approx \frac{A_s(t) - D_s(t_{k-1})}{t - t_{k-1}}, \forall t \in [t_{k-1}, t_k), \forall k \in \mathcal{K},$$

which yields

78
\[ OP(o) \approx \int_0^T \left[ \sum_{s \in S} \left( f\left(\hat{i}_s(t)\right) \frac{w_{as}(t)\hat{h}(t)}{2} \right) \right] dt \] (4.15)

The difference between (4.14) and (4.15) is

\[ \sum_{s \in S} \left[ \sum_{k \in K} \int_{t_k-1}^{t_k} \left( (A_s(t_{k-1}) - D_s(t_{k-1})) \right) dt \right] < cTS, \]

which is a relatively small value compared with \( OP(o) \). Additionally, numerical experiments with exact solution methods show that \( A_s(t_k) - D_s(t_k) = 0 \) most of the time. Therefore, we feel safe to say that this approximation is reasonable, particularly for large scale instances. Solving the investigated problem with approximated objective function (4.15), however, is not much different from solving it with the original one, since both \( \hat{h}(t) \) and \( \hat{i}_s(t), \forall s \in S \) are step functions. To further simplify the formulation, we replace \( \hat{h}(t) \) and \( \hat{i}_s(t), \forall s \in S \) with continuous functions \( h(t) \) and \( i_s(t), \forall s \in S \), respectively, and obtain

\[ OP(o) \approx \int_0^T \left[ \sum_{s \in S} \left( f\left(\hat{i}_s(t)\right) \frac{w_{as}(t)h(t)}{2} \right) \right] dt \] (4.16)

where the operational cost of continuous vehicle formation \( i_s(t) \in [0, I] \) is defined as \( f(\hat{i}_s(t)) = f_i, \text{ s.t. } i_s(t) \in (i - 1, i], \exists i \in I, \forall t \in [0, T], s \in S \). For the convenience of the notation, define \( f(0) = 0 \). Until now, the spatiotemporal correlation between dispatch decision is eliminated, and we can solve the problem for each time point independently. However, the approximated objective function (4.16) still has \( S + 1 \) number of decision variables, for which the solution approach might not be very simple. Thus, we need to further simplify the model formulation. Specifically, we obtain that for \( t \in [t_{k-1}, t_k), \forall k \in K \), \( i_s(t)c \approx \hat{i}_s(t)\hat{c} = \hat{i}_s(t_k)c \approx \int_{t_{k-1}}^{t_k} \tilde{a}_s(t) \, dt \approx (t_k - t_{k-1})\tilde{a}_s(t) \, dt = h(t)\tilde{a}_s(t) \). Applying this relationship into Eq. (4.16) yields the final approximate objective function with only one decision variable as
Clearly, the $h(t)$ value that minimizes Eq. (4.17) also minimizes the integrand at every $t \in [0, T]$, so the original problem can be decomposed across the operational horizon as the following set of sub-problems, each for one time point $t$.

$$
c^*(t) := \min_{h(t)} c_t(h(t)) := \sum_{s \in S} \left( f \left( \frac{h(t) \bar{a}_s(t)}{c} \right) + \frac{w(a_s(t)h(t))}{2} \right), \forall t \in [0, T].
$$

(4.18)

s.t.

$$
h \leq h(t) < \frac{I_c}{\bar{a}_s(t)}, \forall s \in S, t \in [0, T].
$$

(4.19)

where Constraints (4.19) are imposed due to the minimum headway constraints and the vehicle capacity constraints in the original problem. Note that the above unit-time problems at each time point have only one decision variable and two constraints and they can be independently solved analytically at each time $t \in [0, T]$. Thus, these single variable optimization problems are much simpler to solve than the original problem. After solving the unit-time problems across the entire operational horizon, we can apply $c^*(t)$ into the following equation to obtain an approximate objective value to optimal solution $o$:

$$
OP(o) \approx \int_0^T c^*(t) dt.
$$

(4.20)

4.4.2 Analytical Solution

This subsection presents an analytical solution approach to unit-time problems (4.18) - (4.19). The proposed solution approach first divides the feasible region of $h(t)$ into a finite number of sub-regions in each of which a local minimum to objective function (4.18) exists. Then the local minimum in each sub-region can be analytically solved. Finally, a comparison between all local
minima can easily lead us to the global minimum of the unit-time problems. Note that the following discussion applies to each \( t \in [0, T] \), but we omit the index \( \forall t \in [0, T] \) in the formulas for the convenience of the notation.

Figure 4.4: Sub-region division for the feasible region of \( h(t) \) for unit-time problems.

We first present how to divide the feasible region into a finite number of sub-regions where \( c_t(h(t)) \) is strictly unimodal, as shown in Figure 4.4. According to the definition of \( f(i_s(t)) \), for each station \( s \in S \), \( f\left(\frac{h(t)a_s(t)}{c}\right) \) is a step function specified by a set of discrete points (joints between pairs of consecutive pieces of the function) denoted as \( \mathcal{H}_s(t) := \left\{ \frac{ic}{a_s(t)} \left| \frac{ic}{a_s(t)} \leq \min \left\{ \frac{ic}{a_s(t)}, T \right\}, \forall i \in I \right\} \), where the elements are ascendingly ordered. With this, we can divide the feasible region for all possible headway values into a finite number of sub-regions with a set of segregation points denoted as an ascendingly ordered set \( \mathcal{H}(t) := \bigcup_{s \in S} \mathcal{H}_s(t) \).
For the convenience of the notation, denote the $l$-th element in $\mathcal{H}(t)$ as $h_l(t)$ in the following analysis. Define $X_{l-1}(t) := \sum_{s \in S} \left( f \left( \frac{h(t) \Delta_s(t)}{c} \right) \right) > 0$, $Y(t) := \sum_{s \in S} \left( \frac{w_{as(t)} h(t)}{2} \right) \geq 0$, which are both constant over $h(t) \in (h_{l-1}(t), h_l(t)]$ based on the definition of the step function. Then, $c_t(h(t)) = \frac{X_{l-1}(t)}{h(t)} + Y(t) h(t)$ is a strictly convex EOQ (economic-order-quantity) function of $h(t)$. Solving $c_t'(h(t)) = 0$ yields $h(t) = \frac{X_{l-1}(t)}{\sqrt{Y(t)}}$ and we obtain an analytical solution to the optimal $h(t)$ in sub-region $(h_{l-1}(t), h_l(t)]$ denoted as $h_{l-1}^*(t)$ as follows:

\[
h_{l-1}^*(t) = \begin{cases} 
    h_{l-1}(t) & \text{if } \frac{X_{l-1}(t)}{Y(t)} < h_{l-1}(t) \\
    \frac{X_{l-1}(t)}{\sqrt{Y(t)}} & \text{if } h_{l-1}(t) \leq \frac{X_{l-1}(t)}{\sqrt{Y(t)}} \leq h_l(t), \forall l \in \{2, \cdots, |\mathcal{H}(t)|\}.
    h_l(t) & \text{if } \frac{X_{l-1}(t)}{\sqrt{Y(t)}} > h_l(t)
\end{cases}
\]  \tag{4.21}

Then we can obtain the local minimum in sub-region $(h_{l-1}(t), h_l(t)]$, denoted as $c_{l-1}^*(t)$, by applying Eq.(4.21) into $c_t(h(t))$, i.e., $c_{l-1}^*(t) = c_t(h_{l-1}^*(t))$. With the local minima in all sub-regions $(h_{l-1}(t), h_l(t)], \forall l \in \{2, \cdots, |\mathcal{H}(t)|\}$, the global minimum over the entire feasible region can be obtained as

\[
c^*(t) = \min_{l \in \{2, \cdots, |\mathcal{H}(t)|\}} c_{l-1}^*(t)
\]

with the minimizer as

\[
l^*(t) = \arg\min_{l \in \{2, \cdots, |\mathcal{H}(t)|\}} c_{l-1}^*(t).
\]

This yields the optimal headway as $h^*(t) = h_{l^*}(t)$ and $i_s^*(t) = \frac{h^*(t) \Delta_s(t)}{c}, \forall s \in S$. Finally, after solving the unit-time problems $\forall t \in [0, T]$, we can plug $c^*(t), \forall t \in [0, T]$ into Eq. (4.20) to obtain an approximate value of $OP(o)$.  

82
4.4.3 Discretization Method

The analytical solutions from the CA model cannot directly applied to schedule vehicles along the corridor. They have to be converted to discrete solutions. To this end, this subsection presents a discretization method to convert $h^*(t)$ into discrete time points for each dispatch $t_k^*$ and $i_s^*(t)$ into discrete vehicle formation $i_{ks}^*$. To discretize $h^*(t)$, we use the same method proposed in Chapter 3 due to its simplicity and computation efficiency. Note that this process also determines $K^*$ and $\mathcal{K}^*$.

Once the discrete time points $t_k^*, \forall k \in \mathcal{K}$ are obtained, we can move on to discretize $i_s^*(t), \forall s \in \mathcal{S}$. In Chapter 3, we computed the weighted average of all vehicle formations in time interval $[t_{k-1}^*, t_k^*]$ and rounded it to an integer as the vehicle formation for dispatch $k$. For MAV shuttle systems, the discretized results from this approach are not distant from the continuous values since discrete vehicle formations are used in solving the unit-time problems. However, in this chapter, discrete vehicle formations $\hat{i}_s(t), \forall s \in \mathcal{S}$ are approximated by a continuous function $i_s(t), \forall s \in \mathcal{S}$, whose original step function (i.e., $\hat{i}_s(t)$) cannot be well estimated by the weighted average method. Thus, here we propose a greedy heuristic to discretize $i_s^*(t), \forall s \in \mathcal{S}$ to improve the discretization accuracy. Note that as long as the dispatch times are determined, the vehicle formation of each station can be determined independently. Thus, the following analysis can be applied to all $s \in \mathcal{S}$ but we omit this index for the convenience of the notation. Propositions 4.4 and 4.5 indicate that for each dispatch $k$ at station $s$, the optimal vehicle formation should be selected between the lower bound vehicle formation $i_{ks}$ and upper bound vehicle formation $\bar{i}_{ks}$. For $k \in \mathcal{K}^* \setminus \{K^*\}$, if $\bar{a}_{ks} - i_{ks} c_w(t_k^* - t_{k-1}^*) > f_{\bar{i}_{ks}} - f_{i_{ks}}$, indicating that the increase in passenger waiting cost by using the lower bound vehicle formation cannot be offset by the decrease in the operational cost, we select the upper bound vehicle formation, i.e., $i_{ks}^* = \bar{i}_{ks}$. Otherwise, the
lower bound vehicle formation is used, i.e., $i^*_{ks} = i^*_{ks} = \bar{i}_{ks}$. For the last dispatch $K$, the upper bound should always be used since Constraints (4.6) will be violated otherwise (passengers cannot be all cleared at the end of the operational horizon if the lower-bound vehicles formation is used).

The above discretization methods can be formally stated as the pseudocode in Algorithm 4.1.

---

**Algorithm 4.1. Discretization**

**Input:** $h^*(t), i^*_s(t), \forall s \in S, t \in [0,T], h$

1. $K \leftarrow 0$
2. $t^K_\star \leftarrow T$
3. $\mathcal{K} \leftarrow \{K\}$
4. while $t^K_\star \geq h$
5. $K \leftarrow K + 1$
6. $\mathcal{K} \leftarrow \mathcal{K} \cup \{K\}$
7. $t^K_\star \leftarrow \text{argsup}_{t \in [0,t^K_{s-1}]} (h^*(t) = t^K_{s-1} - t)$
8. end while
9. $i^*_{ks} \leftarrow \left\lfloor \frac{\int^{t^K_\star}_{t^K_{s-1}} \bar{a}_s(t) dt}{c} \right\rfloor, \forall s \in S$
10. for $k$ in $\mathcal{K} \setminus \{K\}$
11. for $s$ in $S$
12. $\bar{a}_{ks} \leftarrow \int^{t^K_\star}_{t^K_{s-1}} \bar{a}_s(t) dt, i_{ks} \leftarrow \left\lfloor \frac{\bar{a}_{ks}}{c} \right\rfloor, \bar{i}_{ks} \leftarrow \left\lfloor \frac{\bar{a}_{ks}}{c} \right\rfloor$
13. if $(\bar{a}_{ks} - i_{ks} c) w(t^K_\star - t^K_{s-1}) > f_{ks} - f_{\bar{f}_{ks}}$
14. $i^*_{ks} \leftarrow \bar{i}_{ks}$
15. else
16. $i^*_{ks} \leftarrow i_{ks}$
17. end if
18. end for
19. end for
20. Reverse ($i^*_k, t^*_k) = (t^*_k, i^*_k), \forall k \in \mathcal{K} := \{0,1, \ldots, K\}$

**Output:** $t^*_k, i^*_k, \forall k \in \mathcal{K}, \forall s \in S$
4.5 Discrete Model

To propose a computationally tractable approach to the investigated problem, this section presents a discrete model based on the discrete time horizon. We first present a mixed integer linear programming formulation that can be solved by existing commercial solvers (e.g., Gurobi, CPLEX) in small instances. Then, since the computation time of the commercial solvers increases substantially as the instance size increases, we investigate the theoretical properties of the optimal solution(s) to the investigated problem, which offers a theoretical foundation for designing efficient solution algorithms. Finally, we propose an expedited branch and bound (B&B) algorithm to solve the problem for relatively large instances. The solution space of the proposed B&B algorithm is greatly reduced with the theoretical properties.

4.5.1 Model Formulation

This subsection formulates the investigated problem as a mixed integer linear programming model. The decision variables needed for the model formulation are as follows.

\( x_{ijs}, \forall i \in I, j \in J, s \in S \): equals 1 if a vehicle in formation \( i \) is dispatched at \( s \) at \( j \) and 0 otherwise.

\( b_{j's}, \forall j' \in J, j \in \{j', \cdots, J\}, s \in S \): equals 1 if passengers arriving at \( s \) at \( j' \) have all boarded by \( j \) and 0 otherwise.

\( r_{js}, \forall j \in J, s \in S \): number of passengers on the vehicle right before it departs from \( s \) at \( j \).

\( y_{j's}, \forall j' \in J, j \in \{j', \cdots, J\}, s \in S \): cumulative percentage of passengers arriving at \( s \) at \( j' \) that have boarded by \( j \).

Here the percentage of the passengers \( y_{j's} \) rather than the actual number of passengers is used to formulate the FIFS principal (see the following analysis for the detailed formulas). With these decision variables, we first formulate the constraints in the investigated problem as follows.
(i) Minimum headway requirement. Constraints (4.22) ensure that at most one vehicle can be dispatched within the minimum dispatch headway at the start station due to the safety considerations. Constraints (4.23) require that the dispatch headway at all stations should be equal to that at the start station; the headway requirement is imposed at all stations along the corridor.

\[
\sum_{i \in I} \sum_{j' = j-1}^{j} x_{ij'1} \leq 1, \forall j - 1 \leq j \leq J. \tag{4.22}
\]

\[
\sum_{i \in I} x_{ij1} = \sum_{i \in I} x_{ij1}, \forall j \in J, s \in S \setminus \{1\}. \tag{4.23}
\]

(ii) Onboard passenger conservation. These constraints formulate the number of passengers in a vehicle as it travels from the first to the last stations. Constraints (4.24) define the number of passengers on the vehicle at the start station. Constraints (4.25) indicate that the number of passengers on the vehicle at station \( s \in S \setminus \{1\} \) at time \( j \) equals the number of passengers on the vehicle at \( s - 1 \) at \( j \) plus the number of passengers who board the vehicle at \( s \) at \( j \) minus the number of passengers who alight at \( s \) at \( j \). Finally, Constraints (4.26) require that the number of passengers on the vehicle should always be no greater than the capacity of the vehicle dispatched at \( s \) at any time \( j \).

\[
r_{j1} = \sum_{j' = 0}^{j} a_{j'1}(y_{j'j1} - y_{j'(j-1)1}), \forall j \in J. \tag{4.24}
\]

\[
r_{js} = r_{j(s-1)} + \sum_{j' = 0}^{j} a_{j's}(y_{j'js} - y_{j'(j-1)s}) - \sum_{j' = 0}^{j} \sum_{u \in \delta_S} a_{j'u}p_{j'uS}(y_{j'u} - y_{j'(j-1)u}), \forall j, s \in S \setminus \{1\}. \tag{4.25}
\]

\[
r_{js} \leq c \sum_{i \in I} i x_{ij}, \forall j \in J, s \in S. \tag{4.26}
\]
(iii) Passenger boarding order. These constraints formulate the FIFS rules. Here we define $y_{j's}$ as the cumulative proportion of passengers who arrive at $s$ at $j'$ and board the vehicle at $j$, so passengers are treated as groups corresponding to different arrival times. Further, $b_{j's}$ is defined to indicate whether all passengers arriving at $s$ at $j'$ have boarded by $j$ or not. With this, Constraints (4.27) define that the cumulative percentage of boarding passengers for each arrival timestamp $j'$ is non-decreasing. Constraints (4.28) define variables $b_{j's}$, which indicate that $b_{j's} = 0$ if the cumulative percentage of passengers arriving at station $s$ at $j'$ is less than 1. Constraints (4.29) indicate that passengers arriving at $s$ at $j'$ cannot board if those arriving at previous timestamp $j' - 1$ have not all boarded yet (i.e., $y_{j's} = 0$ must hold if $b_{(j' - 1)s} = 0$). Finally, Constraints (4.30) require that if a vehicle is dispatched at station $s$ at time $j$, all passengers waiting at $s$ will board the vehicle unless there is no capacity. That is, if $b_{sjj} = 0$, then $c \sum_{i \in J} x_{ijs} \leq r_{js}$. Then with Constraints (4.26) we obtain that $r_{js} = c \sum_{i \in J} x_{ijs}$ in this case. If $b_{sjj} = 0$, Constraints (4.30) are not binding.

$$y_{j'(j-1)s} \leq y_{j's}, \forall j' \in J \backslash \{0\}, j \in \{j', \ldots, J\}, s \in S.$$  \hfill (4.27)

$$b_{j's} \leq y_{j's}, \forall j \in J, j' \in \{0, \ldots, j\}, s \in S.$$  \hfill (4.28)

$$y_{j's} \leq b_{(j' - 1)s}, \forall j \in J \backslash \{0\}, j' \in \{1, \ldots, j\}, s \in S.$$  \hfill (4.29)

$$c \sum_{i \in J} i x_{ijs} - r_{js} \leq lc b_{jjs}, \forall j \in J, s \in S.$$  \hfill (4.30)

(iv) Service quality. Constraints (4.31) are imposed to make sure that all passengers arriving at each station along the corridor are served at the end of the operational horizon.

$$y_{j's} = 1, \forall j' \in J, s \in S.$$  \hfill (4.31)

(v) Variable domain. These constraints define the feasible regions for all the decision variables.
\[ x_{ij} \in \{0,1\}, \forall i,j \in J, s \in S. \]  
\[ b_{j'js} \in \{0,1\}, \forall j' \in J, j \in \{j', \ldots, J\}, s \in S. \]  
\[ y_{j'js} \in [0,1], \forall j' \in J, j \in \{j', \ldots, J\}, s \in S. \]  
\[ r_{js} \in [0, l_c], \forall j \in J, s \in S. \]  

Next, we formulate the objective function. Following the approach in Chapter 3, we measure the passenger waiting time as the area between the cumulative passenger arrival curves and cumulative passenger departure curves as

\[ \delta \sum_{j'=0}^{j-1} \left( a_{j's} \left( 1 - y_{j'(j-1)s} \right) \right) + \frac{a_{js}}{2}. \]

Note that the smaller the discretization interval \( \delta \), the higher the approximation accuracy. Further, the total operational cost can be formulated as \( \sum_{s \in S} \sum_{j \in J} \sum_{i \in \mathcal{I}} x_{ij}s f_i \). Therefore, the investigated problem can be formulated as a mixed integer linear programming model as

\[
\min \left[ \sum_{s \in S} \sum_{j \in J} \sum_{i \in \mathcal{I}} x_{ij} s f_i + w \delta \sum_{j \in J \setminus \{0\}} \left( \sum_{j'=0}^{j-1} \left( a_{j's} \left( 1 - y_{j'(j-1)s} \right) \right) + \frac{a_{js}}{2} \right) \right]
\]

subject to Constraints (4.22) – (4.35).

4.5.2 Solution Space Reduction Properties

To prepare for the development of an efficient customized solution algorithm, this subsection analyzes properties that can reduce the solution space of the investigated problem. We introduce some definitions for preparations.

(i) Passenger queue: The passenger queue at station \( s \) at time \( j \), \( q_{js} \), is defined as the number of passengers waiting for boarding at \( s \) at \( j \), i.e.,
\[ q_{js} := \sum_{j' = 0}^{j} a_{j's}(1 - y_{j'js}), \forall j, s \in S. \quad (4.37) \]

(ii) Cross-sectional passenger queue: The cross-sectional passenger queue at station \( s \) at time \( j \), \( q_{js} \), is defined as the number of passengers who are waiting for boarding at stations in the upstream of \( s \) (inclusive) and intend to travel to destination stations in the downstream of \( s \), i.e.,
\[ q_{js} := \sum_{u \in S_{s+1}} \sum_{v \in S_s^+} \sum_{j' = 0}^{j} (a_{j'u} p_{j'uv}(1 - y_{j'ju})), \forall j \in J, s \in S. \quad (4.38) \]

(iii) Actual seat demand: The actual seat demand at station \( s \) at time \( j \), \( a_{js} \), is the number of passengers waiting for boarding at \( s \) and those who board at \( s \)'s upstream stations but have not alighted at \( s \) at \( j \), i.e.,
\[ a_{js} := q_{(j-1)s} + a_{js} + \sum_{u \in S_{s}^-} \sum_{v \in S_{s}^+} \sum_{j' = 0}^{j} (a_{j'u} p_{j'uv}(y_{j'ju} - y_{j'(j-1)u})), \forall j \in J, s \in S. \quad (4.39) \]

(iv) Upper-bound vehicle formation: The upper-bound vehicle formation at station \( s \) at time \( j \), \( \bar{i}_{js} \), is the shortest vehicle formation whose capacity is no smaller than the seat demand at \( s \) at \( j \), i.e.,
\[ \bar{i}_{js} := \min \left\{ \left\lceil \frac{a_{js}}{c} \right\rceil, I \right\}, \forall j \in J, s \in S. \quad (4.40) \]

(v) Lower-bound vehicle formation: The lower-bound vehicle formation at station \( s \) at time \( j \), \( \bar{l}_{js} \), is the longest vehicle formation whose capacity is no greater than the seat demand at \( s \) at \( j \), i.e.,
\[ \bar{l}_{js} := \min \left\{ \left\lfloor \frac{a_{js}}{c} \right\rfloor, l \right\}, \forall j \in J, s \in S. \quad (4.41) \]
With these definitions, we derive the discrete counterparts of Propositions 4.6 and 4.7 as follows.

**Proposition 4.6.** In an optimal solution \( \{x_{ij}, b_{j's}, y'_{j's}, r_{js}\} \) to problem (4.22) \(-\) (4.36), if \( \sum_{i \in J} x_{ij} = 1 \), then \( x_{ij} = 0, \forall i > i_{js}, j \in J, s \in S \).

**Proposition 4.7.** In an optimal solution \( \{x_{ij}, b_{j's}, y'_{j's}, r_{js}\} \) to problem (4.22) \(-\) (4.36), if \( \sum_{i \in J} x_{ij} = 1 \) and the following conditions hold: (i) \( S(f_2 - f_1) < cw_j \delta \); and (ii) \( a_{js} \leq 1c, \forall s \in S \), then \( x_{ij} = 0, \forall i < i_{js}, j \in J, s \in S \).

While the above two propositions focus on the optimal vehicle formations, in the following proposition we analyze the relationship between the vehicle dispatch headway and passenger queue in the optimal solution.

**Proposition 4.8.** A feasible solution \( \{x_{ij}, b_{j's}, y'_{j's}, r_{js}\} \) to problem (4.22) \(-\) (4.36) is not optimal if there exists \( \tilde{j} < j \in J \) such that: (i) \( \sum_{i \in J} x_{i\tilde{j}} = 1, \sum_{i \in J} x_{i1} = 1, \sum_{j'' \in (j, j)} \sum_{i \in J} x_{i'j''1} = 0 \); (ii) \( j - \tilde{j} > j \); (iii) \( q_{(j-1)s} \geq c \sum_{i \in J} i x_{ij} - \sum_{j' = 0}^{j} \sum_{u \in S_t} \sum_{v \in S_t^+} \left( a_{j'u} p_{j'uv}(y'_{j'u} - y'_{(j-1)u}) \right), \forall s \in S \).

**Corollary 4.1.** In an optimal solution \( \{x_{ij}, b_{j's}, y'_{j's}, r_{js}\} \) to problem (4.22) \(-\) (4.36), if there exists \( \tilde{j} < j \in J \) such that: (i) \( \sum_{i \in J} x_{i\tilde{j}} = 1, \sum_{i \in J} x_{i1} = 1, \sum_{j'' \in (j, j)} \sum_{i \in J} x_{i'j''1} = 0 \); (ii) \( q_{(j-1)s} \geq c \sum_{i \in J} i x_{ij} - \sum_{j' = 0}^{j} \sum_{u \in S_t} \sum_{v \in S_t^+} \left( a_{j'u} p_{j'uv}(y'_{j'u} - y'_{(j-1)u}) \right), \forall s \in S \), then \( j - \tilde{j} = j \).

**Corollary 4.2.** In an optimal solution \( \{x_{ij}, b_{j's}, y'_{j's}, r_{js}\} \) to problem (4.22) \(-\) (4.36), if there exists \( \tilde{j} < j \in J \) such that: (i) \( \sum_{i \in J} x_{i\tilde{j}} = 1, \sum_{i \in J} x_{i1} = 1, \sum_{j'' \in (j, j)} \sum_{i \in J} x_{i'j''1} = 0 \); (ii) \( j - \tilde{j} > j \), then there must exist \( \tilde{s} \in S \) satisfying \( q_{(j-1)s} < c \sum_{i \in J} i x_{ij} - \sum_{j' = 0}^{j} \sum_{u \in S_t} \sum_{v \in S_t^+} \left( a_{j'u} p_{j'uv}(y'_{j'u} - y'_{(j-1)u}) \right), \forall s \in S \).
Proposition 4.8 reveals that if the passenger queue at the time point right before a dispatch is greater than the capacity of the dispatched MAV at all stations along the corridor, then dispatching this MAV earlier will always lead to a better solution. Further, the resultant corollaries show more properties of the optimal solution based on this property. Corollary 4.1 shows that in an optimal solution, if the passenger queue at the time point right before a dispatch is greater than the capacity of the dispatched MAV at all stations along the corridor, then the headway between this and the previous dispatch equals the minimum dispatch headway. Corollary 4.2 shows that in an optimal solution, if the headway between a dispatch and the previous dispatch is larger than the minimum dispatch headway, then there must be a station where the passenger queue is greater than the capacity of the dispatched MAV at this station. Any feasible solution not satisfying these two corollaries can be deleted from the solution space.

Finally, we investigate a dominance rule between two feasible solutions in Proposition 4.9. This identifies an approach to determining if a feasible solution can form an optimal solution before it is completely constructed given another feasible solution. To formulate this proposition, we define the bounded number of elapsed time points from the previous dispatch at \( j \) as \( h_j := \min \{ j - j', j \} \) such that \( j' := \arg\max_{j'' \in \{0, \ldots, j\}} \{ \sum_{i \in I} x_{ij''} = 1 \} \), which represents the minimum between the number of time points elapsed from the previous dispatch to \( j \) and the minimum dispatch headway. Note that when \( h_j < j \) no MAVs can be dispatched (Constraints (4.22) and (4.23)). Further, given a solution \( o := \{ x_{js}, b_{j's}, y_{j's}, r_{js} \} \) to problem (4.22) \(-\) (4.36), we define \( z_j(o) \) as its objective value computed up till \( j \), i.e.,

\[
z_j(o) := \sum_{s \in S} \left[ \sum_{j'' \in \{0, \ldots, j\}} \left( \sum_{i \in I} (x_{ij''} s) f_i + w\theta \sum_{j''' \in \{ 1, \ldots, j \}} \left( \sum_{j'' = 0}^{j'''} \left( a_{j's} \left( 1 - y_{j'(j''-1)s} \right) + \frac{a_{j''s}}{2} \right) \right) \right) \right], \forall j.
\]
Proposition 4.9. Given two feasible solutions \( o := \{x_{ij}, b_{js}, y_{js}, r_{js}\} \) and \( \hat{o} := \{\hat{x}_{ij}, \hat{b}_{js}, \hat{y}_{js}, \hat{r}_{js}\} \) (all variables relevant to \( \hat{o} \) will be denoted with a hat accent in the following analysis) to problem (4.22) \(-\) (4.36), we say \( o \) dominates \( \hat{o} \) if there exists \( j \in J \) such that: (i) \( h_j \geq \hat{h}_j \); (ii) \( \bar{a}_{js} \leq \hat{a}_{js}, \forall s \in S \); and (iii) \( z_j(o) < z_j(\hat{o}) \).

4.5.3 Branch and Bound Algorithm

In this subsection, we present an expedited branch and bound (B&B) algorithm to solve (4.22) \(-\) (4.36) more efficiently. Different from standard B&B algorithms, the computation speed of this customized B&B algorithm is greatly enhanced by the state reduction properties discovered in the previous subsection.

4.5.3.1 Algorithm Overview

The framework of the expedited B&B algorithm is shown in Figure 4.5. The process of implementing a B&B algorithm is essentially to construct a tree with a set of nodes indexed as \( n \in \mathcal{N} := \{1, \ldots, N\} \), where \( N \) is the number of nodes in the tree. All nodes are categorized into \( L \) levels based on their depths in the tree, resulting in a set of levels indexed as \( l \in \mathcal{L} := \{1, \ldots, L\} \) and a set of nodes in each level denoted as \( \mathcal{N}_l := \{1, \ldots N_l\}, \forall l \in \mathcal{L} \) where \( N_l \) is the number of nodes at level \( l \). Note that \( \bigcup_{l \in \mathcal{L}} \mathcal{N}_l = \mathcal{N} \). Each node in the B&B tree is indexed as a 2-tuple \((l, n)\) associated with information as follows: parent node index \((p_{ln})\), station index \((s_{ln})\), time index \((j_{ln})\), formation of the MAV dispatched at the parent node \((i_{ln})\), the number of elapsed time points from the previous dispatch \((h_j)\), a vector containing the cumulative percentage of passengers that have boarded at all stations \((\vec{y}_{ln} := \{y_{js}, \forall j' \in [0, l_{ln}], j \in [j', j_{ln}], s \in S\})\), objective value \((c_{ln})\), lower bound \((l_{ln})\), and upper bound \((u_{ln})\). This algorithm constructs the B&B tree with a depth-first search strategy. To this end, all the existing node indexes are stored in two different pools, including a local pool \( \mathcal{P} \) where all child nodes of the current parent node are stored, and a global pool \( \mathcal{Q} \) that contains all
other unfathomed nodes in the tree. For each node \((l, n)\), we compute its lower bound \((b_{ln})\) and upper bound \((\bar{b}_{ln})\) analytically. The global lower bound, denoted as \(\bar{B}\), and the global upper bound, denoted as \(\bar{\bar{B}}\), are defined as \(\bar{B} := \min_{(l,n) \in Q} b_{ln}\) and \(\bar{\bar{B}} := \min_{(l,n) \in Q} \bar{b}_{ln}\), respectively. In the following analysis, let \(l_0, n_0, l_1, n_1\) be the level index of the parent node, node index of the parent node, level index of the child node, and node index of the child node, respectively.

![Figure 4.5: Framework of the expedited B&B algorithm.](image-url)
With the above settings, the algorithm starts from the root node \((0,0)\) where \(p_{00} = -1, s_{00} = S, f_{00} = -1, i_{00} = 0, h_{00} = j, c_{00} = 0, y_{00} = \{y_{s(-1)(-1)} = 0, \forall s \in S\}, b_{00} = 0, b_{00} = \infty. \) Further, we set \(l_0 = 0, n_0 = 0\) (i.e., we set the root node as the next node for branching), \(P = \emptyset, Q = \emptyset, B = \infty, B = 0.\) After the initialization process, the algorithm solves the model by recursively calling the following steps until the convergence criterion is satisfied. First, the level index of the child node is computed as \(l_1 = l_0 + 1\) while the station and time indexes for the child nodes of \((l_0, n_0)\), denoted as \(s_{l_1}(l_0, n_0)\) and \(j_{l_1}(l_0, n_0)\) respectively, are updated as follows:

\[
s_{l_1}(l_0, n_0) := \begin{cases} s_{l_0 n_0} + 1 & \text{if } s_{l_0 n_0} + 1 \leq S \\ 1 & \text{if } s_{l_0 n_0} + 1 > S \end{cases}
\]

(4.42)

\[
j_{l_1}(l_0, n_0) := \begin{cases} j_{l_0 n_0} + 1 & \text{if } s_{l_0 n_0} + 1 \leq S \\ j_{l_0 n_0} + 1 & \text{if } s_{l_0 n_0} + 1 > S \end{cases}
\]

(4.43)

With the updated time-space information, we compute the actual seat demand for the child nodes of \((l_0, n_0)\), denoted as \(\tilde{a}_{l_1}(l_0, n_0)\), using Eq.(4.39), where \(j\) and \(s\) are replaced by \(j_{l_1}(l_0, n_0)\) and \(s_{l_1}(l_0, n_0)\), respectively. Based on Propositions 4.6 and 4.7, Eqs.(4.22) and (4.23), the upper-bound and lower-bound vehicle formations for the child nodes of \((l_0, n_0)\) in the optimal solution(s), denoted as \(i_{l_1}(l_0, n_0)\) and \(\tilde{i}_{l_1}(l_0, n_0)\) respectively, can be computed as follows

\[
\tilde{i}_{l_1}(l_0, n_0) = \min \left\{ \left\lceil \frac{\tilde{a}_{l_1}(l_0, n_0)}{c} \right\rceil, l \right\}.
\]

(4.44)

\[
i_{l_1}(l_0, n_0)
\]

(4.45)

\[
= \begin{cases} \min \left\{ \left\lceil \frac{\tilde{a}_{l_1}(l_0, n_0)}{c} \right\rceil, l \right\} & \text{if } S(f_2 - f_1) < c w j \delta \text{ and } \tilde{a}_{l_1}(l_0, n_0) \leq lc, \forall s \in S \\text{otherwise} \\ 1 & \end{cases}
\]

With these properties and the feasibility conditions in the model, we next solve the set of candidate optimal vehicle formations at node \((l_0, n_0)\). Specifically, Constraints (4.31) require that
all passengers must be cleared at $J$. Adopting vehicle formations smaller than the upper bound cannot serve all passengers (the definition of upper-bound vehicle formation) while using those greater than the upper bound will unnecessarily increase the vehicle operational cost (Proposition 4.7). Therefore, the upper-bound vehicle formation must be dispatched at $J$. Since MAVs must be dispatched at $J$, then Constraints (4.22) – (4.23) indicate MAVs cannot be dispatched from $J - j + 1$ to $J - 1$. At other time points (i.e., $j_{l_0n_0} \leq J - j$), the candidate optimal vehicle formations can be determined in three cases. First, if $h_{l_0n_0} < j - j$, which indicates that the number of elapsed time points from the previous dispatch is smaller than the minimum dispatch headway, and thus no MAVs can be dispatched (otherwise Constraints (4.22) – (4.23) will be violated). Second, if $h_{l_0n_0} = j - j$ and $s_{l_0n_0} = S$, MAVs could be dispatched, and therefore the set of candidate optimal vehicle formations is $\{0, \overline{i}_1(l_0, n_0), \overline{i}_1(l_0, n_0) \}$, where 0 represents that no MAVs are dispatched. Third, if $h_{l_0n_0} = j - j$ and $s_{l_0n_0} < S$, the vehicle formation decisions at the current parent node depends on those at the previous station (i.e., $i_{l_0n_0}$) as required by Constraints (4.23). If no MAVs are dispatched at the previous station (i.e., $i_{l_0n_0} = 0$), then MAVs are not allowed to be dispatched at the current station either. In contrast, if MAVs are dispatched at the previous station (i.e., $i_{l_0n_0} > 0$), then a MAV must be dispatched at the current station. To sum up, the set of candidate optimal vehicle formations at parent node $(l_0, n_0)$ is formulated as

$$j_{l_0n_0}(j_{l_0n_0}, s_{l_0n_0}, i_{l_0n_0}, h_{l_0n_0})$$

$$= \begin{cases} 
\overline{i}_1(l_0, n_0) & \text{if } j_{l_0n_0} = J \\
0 & \text{if } J - j < j_{l_0n_0} \leq J - 1 \text{ or } j_{l_0n_0} \leq J - j, h_{l_0n_0} < j \text{ or } j_{l_0n_0} \leq J - j, s_{l_0n_0} < S, i_{l_0n_0} = 0 \\
\{0, \overline{i}_1(l_0, n_0), \overline{i}_1(l_0, n_0)\} & \text{if } j_{l_0n_0} \leq J - j, s_{l_0n_0} = S, h_{l_0n_0} = j \\
\{\overline{i}_1(l_0, n_0), \overline{i}_1(l_0, n_0)\} & \text{if } j_{l_0n_0} \leq J - j, s_{l_0n_0} < S, i_{l_0n_0} > 0 
\end{cases}$$

$\forall n_0 \in \mathcal{N}_{l_0}, l_0 \in \mathcal{L}$. (4.46)
Note that following standard B&B algorithms, the number of branches at each node where vehicles can be dispatched is \((I + 1)\). By imposing the theoretical bounds to the optimal vehicle formations (though not exhaustive), the proposed algorithm restricts the number of branches at each node to at most three, which greatly reduces the solution space. Next, we enumerate \(i \in \mathcal{J}_{l_0n_0}(j_{l_0n_0}, s_{l_0n_0}, i_{l_0n_0}, h_{l_0n_0})\) and create a child node \((l_1, n_1 := |\mathcal{N}_l| + 1)\) for each \(i\). The information of each resultant child node is updated with the following state transition functions.

\[
\begin{align*}
p_{l_1n_1} &= (l_0, n_0). \\
s_{l_1n_1} &= s_{l_1}(l_0, n_0). \\
j_{l_1n_1} &= j_{l_1}(l_0, n_0). \\
i_{l_1n_1} &= i.
\end{align*}
\]

\[
h_{l_1n_1} = \begin{cases} 
\min\{h_{l_0n_0} + 1, j\} & \text{if } s_{l_0n_0} = S, i = 0 \\
1 & \text{if } s_{l_0n_0} = S, i > 0 \\
h_{l_0n_0} & \text{if } s_{l_0n_0} < S
\end{cases}
\]

\[
y_{j'j_{l_1n_1}s_{l_1n_1}} = \begin{cases} 
\frac{B_{l_1n_1j''} - C_{l_1n_1}}{a_{j's_{l_1n_1}}} & \text{if } j' < j_{l_1n_1} \\
0 & \text{if } j' = j_{l_1n_1}, \forall j' \in \{0, \cdots, j_{l_1n_1}\}.
\end{cases}
\]

where

\[
B_{l_1n_1j''} := \sum_{v \in \mathcal{S}_{l_1n_1}} \sum_{j' = 0}^{j''} \left( a_{j's_{l_1n_1}} p_{j's_{l_1n_1}} v \left(1 - y_{j'(j_{l_1n_1} - 1)u}\right)\right), \forall j'' \in \{0, \cdots, j_{l_1n_1}\}
\]

represents the number of passengers arriving at \(s_{l_1n_1}\) before \(j''\) (inclusive) and still waiting at \(j_{l_1n_1}\)

\[
C_{l_1n_1} := \min \left\{ i_{l_1n_1} c - \sum_{u \in \mathcal{S}_{l_1n_1}} \sum_{v \in \mathcal{S}_{l_1n_1}} \sum_{j' = 0}^{j_{l_1n_1}} \left( a_{j'u} p_{j'u} v \left( y_{j'(j_{l_1n_1} - 1)u} - y_{j'(j_{l_1n_1} - 1)u}\right)\right) \right\}
\]

represents the capacity of the MAV after passengers alighting at \(s_{l_1n_1}\), and
\[ j_{l_1 n_1} := \arg\max_{j'' \in \{0, \ldots, j_{l_1 n_1}\}} \{ B_{l_1 n_1 j''} \leq C_{l_1 n_1} \} \]

represents the time point of the last passenger boarding the vehicle before all space is occupied.

With this information, next we apply theoretical properties in Propositions 4.8, 4.9 as well as bounding techniques to fathom nodes that certainly do not constitute optimal solutions. Specifically, for the newly generated node \((l_1, n_1)\), we first check if the conditions in Proposition 4.8 are satisfied. If yes, this node cannot constitute an optimal solution. Otherwise, we iterate over all nodes \((l', n') \in Q\), check whether \((l', n')\) is dominated by \((l_1, n_1)\) using conditions in Proposition 4.9, and remove it from \(Q\) if it is dominated. Then, we iterate over all remaining nodes \((l', n') \in Q\) again and check whether \((l_1, n_1)\) is dominated by any \((l', n') \in Q\), applying conditions in Proposition 4.9. \((l_1, n_1)\) cannot form an optimal solution if it is dominated. Further, we compute the lower and upper bounds of \((l_1, n_1)\) and compare the lower bound with \(\bar{B}\). If the lower bound is greater than \(\bar{B}\), this node cannot constitute an optimal solution and thus can be fathomed. After the aforementioned check, we add \((l_1, n_1)\) to \(P\) if it may still form an optimal solution. Different from standard B&B algorithms that simply adopt the bounding techniques to fathom nodes, the proposed algorithm takes advantages of the theoretical properties of the optimal solution to the investigated problem. These properties further cut down the solution space of the investigated problem that will grow extremely quickly without them.

When the enumeration is completed, the next step is to select a new parent node for branching and to update the global lower and upper bounds. We first check if \(j_{l_1}(l_0, n_0) = j\) and \(s_{l_1}(l_0, n_0) = S\) or \(P = \emptyset\) are satisfied. If yes, we set the node with the lowest lower bound (i.e., the global lower bound) in the global selection pool as the next parent node. Otherwise, the node with the smallest lower bound in the local selection pool should be selected, i.e.,
\[(l_0, n_0) = \begin{cases} \arg\min_{(l, n) \in Q} \{b_{jn}\}, & \text{if } j_{l_1}(l_0, n_0) = J \text{ and } s_{l_1}(l_0, n_0) = S \text{ or } \mathcal{P} = \emptyset \\ \arg\min_{(l, n) \in \mathcal{P}} \{b_{jn}\}, & \text{otherwise} \end{cases}.\]

The unfathomed nodes in the local selection pool are then added to the global selection pool and the local pool is reinitialized as an empty set for the next branching. With the updated global pool, the global lower and upper bounds are updated based on their definitions. Afterwards, the algorithm moves on to the next branching procedure with the new \((l_0, n_0)\). This algorithm terminates when either the optimality gap (i.e., the relative difference between \(\overline{B}\) and \(\underline{B}\) with \(\overline{B}\) as the benchmark) is less than a predefined threshold or the computation time reaches a given limit.

**4.5.3.2 Lower Bound**

The lower bound of each node \((l, n)\) in the B&B tree is calculated analytically. With the objective value at this node (i.e., \(c_{ln}\)), we first rewrite objective function (4.36) as follows.

\[
\sum_{s \in \mathcal{S}} \left[ \sum_{j \in J} \sum_{i \in I} x_{ijs} f_i + w \delta \left( \sum_{j \in J \setminus \{0\}} \left( \sum_{j' = 0}^{j-1} (a_{j's}(1 - y_{j'(j-1)s}) + \frac{a_{js}}{2}) \right) \right) \right] = c_{ln} + f_{ln}(x_{ijs}) + g_{ln}(y_{j'js}),
\]

where

\[
f_{ln}(x_{ijs}) := \sum_{s \in \mathcal{S}_{ln}^+} \sum_{j \in J} x_{ijs} f_i + \sum_{s \in \mathcal{S}} \sum_{j \in \{j_{ln} + 1, \ldots, J\}} \sum_{i \in I} x_{ijs} f_i
\]

is the operational cost of the system after this node (we call it the future operational cost) and

\[
g_{ln}(y_{j'js}) = w \delta \left\{ \sum_{s \in \mathcal{S}_{ln}^+} \left[ \left( \sum_{j' = 0}^{j-1} (a_{j's}(1 - y_{j'(j-1)s}) + \frac{a_{js}}{2}) \right) \right] \right\} + \sum_{s \in \mathcal{S}} \left( \sum_{j \in \{j_{ln} + 1, \ldots, J\}} \left( \sum_{j' = 0}^{j-1} (a_{j's}(1 - y_{j'(j-1)s}) + \frac{a_{js}}{2}) \right) \right).
\]
denotes the passenger waiting cost of the system after this node (we call this the future passenger waiting cost in the following analysis). Since
\[
\min\{f_i(x_{ij}) + g_i(y'_{js})\} \geq \min\{f_i(x_{ij})\} + \min\{g_i(y'_{js})\},
\]
to find a lower bound to the objective value at \((l, n)\), i.e., \(\min\{f_i(x_{ij}) + g_i(y'_{js})\}\), we can just solve \(\min\{f_i(x_{ij})\}\) and \(\min\{g_i(y'_{js})\}\), respectively. The following propositions offer analytical solutions to the lower bound to the future operational cost and to the future passenger waiting cost, respectively.

**Proposition 4.10.** For a node \((l, n), l \in \mathcal{L}, n \in \mathcal{N}_l\) in the B&B tree with cross-sectional passenger queue \(\bar{q}_{jns}, \forall s \in \mathcal{S}\), passenger arrival demand \(a_{js}, \forall j \in \mathcal{J}, s \in \mathcal{S}\), passenger demand distribution \(p_{jus}, \forall j \in \mathcal{J}, u \in \mathcal{S}, v \in \mathcal{S}_{u}^+\), and cumulative percentage of boarding passengers \(y'_{js}, \forall j' \in \mathcal{J}, j \in \{j', \cdots, J\}, s \in \mathcal{S}\), a lower bound to the future operational cost is
\[
\min\{f_i(x_{ij})\} = \sum_{s \in \mathcal{S}} \left\{ \left[ Q_{ins} := \bar{q}_{jns} + \sum_{u \in \mathcal{S}_{s+1}} \sum_{v \in \mathcal{S}_s^+} \sum_{j' = j_{in}} \left( a_{j'u} p_{j'u} (1 - y'_{j'u}) \right) \right] f_i \right\}.
\]

**Proposition 4.11.** For a node \((l, n), l \in \mathcal{L}, n \in \mathcal{N}_l\) in the B&B tree with time index \(j_{in}\), bounded number of time points elapsed from the previous dispatch \(h_{in}\), passenger queue \(q_{jns}, \forall s \in \mathcal{S}\), and passenger arrival demand \(a_{js}, \forall j \in \mathcal{J}, s \in \mathcal{S}\), we construct an ordered set \(\mathcal{E}_{ins} := \{q_{jns}\} \cup \{a_{js}\}_{j \in \{j_{in}+1, \cdots, J\}}\), \(\forall s \in \mathcal{S}\), where the elements are sorted in an ascending order. Then a lower bound to the future passenger waiting cost is
\[
\min\{g_i(y'_{js})\} = w_d \sum_{s \in \mathcal{S}} \sum_{e \in \{1, \cdots, e_{in}\}} \mathcal{E}_{ins},
\]
where $e$ denotes the $e$-th element in $\mathcal{E}_{\text{lns}}$ and $e_{\text{ln}} := \left\lfloor \frac{j - [l_{\text{ln}} + (j - h_{\text{ln}}) + 1]}{l} \right\rfloor (j - 1)$.

Proposition 4.10 indicates that the lower-bound future operational cost at node $(l, n)$ equals the operational cost obtained if we consolidate the MPs that are needed to accommodate all remaining passengers into the longest MAV formation. This result indeed reflects the economics of scale of the operational cost. Proposition 4.11 indicates that the lower-bound future passenger waiting cost at node $(l, n)$ equals the passenger waiting cost arising from the minimum number of passengers waiting for one time interval if the minimum dispatch headway is adopted between every two consecutive dispatches. With these two propositions, we can compute the lower bound to the objective value at node $(l, n)$ as $b_{ln} = c_{ln} + \min\{f_{ln}(x_{ils})\} + \min\{g_{ln}(y_{js})\}$, which can be done in almost no time.

4.5.3.3 Upper Bound

The upper bound of each node $(l, n)$ in the B&B tree is calculated by finding a feasible solution to problem (4.22) ~ (4.36), which can be achieved with various heuristics. Among these approaches, the classical continuum approximation (CA) approach has been shown to be able to produce highly accurate near-optimal solution to such problems very efficiently. Thus, here we use the CA model presented in Section 4.4 to compute the upper bound. However, different from Section 4.4 where only unsaturated traffic is considered, the travel demand at a node $(l, n)$ can be oversaturated in the investigated problem. Thus, we need to revise the demand curve into one only with unsaturated traffic so that the “local-impact” property required by the CA approach can be satisfied.

To this end, we first solve the problem for $s \in S_{s_{ln}}^+$ if $S_{s_{ln}}^+ \neq \emptyset$. If $i_{ln} = 0$, based on Constraints (4.23) then no MAVs can be dispatched $\forall s \in S_{s_{ln}}^+$; i.e., $x_{ils} = 0, \forall i \in J, s \in S_{s_{ln}}^+$. If
$i_n > 0$, then dispatching the upper-bound vehicle formation suffices to provide a feasible solution. Thus, we compute the upper-bound vehicle formation at $s$ at $j_{ln}$, i.e. $\bar{j}_{lns}$, with Eq. (4.40). Then set $x_{\bar{j}_{lns}lns} = 1, \forall s \in S_{ln}^+$, and $x_{ij_{lns}} = 0, \forall i \in J \setminus \{\bar{j}_{lns}\}, s \in S_{ln}^+$. With these decisions, we solve $y'_{j_{lns}}, \forall j' \in J$ using Eq. (4.52) (with $j_{tn_1}, s_{tn_1}, i_{tn_1}$ replaced by $j_{ln}, s, \sum_{i \in J} x_{ij_{lns}}$, respectively). Then, we compute the resulting system costs and add it to the upper bound, i.e.,

$$\bar{b}_{ln} = \sum_{s \in S_{ln}^+} \left[ w \delta \left( \sum_{j' = 0}^{j-1} \left( a'_{j's} \left( 1 - y'_{j'(ln-1)s} \right) \right) + \frac{a_{j_{lns}}}{2} \right) + \sum_{i \in J} f_i x_{ij_{lns}} \right].$$

Next, we revise the passenger arrival demand $a_{js}, \forall j \in \{j_{ln} + 1, \ldots, J\}, s \in S$ to obtain a problem with only unsaturated arrival demand denoted as $a'_{js}, \forall j \in \{j_{ln} + 1, \ldots, J\}, s \in S$. The passenger distribution $p_{jsv}, \forall j \in \{j_{ln} + 1, \ldots, J\}, s \in S, v \in S_v^+$ also need to be revised as $p'_{jsv}, \forall j \in \{j_{ln} + 1, \ldots, J\}, s \in S, v \in S_v^+$. Following Chapter 3, we define the oversaturation periods as the set of time points when the actual boarding demand cannot be all accommodated even if the maximum service rate (i.e., the longest vehicle formation with the minimum dispatch headway $lc/j$) is adopted. Here the actual boarding demand at a station rather than the passengers arriving at this station is used, because the former represents the actual number of passengers who would like to board the vehicle in corridor systems. With this definition, we revise the passenger arrival demand with the following steps.

First, set the station index $s = 0$.

Second, set $s = s + 1$. Set the first time point of the current iteration $j_0 = j_{ln}$.

Third, find the first time point after $j_0$ with oversaturated actual boarding demand, i.e.,

$$j_1 := \arg\min_{j \in \{j_0, \ldots, J - 1\}} \left\{ \sum_{j' = j}^{j + J} \sum_{u \in S_{s+1}^s} \sum_{v \in S_v^s} \left( a'_{j'u} p'_{j'u} \left( 1 - y'_{j'(lns)} \right) \right) \geq lc \right\}.$$ If $j_1$ does not exist, indicating there is not oversaturated demand after $j_0$ at station $s$, then set $j_1 = J$.

Fourth, if $j_1 < J$, find the first time point after $j_1$ when the oversaturated actual boarding
passenger demand can be cleared, i.e., \( j_2 := \arg\min_{j \in \{j_{1+1}, \ldots, J\}} \left\{ \sum_{j_{1+1} \leq j} \sum_{u \in S_{s+1}} \sum_{v \in S_v^+} (a_{j_u} p_{j_u v} (1 - y_{j_{1+1} s}^u)) \right\} \leq lc(j - j_1) / J \). Note that \( j_2 \) must exist because otherwise the passengers cannot be all accommodated at \( J \) even if we serve them with \( lc / J \), which indicates that the problem is infeasible.

With this construction, there are always passengers left at station \( s \) after each dispatch at any \( j \in [j_1, j_2] \). Thus, \([j_1, j_2]\) is an oversaturated period. Otherwise, if \( j_1 = J \), set \( j_2 = J \).

Fifth, revise the passenger arrival demand. To convert the passenger arrival demand curve to one with unsaturated traffic, we need to replace the arrival demand rate during the oversaturated period with the maximum service rate. If \( j_1 = j_2 = J \) (i.e., there is no oversaturated demand), set \( a_j' = a_j, \forall j \in [j_0, j_1] \). Otherwise, set \( a_j' = a_j, \forall j \in [j_0, j_1] \) and \( a_j' = \frac{lc}{J}, \forall j \in [j_1, j_2] \). With this, \( a_j' \leq \frac{lc}{J}, \forall j \in [j_0, j_2] \) and thus the original passenger arrival demand curve is revised into one with only unsaturated traffic.

Sixth, revise the passenger demand distribution. If \( j_1 = j_2 = J \) (i.e., there is no oversaturated demand), set \( p_j' = p_j, \forall j \in S^+, j \in [j_0, j_1] \) and go to the second step. Otherwise, set \( p_j' = p_j, \forall j \in [j_0, j_1] \); the passenger demand remains unchanged during \([j_0, j_1] \) and thus the demand distribution remains the same as well. For \( j \in [j_1, j_2] \), we need to recompute the passenger demand distribution as the number of passengers who can board the vehicles if the system is served with the maximum service rate. Set \( k = 0 \) and then conduct the following steps.

(i) Find the first time point with passenger queue, i.e., \( j' := \arg\min_{j' \in \{0, \ldots, J \}} \left\{ y_{j'_{1+1} u} < 1 \right\} \). Set \( k = k + 1 \), \( r = 0 \), \( s = 0 \), and \( r_{sv} = 0, \forall v \in S_v^+ \).

(ii) If \( r + \sum_{u \in S_{s+1}} \sum_{v \in S_v^+} (a_{j_u} p_{j_u v} (1 - y_{j_{1+1} s}^u)) < lc \), all passengers arriving at \( s \) at \( j' \) can board the vehicle. Thus, set \( y_{j'_{1+1} s} = 1 \). Otherwise, the number of passengers who can board
the vehicle is limited by the remaining capacity of the vehicle. Set \( y_{j_{ln}s} = (Ic - r - \sum_{u \in S_u^+} \sum_{v \in S_v^+} (a_{j'u}p_{j'u}y_{j_{ln}u} - y_{j'(j_{ln-1})u})) / a_{js} \).

(iii) Set \( r = r + \sum_{u \in S_u^+} \sum_{v \in S_v^+} (a_{j'u}p_{j'u}y_{j_{ln}u} - y_{j'(j_{ln-1})u}) \), \( r_s = r + a_{j's} (y_{j_{ln}s} - y_{j'(j_{ln-1})s}) \), and \( r_{sv} = r_{sv} + r_s p_{j'sv}, \forall v \in S_v^+ \). If \( r < Ic \), set \( j' = j' + 1 \) and go to Step (ii). Otherwise, set \( p'_{j'sv} = r_{sv} / r_s \), \( \forall j \in \left[ j_1 + (k - 1)j_2, j_1 + kj_2 \right], v \in S_v^+ \) and then check if \( k = \left\lceil (j_2 - j_1) / j_1 \right\rceil \). If yes, set \( j_0 = j_2 = j_1 + 1 \), \( a_{js} = a_{j's}, \forall j \in \{ j_{ln} + 1, \ldots, J \}, s \in S \), \( p_{j'sv} = p'_{j'sv}, \forall j \in \{ j_{ln} + 1, \ldots, J \}, s \in S, v \in S_v^+ \) and go to the second step. Otherwise, go to Step (i).

After this passenger demand revision process, the CA method in Section 4.4 can be applied.

### 4.6 Numerical Experiments

This section assesses the proposed models and station-wise docking operation with two sets of numerical experiments. The first set of experiments is built on a hypothetical transit corridor in Mandl’s network (Mandl, 1980) to assess the computation performance of the proposed models, and analyze the system performance with varying input parameters. The second set of experiments apply the proposed models to design future MAV service with traffic demand data collected from the Batong line in the Beijing Subway system. The purpose of these experiments is to study the applicability of the proposed methods to real-world problems and to further reveal some interesting managerial insights. All experiments are run on a DELL Studio PC with 3.60 GHz of Intel Core i7-7700 CPU and 16 GB RAM in a Windows environment. All the algorithms are implemented in Visual C++ 2015.

#### 4.6.1 Case Study 1: A Hypothetical Transit Corridor in Mandl’s Network

We consider a hypothetical transit corridor in Mandl’s network (Mandl, 1980). As shown in Fig. 4.6, Mandl’s network includes a group of cities in Switzerland, consisting of 15 nodes and
21 bidirectional links. Without loss of generality, we select 11 nodes (the node with shaded circles) to construct a unidirectional transportation corridor starting at node 1 and terminating at node 5 (with the direction of each link indicated by the arrows). The time-dependent travel demand is simulated with data in Arbex and da Cunha (2015). Specifically, we first obtain the peak-hour passenger arrival rate between each OD pair by simulating the passenger arrival process as a Poisson process with the average passenger demand rate from Arbex and da Cunha (2015). Then the OD matrix in other periods is obtained by multiplying the ratio of the time-of-day factor of the target period to the factor of the peak hour. We assume that the constructed corridor is served by MAVs that can dock and undock at any station along the corridor. To make sure all passengers can be served at the end of the operational horizon with a minimum dispatch headway of $h = 3$ min, we set $J := [1,2,3]$ and $c = 50$ passengers/pod. Further, the operational cost of MAVs function is $f_i = 1.912 + 29.50i, \forall i \in J$. Besides, the unit-time waiting cost per passenger is set as $w = 0.8$ $$/\text{min}.$

Figure 4.6: Mandl’s network and the hypothetical corridor.
4.6.1.1 Computation Performance

This subsection examines the computation performance of the proposed solution methods. Specifically, we compare the performance of five different solution methods, including a state-of-the-art commercial solver Gurobi, a standard B&B algorithm where the state reduction properties are not applied, the expedited B&B algorithm, the CA method with cost estimated by integral (CA-I), and the CA method with cost estimated via discretization (CA-D). Further, we set the time limit of each solution approach to 1 hour; i.e., we stop the program after running the program for an hour if it has not solved the problem to optimality. To understand how these solution methods performance under different instance sizes, we create 39 instances with different numbers of origin stations (i.e., 2 – 10) and lengths of time horizons (i.e., 0.15, 0.5, 1.0, 1.5, 2.0 hrs) using the simulated demand. For the convenience of the illustration, in the following analysis we denote an instance as IN-Number of origin stations-Length of the operational horizon. For example, IN-2-1.5 denotes an instance with 2 origin stations and an operational horizon of 1.5 hrs.

To evaluate the solution quality of the exact solution methods, we record the optimality gap defined as $\frac{B - B^*}{B^*} \times 100$. Note that the lower the optimality gap, the better the solution quality. For the CA methods, we compute the relative gap between their objective values and those from the exact solution method. Note that the smaller the relative gap, the closer the objective value computed from the CA method is to that from the exact optimal solution, and thus the better the solution quality of the CA method. We also record the computation time for a comparison between the computation speeds of different methods. Comparisons of solution quality are summarized in Table 4.1. A comparison of the computation time is summarized in Table 4.2.

We see from Table 4.1 that Gurobi can only solve the optimal solution for very small instances within an hour. For some instances (e.g., IN-6-0.5), Gurobi obtains a feasible solution
with relatively small optimality gaps within an hour. However, there are eleven instances for which Gurobi cannot even find a feasible solution after running for an hour, which can be attributed to two reasons. For some instances (e.g., IN-7-1), this happens because before starting the branch and bound procedure, Gurobi adopts a collection of problem reductions to downsize the problem and to tighten its formulation (Gurobi, 2020). Yet, the model size in these instances is so large that these problem reductions cannot be completed within an hour, and therefore Gurobi fails to offer even a feasible solution. For other instances (e.g., IN-6-1.5), the model contains so many decision variables and constraints that it is out of the memory assigned to Gurobi on the working machine. The standard B&B algorithm performs much worse than Gurobi. Specifically, only the optimal solutions of 5 instances are successfully solved by the standard B&B algorithm. For the rest of the instances, the optimality gaps are mostly 30%, with a minimum of 19.5% and a maximum of 50.6%. Such a large optimality gap reveals the standard B&B algorithm is far from converging to the true optimal solution after solving the model for an hour. This result can be explained by the large number of branches at each node, leading to a branch and bound tree growing exponentially. Although conventional bounding methods are used to prune nodes, the number of nodes that must be visited is too substantial for the standard B&B algorithm to acquire the optimal solution. After applying the solution space reduction properties, the expedited B&B algorithm successfully solves all the instances reported in Table 4.1 with the preassigned time and memory resources.

For the CA methods, we see that CA-I produces positive or negative relative gaps, indicating that CA-I can either overestimate or underestimate the optimal objective value, which is consistent with the finding in Chapter 3. Despite the signs of the relative gaps in total system costs from CA-I, their absolute values are all under 7% and only one of them are above 5%. These results indicate that the CA-I estimates the total cost of transit corridors with good accuracy even
without discretization. For CA-D, all relative gaps are positive since the discretization method ensures that all solutions are feasible. Further, the relative gaps in the total cost are all below 5% and most of them are under 2%, indicating the CA-D can produce near-optimal solutions to the investigated problem with high accuracy. Also, we can observe that the relative gap (i.e., the accuracy of the CA method) generally decreases as the instance size (or the number of stations) increases, meaning the CA method is appropriate for large-scale problem.

In terms of the computation time, we see that as the instance size increases, the computation time of Gurobi increases from less than one second to more than an hour very quickly, which is consistent with the fact the problem size grows with exponential speed. Thus, existing commercial solvers may not be an ideal solution approach for the problem in engineering applications. Again, the standard B&B algorithm shows much worse performance than Gurobi. In four of the instances for which the standard B&B algorithm can solve the optimal solutions, its computation time is significantly longer than Gurobi. For instances that the standard B&B algorithm cannot find the true optimal solution, it reaches worse (i.e., larger) optimality gaps with longer computation time (in most instances more than 1000 times as slower). Worse still, the computation time of the standard B&B algorithm increases at a faster speed than Gurobi’s, owing to the dramatically increasing number of nodes in the branch and bound tree. However, the expedited B&B algorithm can solve the optimal solutions to all instances in much less time than Gurobi. We see from Table 4.2 that the expedited B&B algorithm solves the problem 99 times as quickly as Gurobi most instances. The computation time of the expedited B&B algorithm in most instances is within 5 mins (i.e., 300 seconds) and the largest case is less than 15 mins. The CA methods could tackle all instances with astonishing efficiency, i.e., within ten milliseconds. Besides, we can also see that the computation time of the CA methods are relatively stable. This is because the CA method needs
to solve a finite number of unit-time problems and thus has only linear time complexity. These results show that the CA methods are also attractive from the perspective of computation efficiency.

Table 4.1: Solution quality of different solution methods. / indicates that the instance is unsolved after running for an hour.

<table>
<thead>
<tr>
<th>Instance</th>
<th>Optimality gap (%)</th>
<th>Relative gap (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gurobi</td>
<td>Standard B&amp;B</td>
</tr>
<tr>
<td>IN-2-0.15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN-3-0.15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN-4-0.15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN-5-0.15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN-6-0.15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN-7-0.15</td>
<td>0</td>
<td>39.500</td>
</tr>
<tr>
<td>IN-8-0.15</td>
<td>0</td>
<td>47.200</td>
</tr>
<tr>
<td>IN-9-0.15</td>
<td>0</td>
<td>48.600</td>
</tr>
<tr>
<td>IN-10-0.15</td>
<td>0</td>
<td>50.600</td>
</tr>
<tr>
<td>IN-2-0.5</td>
<td>0</td>
<td>19.500</td>
</tr>
<tr>
<td>IN-3-0.5</td>
<td>0</td>
<td>27.100</td>
</tr>
<tr>
<td>IN-4-0.5</td>
<td>0</td>
<td>27.900</td>
</tr>
<tr>
<td>IN-5-0.5</td>
<td>0</td>
<td>29.500</td>
</tr>
<tr>
<td>IN-6-0.5</td>
<td>0.007</td>
<td>30.600</td>
</tr>
<tr>
<td>IN-7-0.5</td>
<td>0.009</td>
<td>31.400</td>
</tr>
<tr>
<td>IN-8-0.5</td>
<td>0.008</td>
<td>32.800</td>
</tr>
<tr>
<td>IN-9-0.5</td>
<td>0.931</td>
<td>33.400</td>
</tr>
<tr>
<td>IN-10-0.5</td>
<td>0.864</td>
<td>35.400</td>
</tr>
<tr>
<td>IN-2-1</td>
<td>0.401</td>
<td>28.300</td>
</tr>
<tr>
<td>IN-3-1</td>
<td>1.183</td>
<td>30.500</td>
</tr>
<tr>
<td>IN-4-1</td>
<td>0.970</td>
<td>29.800</td>
</tr>
<tr>
<td>IN-5-1</td>
<td>1.384</td>
<td>29.600</td>
</tr>
<tr>
<td>IN-6-1</td>
<td>1.589</td>
<td>29.500</td>
</tr>
<tr>
<td>IN-7-1</td>
<td>/</td>
<td>30.000</td>
</tr>
<tr>
<td>IN-8-1</td>
<td>/</td>
<td>30.400</td>
</tr>
<tr>
<td>IN-9-1</td>
<td>/</td>
<td>31.400</td>
</tr>
<tr>
<td>IN-10-1</td>
<td>/</td>
<td>32.300</td>
</tr>
<tr>
<td>IN-2-1.5</td>
<td>0.437</td>
<td>31.000</td>
</tr>
<tr>
<td>IN-3-1.5</td>
<td>0.936</td>
<td>32.400</td>
</tr>
<tr>
<td>IN-4-1.5</td>
<td>1.183</td>
<td>31.500</td>
</tr>
<tr>
<td>IN-5-1.5</td>
<td>/</td>
<td>30.000</td>
</tr>
<tr>
<td>IN-6-1.5</td>
<td>/</td>
<td>30.400</td>
</tr>
<tr>
<td>IN-7-1.5</td>
<td>/</td>
<td>31.080</td>
</tr>
<tr>
<td>IN-8-1.5</td>
<td>/</td>
<td>31.100</td>
</tr>
<tr>
<td>IN-2-2</td>
<td>0.381</td>
<td>31.900</td>
</tr>
<tr>
<td>IN-3-2</td>
<td>0.703</td>
<td>33.400</td>
</tr>
<tr>
<td>IN-4-2</td>
<td>/</td>
<td>30.100</td>
</tr>
<tr>
<td>IN-5-2</td>
<td>/</td>
<td>33.500</td>
</tr>
<tr>
<td>IN-6-2</td>
<td>/</td>
<td>31.800</td>
</tr>
</tbody>
</table>
Table 4.2: Computational time of different solution methods

<table>
<thead>
<tr>
<th>Instance</th>
<th>Gurobi</th>
<th>Computation time (CPU sec)</th>
<th>Standard B&amp;B</th>
<th>Expedited B&amp;B</th>
<th>CA-I</th>
<th>CA-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN-2-0.15</td>
<td>0.04</td>
<td>0.028</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-3-0.15</td>
<td>0.04</td>
<td>0.230</td>
<td>0.006</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-4-0.15</td>
<td>0.09</td>
<td>1.739</td>
<td>0.011</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-5-0.15</td>
<td>0.08</td>
<td>13.419</td>
<td>0.024</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-6-0.15</td>
<td>0.2</td>
<td>185.529</td>
<td>0.047</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-7-0.15</td>
<td>0.18</td>
<td>3600</td>
<td>0.068</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-8-0.15</td>
<td>0.33</td>
<td>3600</td>
<td>0.094</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-9-0.15</td>
<td>0.37</td>
<td>3600</td>
<td>0.116</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-10-0.15</td>
<td>0.52</td>
<td>3600</td>
<td>0.142</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-2-0.5</td>
<td>45.600</td>
<td>3600</td>
<td>0.106</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-3-0.5</td>
<td>526.630</td>
<td>3600</td>
<td>0.285</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-4-0.5</td>
<td>177.210</td>
<td>3600</td>
<td>1.114</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-5-0.5</td>
<td>123.180</td>
<td>3600</td>
<td>2.177</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-6-0.5</td>
<td>163.450</td>
<td>3600</td>
<td>6.910</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-7-0.5</td>
<td>1433.070</td>
<td>3600</td>
<td>20.989</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-8-0.5</td>
<td>2432.320</td>
<td>3600</td>
<td>35.332</td>
<td>0.001</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>IN-9-0.5</td>
<td>3600.050</td>
<td>3600</td>
<td>72.269</td>
<td>0.001</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>IN-10-0.5</td>
<td>3600.030</td>
<td>3600</td>
<td>114.526</td>
<td>0.002</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>IN-2-1</td>
<td>3600.060</td>
<td>3600</td>
<td>0.509</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>IN-3-1</td>
<td>3600.260</td>
<td>3600</td>
<td>1.212</td>
<td>0.001</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>IN-4-1</td>
<td>3600.050</td>
<td>3600</td>
<td>4.298</td>
<td>0.001</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>IN-5-1</td>
<td>3600.070</td>
<td>3600</td>
<td>9.120</td>
<td>0.001</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>IN-6-1</td>
<td>3600.030</td>
<td>3600</td>
<td>31.429</td>
<td>0.001</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>IN-7-1</td>
<td>/</td>
<td>3600</td>
<td>83.203</td>
<td>0.001</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>IN-8-1</td>
<td>/</td>
<td>3600</td>
<td>174.016</td>
<td>0.001</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>IN-9-1</td>
<td>/</td>
<td>3600</td>
<td>472.522</td>
<td>0.001</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>IN-10-1</td>
<td>/</td>
<td>3600</td>
<td>812.663</td>
<td>0.001</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>IN-2-1.5</td>
<td>3600.040</td>
<td>3600</td>
<td>1.777</td>
<td>0.001</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>IN-3-1.5</td>
<td>3600.030</td>
<td>3600</td>
<td>3.233</td>
<td>0.002</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>IN-4-1.5</td>
<td>3600.030</td>
<td>3600</td>
<td>9.415</td>
<td>0.002</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>IN-5-1.5</td>
<td>/</td>
<td>3600</td>
<td>22.739</td>
<td>0.002</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>IN-6-1.5</td>
<td>/</td>
<td>3600</td>
<td>76.189</td>
<td>0.002</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>IN-7-1.5</td>
<td>/</td>
<td>3600</td>
<td>433.875</td>
<td>0.002</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>IN-8-1.5</td>
<td>/</td>
<td>3600</td>
<td>557.667</td>
<td>0.002</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>IN-2-2</td>
<td>3600.030</td>
<td>3600</td>
<td>3.979</td>
<td>0.002</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>IN-3-2</td>
<td>3600.040</td>
<td>3600</td>
<td>7.130</td>
<td>0.002</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>IN-4-2</td>
<td>/</td>
<td>3600</td>
<td>20.490</td>
<td>0.002</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>IN-5-2</td>
<td>/</td>
<td>3600</td>
<td>45.681</td>
<td>0.002</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>IN-6-2</td>
<td>/</td>
<td>3600</td>
<td>185.928</td>
<td>0.001</td>
<td>0.006</td>
<td></td>
</tr>
</tbody>
</table>

4.6.1.2 Effectiveness of the MAV Corridor Systems

This subsection investigates the effectiveness of station-wise dynamic capacity adjustment (SDCA) in the proposed MAV corridor system and how the system performance varies as key
input parameters change. Without loss of generality, we select IN-10-2 for the analysis since it has the largest problem size. In each experiment, we vary one input parameter and keep the others the same unless stated otherwise. To demonstrate the effectiveness of SDCA in transit corridors, we compare the results of the proposed system with a benchmark system where only the largest vehicle formation can be dispatched. Further, following Chapter 3, we adopt three metrics to quantitatively evaluate the system performance, namely average load percentage (ALP), total operation cost (TOC), and total waiting cost (TWC). ALP is defined as the average of the ratios of the number of onboard passengers to the vehicle capacity of each dispatch across all stations. TOC and TWC refer to the first and second term in objective function (4.3), respectively. Results of the analysis are summarized in Figure 4.7.

We first investigate the effectiveness of SDCA in the MAV corridor system. As can be seen from Figure 4.7, introducing SDCA in UMT systems can improve the system performance for all the parameter values considered in this subsection, though the difference in some instances (i.e., Figure 4.7 (q) and (r)) is too small to be visually observed. Specifically, by allowing SDCA along the corridor, we increase the average load percentage across all dispatches by up to around 40%, indicating a better resource utilization rate in the system. Besides, the total passenger waiting cost and operational cost are both reduced in almost all instances. In one exception where the total waiting cost increases, i.e., Figure 4.7 (e) when the minimum dispatch headway is 6 mins, the decrease in the other cost component is larger so that the total system cost is still decreased. Thus, introducing station-wise dynamic capacity adjustment via MAVs in UMTs can increase the vehicle utilization rate and decrease total system cost.

Next, we analyze the effects of each input parameter on the effectiveness of SDCA. Figure 4.7 (a)(c) show that as the maximum vehicle formation \( I \) increases, both the ALP gap (i.e.,
the vertical distance between the ALP curves) and the TOC gap (i.e., the vertical distance between the TOC curves) increase while the TWC gap (i.e., the vertical distance between the TWC curves) first experiences a plummet and then a steady growth. These observations suggest that the improvement brought by SDCA is the least when $I = 4$ and gets better as $I$ deviates from this value. When $I$ is less than 4, the passenger waiting cost in the benchmark system is extremely high since the inherent mechanism of the optimization model forces vehicles to be dispatched with longer headways. In the proposed MAV system, however, passengers are served with shorter vehicles and smaller headways, thus substantially reducing the passenger waiting cost. Once $I$ reaches 4, the proposed system design remains the same, but the total system cost in the benchmark system keeps increasing, therefore strengthening the effectiveness of the proposed SDCA operation. Since the total system cost in the proposed system reaches the minimum and remains the same after $I$ reaches 4, we set $I = 5$ in the following analysis to allow a larger feasible region as well as a better performance.

Figure 4.7 (d)-(f) reveal that the effectiveness of SDCA is less evident as the minimum dispatch headway $h$ increases. While the ALP gap remains relatively stable, both the TWC gap and the TOC gap (i.e., the vertical distance between the TOC curves) narrow down as $h$ increases. The reason is that the number of passengers waiting for boarding at all stations for each dispatch increases with the minimum time difference between every two consecutive dispatches, therefore raising the probability of using longer vehicles in the proposed MAV-based system. As a result, the capability to flexibly adjust vehicle capacity based on the passenger demand level is limited as $h$ increases, thus weakening the effectiveness of SDCA. Thus, reducing the minimum dispatch headway in transportation corridors, if possible, can improve the effectiveness of SDCA, which is consistent with the finding in Chapter 3.
Figure 4.7: Sensitivity analysis of Case study 1. Blue lines with square dots represent results from the proposed MAV system. Red lines with triangle dots represent results from the benchmark system. Dots on the border of the graphs represent values larger than the maximum values.
Figure 4.7 (g)-(i) imply that the increase in the unit-time passenger waiting cost $w$ amplifies the effectiveness of SDCA, as can be seen from the increasing ALP and TOC gaps. As $w$ increases, the operational cost and passenger waiting cost experience gradual growth in both the benchmark and proposed systems, since vehicles are dispatched more frequently to serve passengers with higher unit-time waiting cost. The total waiting cost in both systems seems to increase at the same pace, as can be seen from the relatively stable TWC gap in Figure 4.7 (h). Yet, the increase in the operational cost is almost negligible in the MAV-based system because shorter vehicles can be used to keep the operational cost small even though vehicles are dispatched more frequently.

The variations in the parameter representing the fixed energy cost regardless of vehicle capacity $C^F$ do not impose a substantial impact on the effectiveness of SDCA, with only mild fluctuations as shown in Figure 4.7 (i)-(k). This parameter increases the operational cost of each vehicle formation equally and thus the solutions are barely changed regardless of the parameter variations. However, the other two parameters in the energy cost function exhibit considerable influences on the effectiveness of SDCA; i.e., despite a slight decrease in the ALP gap, the increase in these two parameters makes SDCA save costs more substantially. For the parameter representing the variable energy cost dependent on the vehicle capacity, i.e, $C^V$, its increase results in a steady growth in the TWC gap while the TOC gap stays relatively stable (Figure 4.7 (n),(o)), indicating a larger gap between the total system costs. For the unitless parameter $\alpha$ in the energy cost function, we can see from Figure 4.7 (q)-(r) that the cost saving is very limited when $\alpha$ is less than 0.4 (both the TWC and TOC gap are almost 0). This is because when the value of $\alpha$ is so small, the operational costs of different vehicle formations are almost the same, thus rendering no substantial difference to dispatch long and short vehicles. As $\alpha$ increases, SDCA starts to bring observable savings in the operational cost because in this case, it is more favorable to dispatch
short vehicles when passenger demand is low. Moreover, when $a$ reaches around 0.7, the passenger waiting cost is reduced as well, resulting in a more substantial saving in the total system cost. Thus, it is crucial to analyze the extent of the economics of scale in the operational cost in UMT systems before introducing the SDCA paradigm since otherwise system improvement can hardly be achieved.

4.6.2 Case Study 2: MAV Service in Batong Line, Beijing Subway System

In this subsection, we apply the CA method to design future MAV service for the Batong line in the Beijing Subway system. We select the direction from Station 13 to 1 for the experiments in this case study. Since a large portion of passengers originating from these stations are destined to stations on other subway lines in the network, we treat all other stations in the network as a virtual destination station in the corridor. Further, oversaturated traffic is not the focus of this study so we consider an operational horizon starting from 11:00 a.m. and terminating at 23:00 p.m., during which only unsaturated traffic is present. With this, we count the passenger demand between each OD pair per minute using the smartcard data and obtain the passenger arrival demand passing through each station during each minute over the investigated operational horizon, as shown in Figure 4.8. In this case study, we envision a future scenario where the operator introduces the MAV technology in the system so that vehicles can change their capacity flexibly at any station across the corridor to serve the passengers. The same parameter settings in Case study 1 in Chapter 1 are used here. In each experiment, we vary only one parameter and remain the others the same as the default values to investigate the sensitivity. To evaluate the effectiveness of SDCA, we compare the proposed MAV-based system with a benchmark system where only one vehicle formation can be dispatched.
Figure 4.8: Passenger demand rate during the unsaturated period of Batong line in Beijing Subway system.

Figure 4.9: Optimal design for Batong line in Beijing Subway System with a minimum dispatch headway of 9 mins (top row), 6 mins (middle row), and 3 mins (bottom row).

We first plot the optimal design for three selected instances in Figure 4.9. We find that the optimal vehicle formations change dramatically across time and space in all three instances; i.e.,
long vehicles are dispatched at stations and time periods with intensive passenger demand while short vehicles are used for stations and time periods with relatively low passenger demand. Further, the larger the minimum dispatch headway, the higher the possibility of using long vehicles and thus the lower the flexibility to adjust vehicle capacity.

Next, we present the ALP, TWC, and TOC of all experiments in Figure 4.10. We see that introducing station-wise dynamic capacity adjustment can improve the performance of the Batong line in the Beijing Subway across all instances, with the average load percentage always being increased, the total passenger waiting cost being decreased or remaining the same, and the total operational cost always being decreased. The difference between this case and the previous case is that in addition to a few scenarios, the total passenger waiting costs are (almost) the same with and without SDCA. This is because the passenger demand in the Batong line is of a different magnitude compared with that in Mandl’s network. As a result, the passenger waiting cost is usually substantially larger than the operational cost, so the optimization model inherently adjusts the vehicle dispatch plan to maintain the passenger waiting cost at a relatively small value. This observation also reveals that the condition in the theoretical propositions may be satisfied in various UMT systems with similar cost structure as that in the Beijing Subway system, rendering a board applicability domain of the proposed method. As for each input parameter, we see that they influence the system performance metrics in a similar way as they influence the system performance in the previous case study.

Finally, the mean computation times for CA-I and CA-D across all experiments are 0.027 and 1.129 seconds, respectively, which is very efficient for an instance with 13 stations and 720 time indexes. The results verify the applicability of the proposed CA method in addressing real-world problems that are extremely difficult to solve efficiently with exact modeling methods.
Figure 4.10: Sensitivity analysis of Case study 2. Blue lines with square dots represent results from the proposed MAV system. Red lines with triangle dots represent results from the benchmark system. Dots on the border of the graphs represent values larger than the maximum values.
4.7 Chapter Summary

This chapter investigates methodologies for designing MAV corridor systems enabling station-wise docking. To this end, a macroscopic CA model and a microscopic discrete model are proposed. Numerical experiments are conducted to assess the models and the effectiveness of the station-wise docking. The main findings are:

(i) The CA model can solve near-optimal solutions that can satisfy the requirements of most engineering applications very efficiently (less than ten milliseconds) compared with the discrete modeling method (which may not even yield a feasible solution in several hours).

(ii) The expedited B&B algorithm proposed in this paper is able to solve the discrete model for relatively large problem instances within hundreds of seconds. The correctness of the B&B algorithm is verified since it produces the same solutions as Gurobi.

(iii) Theoretical properties reveal that the optimal vehicle formations should be either the upper- or lower-bound vehicle formation. There is a quantitative relationship between the passenger queue and vehicle dispatch headway in the optimal solution(s). Further, there is a dominance rule that can be applied to compare two feasible solutions. These are the foundation for designing efficient solution algorithms.

(iv) The effectiveness of the MAV-based station-wise docking operation is verified. Allowing vehicles to adjust their capacity flexibly according to the passenger demand at any station can reduce the total system cost up. Further, input parameters related to the vehicle operational cost and passenger waiting cost affect the system cost substantially. However, the system is not much affected by the variations in the capacity of a single modular pod and the minimum dispatch headway.

Overall, this paper fills the methodological void of operational design for MAV-based transportation corridors enabling station-wise docking. It provides analytical and numerical methods for solving realistic problem instances and offer managerial insights to UMT operators. Additionally, it adds knowledge on the impacts of MAV technologies on transportation corridors.
Chapter 5: Inequality Impacts of AV Systems

5.1 Overview

The equity impacts of emerging vehicle technologies are of concerns for many people. As a first step to study the equity dimension of the emerging vehicle technologies, this chapter studies the inequality impacts of AV systems. Note that equity and inequality are two different concepts. Equity refers to the fairness of the distribution of impacts among people (Litman, 2002). Inequality concerns about whether the impacts that homogeneous individuals receive is equal or not (García-Valiñas et al., 2005). When studying equity for different racial/ethnical groups, equality is often used to represent equity under the assumption that it is not fair for one group to have worse outcomes simply because of their race/ethnicity. However, when thinking about equity for groups defined by other sociodemographic, equality would be unfair (inequitable). For example, it would be unfair to expect people who can’t walk have equal access to wheelchairs as people who walk well. Thus, whether equality represents equity should be determined via moral judgement or justice theory. The focus of this study is to propose a method to quantify the inequality effects in terms of how outcomes from AV systems are distributed in society. How these inequality effects should be translated into equity outcomes is a research question left for future studies.

The AV systems are selected here for analysis for a number of reasons. First, AVs are acknowledged to be a critical component of next-generation transportation systems because of their great potentials in bringing substantial mobility, environmental, and public health benefits to society (Fagnant and Kockelman, 2014). Further, AVs, among other emerging vehicle systems, are promising to be realized in the future. With the rapid development of the AV technologies and
their potential substantial benefits, more people have come to question their equity/inequality impacts; specifically, whether costs / benefits brought by the AV technologies will be distributed fairly among society. Finally, the AV technologies play a crucial role in the development of other AV-related technologies, including the MAV transportation systems. Building a successful AV system is the necessary condition for implementing the MAV transportation systems.

Although equity studies have been conducted for conventional transportation systems (e.g. Delbosc and Currie, 2011; Guzman et al., 2017; Kaplan et al., 2014) and some shared-mobility systems (e.g. Chen et al., 2019b, Qian and Niemeier, 2019), analytical methods and empirical results on the equity/inequality impacts of AV systems can hardly be found in the literature. This chapter aims to study the inequality impacts of AV systems with a model-centered disaggregated approach. The disaggregated approach takes advantage of state-of-the-art planning models that yield individual trip diary information as a standard output. The approach starts from simulating individual travel demand with a sophisticated agent-based simulation model built from a multi-agent transport simulator, MATSim. Then, a list of cost / benefit factors and their corresponding measures are identified to quantify the outcomes of the AV systems. Finally, a series of statistical and econometric methods are applied to analyze the distribution of the costs / benefits from both the horizontal and vertical perspectives. From the horizontal perspective, we compute the Gini index and apply geographic mapping analysis. For the vertical perspective analysis, we compare summary statistics and calculate a subgroup inequality index. We apply the proposed model-centered framework to study AV systems in the Tampa Bay Region in the 2040 planning scenario. Results from these analyses will give us an answer about how costs / benefits of future AV systems will be distributed among different population groups in the Tampa Bay Region.
The proposed methodological framework stands out from existing ones to assessing the distribution of transportation costs/benefits and makes contributions to the literature as follows.

(i) This approach applies high resolution travel itinerary and sociodemographic data (e.g., detailed simulations of individual travelers) to identify user groups with multi-dimensional metrics (e.g., age, income, spatial distribution, etc.). The use of disaggregated data unveils individual heterogeneity that might be absorbed by existing methods with aggregated data. Further, the application of high-resolution data improves the scalability of the approach. Individual-level data can be easily aggregated to different levels (e.g., aggregated OD trips, census tract demographic data), enabling us to analyze the costs/benefits at whichever level the planning agencies desire.

(ii) The proposed model-centered methodology is built on high-definition planning models (i.e., the Tampa Bay Region Activity-Based Model, TBABM), which is consistent with state of the practice in transportation planning agencies. Specifically, to capture the uniqueness in AV systems, we customize MATSim to incorporate private AV operations and the impact of individual demographics on travel decisions. The consideration of these factors renders the methodology appropriate for AV analysis.

(iii) This approach combines horizontal and vertical analysis methods (which have been popular in transportation equity studies) to assess the inequality impacts of AV systems. The horizontal and vertical analysis will complement each other and offer policy makers a more complete picture of the inequality impacts of future AV scenarios. Empirical analysis results for the Tampa Bay Region will offer important managerial insights for local transportation authorities to formulate relevant policies for AV systems in a more equitable manner in the future.

The remainder of this chapter is organized as follows. Section 5.2 presents the proposed methodological framework and the case study application. Section 5.3 presents the main results.
from the case study application and discusses the inequality impacts as well as the limitations of this study. Finally, Section 5.4 concludes this chapter.

5.2 Methods

This section presents the methodological framework for analyzing the inequality impacts of AV systems, including the simulation framework, outcome measurements, and inequality analysis approach. Next, a case study application of the methodological framework in the Tampa Bay Region will be presented.

5.2.1 Methodological Framework

5.2.1.1 AV System Simulation

To estimate how AVs will affect individual travel behavior and the associated costs / benefits, we first develop a sophisticated agent-based simulation model for simulating AV systems. The simulation model is developed based on a state-of-the-art multi-agent transport simulator, MATSim, as shown in Figure 5.1. MATSim is an open-source, activity-based, multi-agent micro simulation software of transportation systems. It optimizes the travel itineraries for every agent in the simulated reality via co-evolutionary algorithms (Axhausen et al., 2016). MATSim is capable of generating detailed descriptions of individual travel behavior and being flexibly extended to include new functionalities via a modular programming paradigm. As a result, MATSim has been applied to simulate various transportation systems such as highway systems (Gurram et al., 2019), electric vehicle systems (Waraich et al., 2013), and AV systems (Liu et al., 2017). Following these studies, in this dissertation we also select MATSim as a basis for developing a simulation model for AV systems.

The MATSim simulation model takes as inputs the road network, the transit network and schedules, AV infrastructures (which is necessary for AV simulation), as well as individual travel
itineraries and sociodemographic. In the road network, we specify the coordinates of each node (i.e., intersection) and a series of information for each link (i.e., road), including the starting and terminating nodes, length, free-flow travel speed, capacity, number of lanes, and modes allowed to travel on the link. For transit operations, we provide information of the network structure, transit schedule, and transit fleet. Further, to allow the simulation of AVs, we also specify infrastructure related to AVs, which typically include links where AVs can travel and the AV fleet. Finally, for each agent, we provide the associated travel itinerary, which is called “plan” in MATSim, and optionally the sociodemographic information. The individual travel itineraries are necessary for MATSim to initialize the simulation while the optional sociodemographic information is used to update individual travel decisions.

Figure 5.1: Methodological framework for analyzing the inequality impacts of AVs.
With the inputs, MATSim starts updating individual plans via modifying the original plans, e.g., switching to other modes, changing activity duration, and rerouting. The updated plan will then be scored using scoring functions quantifying the utility that an individual traveler obtains by executing the plan. At each iteration, the plan with the highest score will be executed while those with the lowest scores are discarded, depending on how many plans can be stored for an individual. After running for a predefined number of iterations, the simulation model eventually generates the optimized plans and travel routes for all simulated individuals as well as the status of each link in the network.

A key challenge in applying MATSim for simulating AV systems is that the default scoring (or utility) functions do not consider individual sociodemographic. In its default setting, the entire population shares a set of parameters. As a result, the default scoring functions cannot reflect the impacts of individual sociodemographic (e.g., age, income) on their travel decisions. However, an important characteristic of AVs is that they may improve mobility for seniors and children who are uncomfortable or unable to drive (National Highway Safety Administration, 2017). If we use the default scoring functions in MATSim for the simulation, it is possible that a kid aged below 16 drives alone for their trips even in a non-AV scenario. To address this issue, we write a Java script to customize MATSim’s scoring functions. The idea of the customized scoring function is to add a penalty term to the default scoring function in MATSim. The penalty term will be an extremely large value if the simulator assigns driving as the mode of travel for an individual who cannot drive during the innovation process. In contrast, no penalty will be incurred if an agent capable of driving is assigned to drive during a trip.

Additionally, previous studies simulating AV systems with MATSim (e.g., Liu et al., 2017) have predominantly considered shared AV services using the dynamic transport services module.
With the shared AV services, travelers do not own AVs but would order a shared AV when they would like to travel. However, shared AV services is not the only possible operational model for AV systems. There has been evidence that individuals would like to use private AVs regardless of the availability of shared AV services (Haboucha et al., 2017; Zhang et al., 2018). Thus, it is important to incorporate the option of private AVs into the simulation to analyze the overall inequality impacts of the AV technologies. To this end, we take advantage of the shared AV functionality to allow private AV simulation in MATSim. Specifically, we estimate households that are willing to purchase private AVs as the top $\times$ percentage of all households ranked by their expenditure on cars between 2015 and 2018, where $\times$ is AV penetration rate in a specific study scenario. Then, we add the private AVs of each household into the shared AV fleet. However, instead of using the default strategy to match passengers with shared AVs, we introduce a condition to ensure that the private AVs of a household only serve members in this household. This way, operational characteristics of shared AVs (e.g., passenger-vehicle matching, empty vehicle hauling, reduced operational headways) can be directly applied to simulate the behavior of private AVs.

We implemented the above functionalities by writing Java scripts and adding them into the source codes of the core distribution and the dynamic transport services distribution in MATSim. With this, we compile all scripts into an executable jar-file that can be run on a local computer.

5.2.1.2 Transportation Outcome Metrics

To quantify the impacts of AV systems on travelers, we specify a series of cost / benefit metrics in terms of accessibility, mobility, and affordability in this section.

The purpose of a transportation system is to offer people/goods access to their destinations. Thus, accessibility, defined as the potential to reach opportunities for interactions (Hansen, 1959), is an important measure for assessing transportation systems. In this section, we consider three
types of opportunities, including job locations, education resources, and social activity locations. For each type of the opportunities, we measure accessibility with a binary function following El-Geneidy et al., (2016). Specifically, the travel itinerary from MATSim contains a set of trips an agent visits over a typical weekday. For each trip, the origin, destination, travel time, mode of travel, and route information are provided. With this information, an individual’s accessibility to a specific type of opportunity is computed as follows. First, the accessibility of a trip is defined as 1 if the travel time of this trip is no more than a given time threshold (defined by transportation planners case by case) and otherwise 0. Then, an individual’s accessibility to a type of opportunity is computed as the average of the accessibility of all trips whose destinations are facilities offering this type of opportunity.

While accessibility describes the activity range of an individual given her home location, available transportation services and the spatial distributions of resources/opportunities, mobility measures the efforts required to complete trips. That is, mobility measures the ease for individuals to reach selected destinations. Therefore, we use mobility as one of the performance metrics of transportation systems as well. There are multiple metrics that can be used to quantify mobility. In this dissertation, we use the person miles traveled per day as the mobility metric because of its simplicity and popularity in planning practice (National Research Council, 2002; Blumenberg et al., 2013). Note that the larger the mobility metric, the longer the distance one has to travel to reach her activity locations. Thus, a larger value of the person miles traveled indicates lower mobility.

Financial costs often determine which users benefit from technologies. Too high a user cost may exclude low-income users from using the technology since they cannot afford it. Thus, we include two affordability metrics to quantify the impact of a transportation systems on users’ affordability to travel, including the travel cost and the travel cost to income ratio. The travel cost
metric refers to an individual’s total monetary cost incurred by travels over a simulated weekday. For trips commenced by private cars (human-driven or autonomous), the travel cost is the product of the travel distance and the unit-distance travel cost. For the unit-distance travel cost, we use 0.608 $/mile for human-driven cars (AAA Association Communication, 2019) and 0.630 $/mile for private AVs (Bösch et al., 2018). For public transportation, the travel cost refers to the ticket fees a traveler pays to use the public transportation. The fare structure of transit systems can be very complicated in a metropolitan area, including the case study area in this study (Center for Urban Transportation Research, 2011). After communicating with local transit operators, we use 2$ as the transit ticket fee. For walk and bicycles, there are no monetary costs incurred during the travel so their travel costs are 0. Note that we only consider the monetary costs incurred by operating the vehicles while other costs (e.g., vehicle purchase cost) are omitted. The travel cost to income ratio is the ratio of the travel cost to the daily income of that individual. Also, the larger the travel cost and travel cost to income ratio, the lower one’s affordability.

Finally, we want to note that although only accessibility, mobility, and affordability are analyzed in this dissertation, other impacts such as those that quantify public health impacts can be analyzed with the same methodological framework should the appropriate impact metrics were identified.

5.2.1.3 Inequality Analysis Methods

This section presents methods to analyzing the inequality impacts of the AV systems with the outcome metrics defined in the previous subsection. Following other transportation equity studies (e.g., Chen et al., 2019b), we study the inequality impacts of AV systems from both the horizontal and vertical perspectives. The horizontal perspective is concerned with the distribution of transportation outcomes among homogeneous individuals or groups spread across space. The
vertical perspective accounts for differences in individuals’ sociodemographic characteristics, needs, and/or capabilities. With these definitions, we adopt a series of analysis methods from Chen et al. (2019b) to quantify the horizontal and vertical inequality impacts.

For the horizontal perspective analysis, we compute the Gini index and also compare mapped geographic differences. The Gini index is a quantitative measure of inequality that offers an overall description of the degree of inequality of resources. It has been widely applied for equity/inequality analysis in the transportation filed (Delbosc and Currie, 2011). Here we use it to quantify the overall inequality in the distribution of different transportation outcomes. Gini index ranges between 0 and 1, with 0 indicating perfect equality and 1 perfect inequality. We compute the Gini index with the ineq package in R statistical software. Also, we compute the mean outcome values of all individuals in a census tract as the outcome value of that census tract. Then we plot the spatial distribution of the outcome in ArcGIS to provide a visual understanding of distributions of impacts.

For the vertical perspective analysis, we compute the summary statistics and the subgroup inequality index. We compute the minimum, median, mean, maximum, and the relative gap between the subgroup mean and the population mean of the outcomes investigated. Population subgroups are categorized by sociodemographic such as gender, age, income, and so on. Subsequently, to quantify the degree of inequality, we apply the subgroup inequality index to conduct the inequality analysis. The subgroup index of inequality is developed by Stuart et al. (2009) that quantifies the degree to which a specific subgroup is disproportionately exposed to costs (e.g. pollutant emissions) and/or benefits (e.g. regulatory monitoring) of an environmental management system. A positive subgroup inequality index indicates that members of the subgroup (on average) are disproportionately exposed to the hazard (or benefit) at the level considered.
Instead, a negative subgroup inequality index means they are disproportionately exposed at this level. Chen et al. (2019b) applied the subgroup inequality index to measure the vertical inequality of the bike-sharing accessibility in southern Tampa. Following these studies, we compute the subgroup inequality index at and above four different outcome levels, including the minimum, 25th percentile, median, and the 75th percentile. For example, to compute the subgroup inequality index at and above the median job accessibility for rural residents, we first compute the proportion of rural residents among individuals whose job accessibility is equal to or greater than the median job accessibility, say A%. Then we compute the proportion of rural residents among the population, say B%. The subgroup inequality index for rural residents at and above the median job accessibility is then the common logarithm of the ratio of the proportion of rural residents among individuals whose job accessibility is equal to or greater than the median job accessibility to the proportion of rural residents among the population, i.e., $\log_{10} A/B$. The subgroup inequality indexes for other transportation outcomes at other levels are computed in a similar approach.

5.2.2 Case Study Application

We apply the proposed methodological framework to study the inequality impacts of the AV systems in the Tampa Bay Region in the 2040 planning year. A description of the study area, datasets used for the analysis, and the study scenarios are provided below.

5.2.2.1 Study Area

The case study focuses on Tampa Bay, Florida. The study region is covered by the Florida Department of Transportation (FDOT) District 7 travel demand model (see Figure 4.2). According to the US Census Bureau (2019), the Tampa Bay Region had a metropolitan area population of 3.1 million in 2018. The area has a low population density and is car reliant. It has a highly diverse population. Specifically, the largest racial/ethnic groups within the population are 64.9% white
(with 25.7% Hispanic or Latino), 24.2% African American, and 4.2% Asian. 16.5% of the population is foreign-born as of 2018. About 12 percent of the population is over the age of 65. Of the population under age 65, 8.3 percent have a disability as of 2018. The per-capita annual income, averaged from 2014-2018, is $34,570, with a median household income of $50,909. The percent of persons in poverty in Tampa is 19.5%. The high diversity in population demographic makes the area well suited for an inequality study.

Figure 5.2: Geographic extent of Tampa Bay study region (source: Google Maps)

Additionally, Florida has attracted various companies (e.g., Waymo, Argo Ai, and Ford) testing AV business models, including on-demand mobility and package delivery. In 2012, Florida began testing the AV technology. Downtown Tampa has planned an AV shuttle for Marion Street,
a north south roadway in the center of downtown. In 2019, Florida became the first state to pass legislation allowing automated vehicles to operate on their roadways without requiring a human driver behind the wheel. With these advancements, AV utilization in the near term is therefore more likely in the Tampa Bay Region than in locations where testing is not taking place. Thus, an study of the inequality impacts of AV systems in the Tampa Bay Region would be very beneficial for local transportation authorities.

5.2.2.2 Data Sources

To study the inequality impacts of AV systems in the study area with the proposed method, we use multiple different datasets. These include information on the transportation network, individual travel demand, and individual demographics from the Tampa Bay Activity-Based Model (TBABM) and consumer expenditure survey data from the US Bureau of Labor Statistics (2019).

Data on transportation network, individual travel demand and demographics are extracted from the 2040 scenario in the TBABM, which was developed for the Tampa Bay Region as part of the Federal Highway SHRP2 project (Gurram et al., 2019). The transportation network data offer information on the highway and transit network structure as well as transit operational information (e.g., vehicle dispatch frequencies). We use these data to generate the road network, the transit schedule, and the transit fleet input files required by MATSim. The synthetic population from the TBABM represents the real population in the FDOT District 7. The travel demand for individuals are described as daily travel itineraries, which contain the trip sequence, the origin, the destination, and the timing of each trip. The individual sociodemographic necessary for the customized scoring function and for the vertical inequality analysis is also collected. We consider age (below 65, above 65), income (low income defined as with an annual household income lower
than $35,000, upper income with an annual household income higher than $35,000), gender (male, female), and residence land use property (rural, urban). Further, the TBABM also provides information on the census tracts as residence locations, which allows us to aggregate individual outcomes for the spatial mapping analysis. Note that we simulate the AV system for the FDOT District 7 (the area covered by the TBABM; it is larger than the inequality analysis area defined in the last subsection) to make the MATSim model comparable to the TBABM. This treatment also mitigates the boundary effects since individual travel behaviors at the boundary of the Tampa Bay Region are also captured by the simulation model.

The consumer expenditure survey data is collected from the U.S. Bureau of Labor Statistics. It provides information on expenditures, income, and demographic characteristics of consumers in the United States. We extract the household expenditure on automobiles for households in Florida and the corresponding annual household income during 2015 and 2018 (these are used because they are the most up-to-date data when this study is being conducted). The household expenditure on automobiles are ranked to estimate the AV ownership for the synthetic population as described in Section 5.2.1.1. Then, the private AVs of each household and shared AVs are combined to generate an AV fleet file required to simulate AV systems in MATSim.

5.2.2.3 Study Scenarios

To explore the potential inequality impacts of AVs in the Tampa Bay Region, we compare the system performance of a base scenario and two AV scenarios. The base scenario is converted from the TBABM and does not offer AV services. The purpose of this scenario is to offer a benchmark for investigating how AVs will affect the inequality impacts of transportation systems in the study area. In this scenario, available modes of transportation include private car, transit, walk, bicycle, and school bus. The first AV scenario considers the near future where a small
portion (10%) of the households have purchased private AVs. The second AV scenario considers the relatively far future when 30% of the households in the study area have purchased private AVs. Ideally, dozens of scenarios could be set up representing different market penetration rates and service models of AVs to investigate how these factors would affect the inequality performance. The purpose of this study is to begin this work and to develop a template/framework for considering the inequality impacts of AV technologies deployment. Thus, we limit our analysis to two AV scenarios. A comprehensive study on all possible scenarios are left for future studies. The models are run on the high performance computing cluster at the University of South Florida.

5.3 Results and Discussions

This section presents and discusses the results of our analysis. Section 5.3.1 presents the summary statistics of the study data and transportation outcomes. Section 5.3.2 presents the results on the horizontal perspective analysis, i.e., the distribution of the transportation outcomes among individuals regardless of their sociodemographic and among the geographic space. Section 5.3.3 presents the results on the vertical perspective analysis, i.e., the distribution of the transportation outcomes among different population subgroups. Finally, Section 5.3.4 discusses the limitations of this study.

5.3.1 Summary Statistics

Table 5.1 presents the summary statistics of the study data and the transportation outcome metrics in the base and AV scenarios for the population in the study area.

We see that in the base scenario, the population in Tampa makes 4.2 trips and travels 19.01 miles a day on average. The average job accessibility is 0.74, which is lower than the average accessibility to schools (0.86) and social activities (0.84). The average travel cost is 10.05$/day and the travel cost to income ratio is 0.13 on average, which is consistent with the fact that the
transportation expenditure accounts for 10% ~ 20% of household income in the U.S (Litman, 2013). In the AV scenarios, the average number of trips remain the same. This is because MATSim does not change the trip sequence and the origin and destination of each trip in individual travel itineraries. However, the mean values of the accessibility to all types of opportunities, person miles traveled, travel cost and income to travel cost are all reduced. Further, as the AV proportion increases from 10% to 30%, the magnitude of the reduction in these mean values slightly increases.

The above observation is consistent with findings from previous simulation studies (e.g., Childress et al., 2015). This can be attributed to the low AV market penetration rates and empty vehicle trips induced by the AV operations. Specifically, although AVs can increase roadway capacity by reducing the headways between vehicles, the low market penetration rates limit AVs’ capability in increasing roadway capacity (Talebpour and Mahmassani, 2016). Further, the adoption of AV incurs empty trips to reposition the vehicles between activity locations to serve different members in a household (Levin et al., 2019). As a result, more vehicles travel in the network while the roadway capacity is slightly increased compared with the base scenario. This results in an increase in the travel time, which in turn produces a reduction in the accessibility outcomes. To keep the score (i.e., the utility that individuals received by commencing her daily activities and travels) of each individual as high as possible, the optimization mechanism in MATSim then decreases individual travel distance, which is translated to a decrease in the person miles traveled. Note that individuals are not involved during the empty vehicle trips. As a result, although AVs are usually found to increase the vehicle miles traveled (Cohn et al., 2019), the extract travel distance related to the empty vehicle travels does not contribute to individual travel distance and the mobility metric. Because the travel cost of private car trips is defined as a linear function of the travel distance and over 90% of the trips in the study area are commenced by private
cars, the reduction in the person miles traveled leads to a decrease in the travel cost. Finally, the individual daily incomes remain constant across all scenarios, and thus the travel cost to income ratio is reduced.

These results indicate that if AVs are operated as private vehicles and have a relatively low market penetration rate, they will not bring accessibility benefits to society as expected but the mobility and affordability benefits (i.e., travel cost) will be increased.

Table 5.1: Summary statistics of the study data and transportation outcome metrics.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>25th %tile</th>
<th>Median</th>
<th>Mean</th>
<th>75th %tile</th>
<th>Max</th>
<th>Sample size*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base scenario</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of trips</td>
<td>2.00</td>
<td>2.00</td>
<td>4.00</td>
<td>4.20</td>
<td>5.00</td>
<td>19.00</td>
<td>259840</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>0.00</td>
<td>0.50</td>
<td>1.00</td>
<td>0.74</td>
<td>1.00</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td>School accessibility</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.86</td>
<td>1.00</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td>Social accessibility</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.84</td>
<td>1.00</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td>Person miles travelled</td>
<td>0.00</td>
<td>7.25</td>
<td>14.89</td>
<td>19.01</td>
<td>26.49</td>
<td>161.65</td>
<td>259840</td>
</tr>
<tr>
<td>Travel cost</td>
<td>0.00</td>
<td>4.00</td>
<td>7.94</td>
<td>10.05</td>
<td>13.83</td>
<td>91.80</td>
<td>259840</td>
</tr>
<tr>
<td>Cost to income ratio</td>
<td>0.00</td>
<td>0.02</td>
<td>0.05</td>
<td>0.13</td>
<td>0.12</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td><strong>10% AV scenario</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of trips</td>
<td>2.00</td>
<td>2.00</td>
<td>4.00</td>
<td>4.20</td>
<td>5.00</td>
<td>19.00</td>
<td>259840</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.64</td>
<td>1.00</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td>School accessibility</td>
<td>0.00</td>
<td>0.50</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td>Social accessibility</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.73</td>
<td>1.00</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td>Person miles travelled</td>
<td>0.00</td>
<td>5.36</td>
<td>11.35</td>
<td>15.15</td>
<td>20.86</td>
<td>148.80</td>
<td>259840</td>
</tr>
<tr>
<td>Travel cost</td>
<td>0.00</td>
<td>4.00</td>
<td>7.01</td>
<td>9.12</td>
<td>12.20</td>
<td>85.52</td>
<td>259840</td>
</tr>
<tr>
<td>Travel cost to income ratio</td>
<td>0.00</td>
<td>0.02</td>
<td>0.05</td>
<td>0.12</td>
<td>0.11</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td><strong>30% AV scenario</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of trips</td>
<td>2.00</td>
<td>2.00</td>
<td>4.00</td>
<td>4.20</td>
<td>5.00</td>
<td>19.00</td>
<td>259840</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.60</td>
<td>1.00</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td>School accessibility</td>
<td>0.00</td>
<td>0.50</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td>Social accessibility</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.72</td>
<td>1.00</td>
<td>1.00</td>
<td>259840</td>
</tr>
<tr>
<td>Person miles travelled</td>
<td>0.00</td>
<td>5.05</td>
<td>11.02</td>
<td>14.99</td>
<td>20.71</td>
<td>150.39</td>
<td>259840</td>
</tr>
<tr>
<td>Travel cost</td>
<td>0.00</td>
<td>4.00</td>
<td>6.72</td>
<td>9.01</td>
<td>12.10</td>
<td>91.44</td>
<td>259840</td>
</tr>
<tr>
<td>Travel cost to income ratio</td>
<td>0.00</td>
<td>0.02</td>
<td>0.05</td>
<td>0.12</td>
<td>0.11</td>
<td>1.00</td>
<td>259840</td>
</tr>
</tbody>
</table>

*Number of individuals in the synthetic population

5.3.2 Inequality Analysis from the Horizontal Perspective

Table 5.2 summarizes the Gini index values for different transportation outcomes. Figures 5.2 – 5.6 present the spatial distribution of the accessibility, the person miles traveled, the travel cost, and the travel cost to income ratio, respectively.
We see from Table 5.2 that in the 2040 base scenario, the accessibility to job locations are relatively equally distributed among the population in the Tampa Bay Region, with a Gini index of 0.26. Figure 5.3 (top row, left column) further shows that individuals living in the north border and southeast corner have lower job accessibility than the remaining of the population. The distributions of the school accessibility and social activity accessibility are more equal, as indicated by the smaller Gini index values (i.e., 0.14 and 0.16). This result can also be found by comparing the spatial distributions in the subfigures in the left column of Figure 5.3. Specifically, compared with Figure 5.3 (top row, left column), Figures 5.3 (top row, middle and right columns) show less diverse colors; only several census tracts in the north border and southeast corner has low school and social activity accessibility values. Further, the adoption of private AVs consistently reduces individual accessibility to the three types of opportunities investigated, since the figures on the middle and right columns in Figure 5.3 all show lighter colors. However, the extent of the reduction is not the same for different individuals. For example, for individuals living in east side of the study area and with a job accessibility over 0.7, the job accessibility of some individuals is decreased to 0.6-0.7 and the job accessibility of others stays above 0.7. Consequently, individual differences in accessibility to jobs, school, and social activity opportunities all increases, leading to a more uneven distribution of these benefits from the horizontal perspective, evidenced by larger values of the Gini index. Thus, although the empty AV trips lower accessibility in the study area in general, its impacts are different on individuals in different geographic locations. Additionally,
regardless of a more disperse distribution of the accessibility, individuals living in the north border and southeast corner are still underrepresented. These observations also confirms our findings in the previous subsection.

Figure 5.3: Spatial distribution of accessibility. Top row: job accessibility, middle row: school accessibility, bottom row: social activity accessibility; left column: base scenario, middle column: 10% AV scenario, right column: 30% AV scenario.

The distribution of the person miles traveled is less equal compared with the distribution of accessibility in the 2040 base scenario, with a higher Gini index of 0.44. Individuals who live
in the east and north side of the study area generally have higher person miles traveled per day, as show in Figure 5.4 (left column). This observation indicates that individuals living in these areas travel longer distances (or make more efforts) to complete their daily activities; these areas are underrepresented in terms of the mobility benefits. Note that these are also the areas where individuals generally have lower accessibility, indicating that the results are internally consistent. In the AV scenarios, we obtain higher Gini indexes of 0.46 and 0.47, revealing that private AVs with a low market penetration rate produces a slightly more disperse distribution of the mobility benefits. Indeed, we see from Figure 5.4 (middle and right columns) that the person miles traveled per day is decreased in most census tracts (due to MATSim’s optimization mechanism as discussed above). Yet, for a few census tracts in the east border and area center, the person miles traveled remains the same level regardless of the adoption of the AVs. Probably individuals with AVs do not reside in these census tracts or the change in the person miles travelled of the individuals in these census tracts is too subtle to be reflected in the maps. Thus, the distribution of the person miles traveled is slightly more uneven in the study area.

Figure 5.4: Spatial distribution of person miles traveled (miles/day). Left column: base scenario, middle column: 10% AV scenario, right column: 30% AV scenario.
In terms of the travel cost, a Gini index value of 0.43 reveals that the overall degree of inequality of its distribution is similar to that of the distribution of the travel distance in the base scenario. The spatial distribution of the travel cost is very similar to that of the travel distance as well, as seen in Figure 5.5 (left column). Again, this is because the travel cost of driving (human-driven or autonomous vehicles) is a linear function of the travel distance and the majority of the trips (over 90%) are commenced via driving. Thus, the travel cost approximately linearly increases with the travel distance in the base scenario. Further, the introduction of the AVs trips does not much affect the distribution of the travel cost among individuals, as indicated by a Gini index of 0.42 and 0.44 in the 10% and 30% AV scenarios, respectively. We observe from Figure 5.4 that many census tracts (e.g., in the southwest corner and region center) witness a decline in the person miles traveled while the travel cost shown in Figure 5.5 remains at the same level. This is because the travel distance in these census tracts is reduced (AV trips), but the unit-distance travel cost of these AV trips is slightly higher than that of the private car trips in the base scenario (0.630 vs. 0.608$/mile). Thus, the reduction in the travel distance is compensated by the increase in the unit-distance travel cost. As a result, the individual travel costs in these census tracts are changed very little, which agrees with the fact that the mean individual travel cost decreases by around 1$ from the base to the AV scenarios (see Table 5.1). Thus, the spatial distribution of the travel cost also remains very similar, albeit changes in several census tracts.

Finally, we see from Table 5.2 that the distribution of the travel cost to income ratio is the most uneven in Tampa Bay Region, with Gini index values of 0.65, 0.65, and 0.66 at the base, 10% AV, and 30% AV scenario, respectively. Census tracts with different ratio values are scattered across the study region, as shown in Figure 5.6. However, in all scenarios, census tracts in the east border generally have higher values of the travel cost to income ratio, indicating that residents in these
areas have to spend a higher proportion of their income in transportation. Further, AVs do not change the distribution of the travel cost to income ratio since neither individual travel cost nor daily income is substantially changed.

Figure 5.5: Spatial distribution of travel costs ($/day). Left column: base scenario, middle column: 10% AV scenario, right column: 30% AV scenario.

Figure 5.6: Spatial distribution of the travel cost to income ratio (%). Left column: base scenario, middle column: 10% AV scenario, right column: 30% AV scenario.
5.3.3 Inequality Analysis from the Vertical Perspective

Table 5.3 presents summary statistics of different transportation outcomes by population subgroups. Figures 5.7 ~ 5.9 present the subgroup inequality indexes at different outcome levels.

From Table 5.3, we learn that the distribution of the various transportation outcomes among individuals in most population subgroups is similar to that among individuals in the entire study area. That is, for the majority of population subgroups, values of the summary statistics are close to those of the entire population. Regardless of the overall similarity, there are differences in the outcomes that different population subgroups receive from the transportation system. Particularly, the difference is substantial for some groups.

Such differences can be identified by comparing the mean values of different subgroups and the population mean, which is quantified by the gap column in Table 5.3. We first analyze the distribution for the base scenario. From the mean outcome values, we see that rural residents and individuals with an annual household income greater than $35,000 (i.e., the upper income group) consistently have lower mean accessibility to jobs, schools, and social activities. It is surprising that the upper income group is underrepresented in accessibility because this group is typically found to be overrepresented in transportation systems (El-Geneidy et al., 2016; Chen et al., 2019). This is because the high income population do not reside in the city center where most jobs, schools, and social activities locations are distributed (Nielsen Sitewise, 2015). Thus, they have to travel relatively longer time to reach these locations. Another surprising finding is that individuals over 65 have higher accessibility to jobs. This is because the seniors mainly reside in the east of the area where the job accessibility is high (Figure 5.3, top row left column) while the residence locations of those aged below 64 scatter across the study area. Further, people over 65 have lower
mean accessibility to school and social activities. In terms of gender, males have lower accessibility to jobs while females have lower accessibility to schools and social activity locations.

Regarding the person miles traveled, we see that individuals below 65, people having an annual household income higher than $35,000 (i.e., the upper income group), males, and rural residents have higher mean values than the population mean. Note that these subgroups also have lower accessibility to jobs, schools, and social activities on average, indicating that the results are consistent. This is because the accessibility metric is computed as a function of travel time and travel time is dependent of the travel distance (and thus the person miles traveled). As a result, individuals who have higher person miles traveled tend to experience lower accessibility. Further, since the travel cost is also dependent on the travel distance and the mode of transportation, these groups also have higher mean travel cost. However, despite the higher travel cost, the low income group and females have a higher travel cost to income ratio than the population mean, indicating that they spend more of their income on transportation. Thus, a lower travel cost does not necessarily mean that these groups have better affordability to the transportation services.

Turning to the 2040 AV scenarios, we see that the disparity directions are mostly the same as those in the base scenario with a few exceptions. For example, with the private AV operations, the low-income population tends to have lower mean accessibility to schools than the population mean. This is probably because with private AVs, the low-income population travel slower than in the base scenario and thus have lower accessibility to schools. Although the disparity directions of different transportation outcomes remain mostly the same, the magnitude of the disparity is different. Further, the change in the disparity magnitude can go both directions (i.e., increase or decrease), as evidenced by the change in the absolute value of the gaps from the base scenario to the AV scenarios in Table 5.3. For example, the absolute value of the gap between the mean job
accessibility of rural residents and the population decreases from 14.88% in the base scenario to 11.20% in the 30% AV scenario, indicating a less deviate distribution of the job accessibility between rural and urban residents. However, the absolute value of the gap between the person miles traveled of males and the population increases from 1.62% in the base scenario to 1.64% in the 30% AV scenario, indicating a more deviate distribution of the person miles traveled between males and females. These results show that operating private AVs with a low market penetration rate may not change which groups receive more benefits or bear most costs on average, but it can change the benefit / cost gaps between different subgroups. Further, whether the gap is widened or bridged depends on the specific outcomes and the groups being analyzed.

While comparing means offers a general picture of inequality between different subgroups, the subgroup inequality index offers a detailed description of the outcome distribution at different outcome levels. We see from Figures 5.7 ~ 5.9 that the disparity direction is consistent across all levels for all the three types of accessibility. That is, a group is overrepresented across all accessibility levels if it is overrepresented at one accessibility level. We also see that the disparity direction indicated by the subgroup inequality index agrees with that shown by comparing means, verifying that these two approaches are consistent. Further, by comparing the absolute value of the subgroup inequality index (i.e., the distance of the curve to the horizontal axis), we see that the accessibility is more unevenly distributed between rural and urban residents than subgroups defined by other sociodemographic attributes (e.g., gender). Indeed, the difference in the subgroup inequality index between income and gender groups in terms of accessibility is too subtle to be observed. This observation offers transportation planners an important implication – to reduce the accessibility gaps between different population groups in the transportation system with private AV services, policies should be designed to bridging the gaps between rural and urban residents.
The distributions of other transportation outcomes (i.e., person miles traveled, travel cost, and travel cost to income ratio) are more disperse compared with that of the accessibility, as shown in Figures 5.7 ~ 5.9. For these outcomes, the gaps are more evident between the rural and urban residents, the low and upper income groups, and people aged below and above 65. The difference between males and females are less substantial. Again, priorities should be given to subgroups with more evident gaps to achieve more effective policies. Further, subgroups are not necessarily consistently over/underrepresented at certain outcome levels. For example, rural residents are overrepresented above the median when it comes to the travel cost to income ratio but are underrepresented above the 75th percentile in all scenarios. This observation indicates that to reduce the outcome gap between such subgroups, transportation planners should target at different population groups at different outcome levels. Otherwise, the policy may be ineffective or even bring counter effects. Consider the case where policy makers are to reduce the gap in the travel cost to income ratio between rural and urban residents in the base scenario. If the policy makers ignore the variations in the disparity direction across different outcome levels and formulate the policies by simply comparing means, then the polices are likely to favor all rural residents since they have to pay a higher proportion of their daily income in transportation on average. However, such policies would also benefit rural residents who are already overrepresented above the 75th percentile of the travel cost to income ratio, and thus may eventually render rural residents the underrepresented group across all levels of this outcome. Instead, an effective policy should favor rural residents among those whose travel cost to income ratio is below the 75th percentile and favor urban residents among the remaining population. However, to divide the population into groups of different person miles traveled might not be simple and may require other sophisticated methods for the analysis.
Table 5.3: Summary statistics of transportation outcomes by population subgroups.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Gap* (%)</th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Gap* (%)</th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Gap* (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Job accessibility</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Base scenario</strong></td>
<td>0.00</td>
<td>1.00</td>
<td>0.74</td>
<td>1.00</td>
<td>/</td>
<td>0.00</td>
<td>1.00</td>
<td>0.64</td>
<td>1.00</td>
<td>/</td>
<td>0.00</td>
<td>1.00</td>
<td>0.60</td>
<td>1.00</td>
<td>/</td>
</tr>
<tr>
<td>under 65 population</td>
<td>0.00</td>
<td>1.00</td>
<td>0.74</td>
<td>1.00</td>
<td>-0.09</td>
<td>0.00</td>
<td>1.00</td>
<td>0.64</td>
<td>1.00</td>
<td>-0.10</td>
<td>0.00</td>
<td>1.00</td>
<td>0.60</td>
<td>1.00</td>
<td>-0.71</td>
</tr>
<tr>
<td>over 65</td>
<td>0.00</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>1.35</td>
<td>0.00</td>
<td>1.00</td>
<td>0.65</td>
<td>1.00</td>
<td>1.40</td>
<td>0.00</td>
<td>1.00</td>
<td>0.67</td>
<td>1.00</td>
<td>10.26</td>
</tr>
<tr>
<td>low income</td>
<td>0.00</td>
<td>1.00</td>
<td>0.79</td>
<td>1.00</td>
<td>6.01</td>
<td>0.00</td>
<td>1.00</td>
<td>0.67</td>
<td>1.00</td>
<td>5.21</td>
<td>0.00</td>
<td>1.00</td>
<td>0.64</td>
<td>1.00</td>
<td>5.93</td>
</tr>
<tr>
<td>upper income</td>
<td>0.00</td>
<td>1.00</td>
<td>0.73</td>
<td>1.00</td>
<td>-1.60</td>
<td>0.00</td>
<td>1.00</td>
<td>0.63</td>
<td>1.00</td>
<td>-1.39</td>
<td>0.00</td>
<td>1.00</td>
<td>0.59</td>
<td>1.00</td>
<td>-1.58</td>
</tr>
<tr>
<td>male</td>
<td>0.00</td>
<td>1.00</td>
<td>0.73</td>
<td>1.00</td>
<td>-1.85</td>
<td>0.00</td>
<td>1.00</td>
<td>0.63</td>
<td>1.00</td>
<td>-2.25</td>
<td>0.00</td>
<td>1.00</td>
<td>0.59</td>
<td>1.00</td>
<td>-1.82</td>
</tr>
<tr>
<td>female</td>
<td>0.00</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td>2.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.66</td>
<td>1.00</td>
<td>2.42</td>
<td>0.00</td>
<td>1.00</td>
<td>0.62</td>
<td>1.00</td>
<td>1.96</td>
</tr>
<tr>
<td>rural</td>
<td>0.00</td>
<td>1.00</td>
<td>0.63</td>
<td>1.00</td>
<td>-14.88</td>
<td>0.00</td>
<td>0.67</td>
<td>0.55</td>
<td>1.00</td>
<td>-13.73</td>
<td>0.00</td>
<td>0.50</td>
<td>0.54</td>
<td>1.00</td>
<td>11.20</td>
</tr>
<tr>
<td>urban</td>
<td>0.00</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>1.75</td>
<td>0.00</td>
<td>1.00</td>
<td>0.65</td>
<td>1.00</td>
<td>1.61</td>
<td>0.00</td>
<td>1.00</td>
<td>0.61</td>
<td>1.00</td>
<td>1.31</td>
</tr>
<tr>
<td><strong>School accessibility</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Base scenario</strong></td>
<td>0.00</td>
<td>1.00</td>
<td>0.86</td>
<td>1.00</td>
<td>/</td>
<td>0.00</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td>/</td>
<td>0.00</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>/</td>
</tr>
<tr>
<td>under 65 population</td>
<td>0.00</td>
<td>1.00</td>
<td>0.86</td>
<td>1.00</td>
<td>0.04</td>
<td>0.00</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td>0.04</td>
<td>0.00</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>0.02</td>
</tr>
<tr>
<td>over 65</td>
<td>0.00</td>
<td>1.00</td>
<td>0.79</td>
<td>1.00</td>
<td>-8.90</td>
<td>0.00</td>
<td>1.00</td>
<td>0.70</td>
<td>1.00</td>
<td>-8.36</td>
<td>0.00</td>
<td>1.00</td>
<td>0.72</td>
<td>1.00</td>
<td>-3.86</td>
</tr>
<tr>
<td>low income</td>
<td>0.00</td>
<td>1.00</td>
<td>0.86</td>
<td>1.00</td>
<td>0.10</td>
<td>0.00</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td>-0.63</td>
<td>0.00</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>-0.28</td>
</tr>
<tr>
<td>upper income</td>
<td>0.00</td>
<td>1.00</td>
<td>0.86</td>
<td>1.00</td>
<td>-0.06</td>
<td>0.00</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td>0.38</td>
<td>0.00</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>0.17</td>
</tr>
<tr>
<td>male</td>
<td>0.00</td>
<td>1.00</td>
<td>0.87</td>
<td>1.00</td>
<td>0.82</td>
<td>0.00</td>
<td>1.00</td>
<td>0.77</td>
<td>1.00</td>
<td>0.90</td>
<td>0.00</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td>0.73</td>
</tr>
<tr>
<td>female</td>
<td>0.00</td>
<td>1.00</td>
<td>0.86</td>
<td>1.00</td>
<td>-0.83</td>
<td>0.00</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>-0.91</td>
<td>0.00</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>-0.74</td>
</tr>
<tr>
<td>rural</td>
<td>0.00</td>
<td>1.00</td>
<td>0.78</td>
<td>1.00</td>
<td>-10.11</td>
<td>0.00</td>
<td>1.00</td>
<td>0.69</td>
<td>1.00</td>
<td>-9.18</td>
<td>0.00</td>
<td>1.00</td>
<td>0.69</td>
<td>1.00</td>
<td>-8.31</td>
</tr>
<tr>
<td>urban</td>
<td>0.00</td>
<td>1.00</td>
<td>0.87</td>
<td>1.00</td>
<td>1.29</td>
<td>0.00</td>
<td>1.00</td>
<td>0.77</td>
<td>1.00</td>
<td>1.17</td>
<td>0.00</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td>1.06</td>
</tr>
</tbody>
</table>

* Relative gap computed as (subgroup mean – population mean) / population mean * 100. Underrepresented groups are shaded blue.
Table 5.3 (continued).

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Gap* (%)</th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Gap* (%)</th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Gap* (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social accessibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>population below 65</td>
<td>0.00</td>
<td>1.00</td>
<td>0.84</td>
<td>1.00</td>
<td>/</td>
<td>0.00</td>
<td>1.00</td>
<td>0.73</td>
<td>1.00</td>
<td>0.02</td>
<td>0.00</td>
<td>1.00</td>
<td>0.72</td>
<td>1.00</td>
<td>-1.01</td>
</tr>
<tr>
<td>over 65</td>
<td>0.00</td>
<td>1.00</td>
<td>0.84</td>
<td>1.00</td>
<td>-1.40</td>
<td>0.00</td>
<td>1.00</td>
<td>0.73</td>
<td>1.00</td>
<td>-0.07</td>
<td>0.00</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>3.98</td>
</tr>
<tr>
<td>low income</td>
<td>0.00</td>
<td>1.00</td>
<td>0.85</td>
<td>1.00</td>
<td>1.15</td>
<td>0.00</td>
<td>1.00</td>
<td>0.74</td>
<td>1.00</td>
<td>1.98</td>
<td>0.00</td>
<td>1.00</td>
<td>0.74</td>
<td>1.00</td>
<td>2.30</td>
</tr>
<tr>
<td>upper income</td>
<td>0.00</td>
<td>1.00</td>
<td>0.84</td>
<td>1.00</td>
<td>-0.50</td>
<td>0.00</td>
<td>1.00</td>
<td>0.72</td>
<td>1.00</td>
<td>-0.87</td>
<td>0.00</td>
<td>1.00</td>
<td>0.71</td>
<td>1.00</td>
<td>-1.01</td>
</tr>
<tr>
<td>male</td>
<td>0.00</td>
<td>1.00</td>
<td>0.84</td>
<td>1.00</td>
<td>0.08</td>
<td>0.00</td>
<td>1.00</td>
<td>0.73</td>
<td>1.00</td>
<td>-0.01</td>
<td>0.00</td>
<td>1.00</td>
<td>0.72</td>
<td>1.00</td>
<td>0.09</td>
</tr>
<tr>
<td>female</td>
<td>0.00</td>
<td>1.00</td>
<td>0.84</td>
<td>1.00</td>
<td>-0.08</td>
<td>0.00</td>
<td>1.00</td>
<td>0.73</td>
<td>1.00</td>
<td>0.01</td>
<td>0.00</td>
<td>1.00</td>
<td>0.72</td>
<td>1.00</td>
<td>0.09</td>
</tr>
<tr>
<td>rural</td>
<td>0.00</td>
<td>1.00</td>
<td>0.74</td>
<td>1.00</td>
<td>-12.60</td>
<td>0.00</td>
<td>1.00</td>
<td>0.63</td>
<td>1.00</td>
<td>-12.63</td>
<td>0.00</td>
<td>1.00</td>
<td>0.63</td>
<td>1.00</td>
<td>11.91</td>
</tr>
<tr>
<td>urban</td>
<td>0.00</td>
<td>1.00</td>
<td>0.86</td>
<td>1.00</td>
<td>1.67</td>
<td>0.00</td>
<td>1.00</td>
<td>0.74</td>
<td>1.00</td>
<td>1.67</td>
<td>0.00</td>
<td>1.00</td>
<td>0.73</td>
<td>1.00</td>
<td>1.58</td>
</tr>
<tr>
<td>Person miles traveled (miles/day)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>population below 65</td>
<td>0.00</td>
<td>14.89</td>
<td>19.01</td>
<td>161.65</td>
<td>/</td>
<td>0.00</td>
<td>11.35</td>
<td>15.15</td>
<td>148.80</td>
<td>2.96</td>
<td>0.00</td>
<td>11.02</td>
<td>14.99</td>
<td>150.39</td>
<td>1.71</td>
</tr>
<tr>
<td>over 65</td>
<td>0.00</td>
<td>15.56</td>
<td>19.61</td>
<td>153.53</td>
<td>3.17</td>
<td>0.00</td>
<td>11.82</td>
<td>15.60</td>
<td>148.80</td>
<td>10.49</td>
<td>0.00</td>
<td>11.28</td>
<td>15.25</td>
<td>150.39</td>
<td>1.71</td>
</tr>
<tr>
<td>low income</td>
<td>0.06</td>
<td>12.07</td>
<td>16.36</td>
<td>161.65</td>
<td>-13.93</td>
<td>0.05</td>
<td>9.53</td>
<td>13.18</td>
<td>119.75</td>
<td>-13.04</td>
<td>0.02</td>
<td>9.95</td>
<td>13.87</td>
<td>128.20</td>
<td>-7.52</td>
</tr>
<tr>
<td>upper income</td>
<td>0.01</td>
<td>12.21</td>
<td>16.28</td>
<td>153.53</td>
<td>-14.33</td>
<td>0.01</td>
<td>9.45</td>
<td>13.09</td>
<td>128.12</td>
<td>-13.63</td>
<td>0.01</td>
<td>9.20</td>
<td>13.00</td>
<td>150.39</td>
<td>13.32</td>
</tr>
<tr>
<td>male</td>
<td>0.00</td>
<td>16.21</td>
<td>20.26</td>
<td>161.65</td>
<td>6.56</td>
<td>0.00</td>
<td>12.31</td>
<td>16.10</td>
<td>148.80</td>
<td>6.24</td>
<td>0.00</td>
<td>11.94</td>
<td>15.91</td>
<td>144.93</td>
<td>6.10</td>
</tr>
<tr>
<td>female</td>
<td>0.01</td>
<td>15.20</td>
<td>19.32</td>
<td>161.65</td>
<td>1.62</td>
<td>0.01</td>
<td>11.58</td>
<td>15.40</td>
<td>145.35</td>
<td>1.65</td>
<td>0.01</td>
<td>11.25</td>
<td>15.24</td>
<td>144.93</td>
<td>1.64</td>
</tr>
<tr>
<td>rural</td>
<td>0.00</td>
<td>14.59</td>
<td>18.71</td>
<td>153.53</td>
<td>-1.55</td>
<td>0.00</td>
<td>11.12</td>
<td>14.92</td>
<td>148.80</td>
<td>-1.58</td>
<td>0.00</td>
<td>10.81</td>
<td>14.76</td>
<td>150.39</td>
<td>-1.57</td>
</tr>
<tr>
<td>urban</td>
<td>0.01</td>
<td>20.03</td>
<td>24.65</td>
<td>148.45</td>
<td>29.67</td>
<td>0.01</td>
<td>15.88</td>
<td>20.22</td>
<td>128.12</td>
<td>33.43</td>
<td>0.01</td>
<td>15.64</td>
<td>20.17</td>
<td>143.29</td>
<td>34.53</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>14.29</td>
<td>18.25</td>
<td>161.65</td>
<td>-3.98</td>
<td>0.00</td>
<td>10.84</td>
<td>14.48</td>
<td>148.80</td>
<td>-4.48</td>
<td>0.00</td>
<td>10.50</td>
<td>14.30</td>
<td>150.39</td>
<td>-4.63</td>
</tr>
</tbody>
</table>

* Relative gap computed as (subgroup mean – population mean) / population mean * 100. Underrepresented groups are shaded blue.
Table 5.3 (continued).

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Gap* (%)</th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Gap* (%)</th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Gap* (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Travel cost ($/day)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Base scenario</td>
<td>0.00</td>
<td>7.94</td>
<td>10.05</td>
<td>91.80</td>
<td>/</td>
<td>0.00</td>
<td>7.01</td>
<td>9.12</td>
<td>85.52</td>
<td>/</td>
<td>0.00</td>
<td>6.72</td>
<td>9.01</td>
<td>91.44</td>
</tr>
<tr>
<td></td>
<td>10% AV scenario</td>
<td>0.00</td>
<td>8.19</td>
<td>10.31</td>
<td>91.80</td>
<td>2.55</td>
<td>0.00</td>
<td>7.25</td>
<td>9.35</td>
<td>85.52</td>
<td>2.42</td>
<td>0.00</td>
<td>6.82</td>
<td>9.11</td>
<td>91.44</td>
</tr>
<tr>
<td></td>
<td>30% AV scenario</td>
<td>0.00</td>
<td>6.81</td>
<td>8.93</td>
<td>76.35</td>
<td>-11.21</td>
<td>0.00</td>
<td>6.09</td>
<td>8.15</td>
<td>69.32</td>
<td>-10.66</td>
<td>0.00</td>
<td>6.31</td>
<td>8.54</td>
<td>72.94</td>
</tr>
<tr>
<td></td>
<td>Travel cost to income ratio (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Base scenario</td>
<td>0.00</td>
<td>0.05</td>
<td>0.13</td>
<td>1.00</td>
<td>/</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
<td>/</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>10% AV scenario</td>
<td>0.00</td>
<td>0.05</td>
<td>0.13</td>
<td>1.00</td>
<td>1.87</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
<td>2.19</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>30% AV scenario</td>
<td>0.00</td>
<td>0.15</td>
<td>0.28</td>
<td>1.00</td>
<td>-8.21</td>
<td>0.00</td>
<td>0.14</td>
<td>0.27</td>
<td>1.00</td>
<td>-9.66</td>
<td>0.00</td>
<td>0.14</td>
<td>0.27</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.00</td>
<td>0.06</td>
<td>0.12</td>
<td>1.00</td>
<td>-57.25</td>
<td>0.00</td>
<td>0.03</td>
<td>0.05</td>
<td>0.65</td>
<td>-59.44</td>
<td>0.00</td>
<td>0.03</td>
<td>0.05</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
<td>-3.75</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
<td>-3.84</td>
<td>0.00</td>
<td>0.05</td>
<td>0.11</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.00</td>
<td>0.06</td>
<td>0.13</td>
<td>1.00</td>
<td>3.59</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
<td>3.67</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.00</td>
<td>0.06</td>
<td>0.13</td>
<td>1.00</td>
<td>-0.76</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
<td>-2.95</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.00</td>
<td>0.05</td>
<td>0.13</td>
<td>1.00</td>
<td>-0.10</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
<td>0.39</td>
<td>0.00</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
</tr>
</tbody>
</table>

* Relative gap computed as (subgroup mean – population mean) / population mean * 100. Underrepresented groups are shaded blue.
Figure 5.7: Subgroup inequality index versus outcome levels for job accessibility and school accessibility. Top three rows represent the job accessibility at the base, 10% AV, and 30% AV scenarios, respectively. Bottom three rows represent the school accessibility at the base, 10% AV, and 30% AV scenarios, respectively.
Figure 5.8: Subgroup inequality index versus outcome levels for social activity accessibility and person miles traveled. Top three rows represent the social activity accessibility at the base, 10% AV, and 30% AV scenarios, respectively. Bottom three rows represent the person miles traveled at the base, 10% AV, and 30% AV scenarios, respectively.
Figure 5.9: Subgroup inequality index versus outcome levels for travel cost and travel cost to income ratio. Top three rows represent the travel cost at the base, 10% AV, and 30% AV scenarios, respectively. Bottom three rows represent the travel cost to income ratio at the base, 10% AV, and 30% AV scenarios, respectively.
5.3.4 Limitations

The current study has a few limitations. This study investigates the inequality effects of the outcomes of the future AV systems from the horizontal and vertical perspectives. This treatment is similar to many previous studies that adopted the horizontal and/or vertical analysis method to approach transportation equity (e.g., Guzman et al., 2017; Chen et al., 2019). However, as stated previously, inequality does not equate to equity. The inequality effects simply tell us how the observable outcomes are distributed among the population. Further, this study does not consider other aspects of equity such as procedural equity and restorative equity. To arrive at conclusions concerning the equity effects of the system, moral judgements or justice theory should be applied to determine if the quantified distributions meet the equity criteria (imposed by the transportation planners or the public) or not (van der Veen et al., 2020).

Inherent drawbacks of the agent-based transportation simulation models possibly introduce estimation bias in the analytical results. Transportation planning models are built on realistic traffic survey data; only individuals who travel within in study area are captured in the models. However, traditionally transportation-disadvantaged groups are observed to travel less or not travel at all due to various barriers to transportation systems (Huang et al., 2017). As a result, the demand for these individuals to reach activity locations are not well represented in the analysis. The sampling method adopted in the traffic survey may also affect the representativeness of some population groups. To address this issue, a possible solution is to look at equity from a different perspective. Instead of quantifying the transportation outcomes with observable individual travel behavior, one could consider individuals’ potential access to job opportunities and essential services (e.g., grocery stores, hospitals, and schools). The travel time from the simulation model can be used in Hensen’s cumulative accessibility formula to quantify the accessibility. This way, the demand of
all individuals to jobs and essential services can be captured, whether they travel or not. Readers are referred to Qian and Niemeier (2019) for details of such an approach.

The other inherent drawback of the agent-based simulation model is its inability to capture changes in the travel behavior of the disadvantaged populations. The advent of AVs is expected to alter individual travel behavior, including those who are traditionally transportation-disadvantaged. However, since the traffic simulation model may not capture the travel behavior of these groups, it cannot take into account how people in these groups will behave differently than that currently observed. The parameters we use in the simulation model are based on historical observations and may thus lead to biased estimation results. For example, we use the consumer expenditure data from 2015 to 2018 to estimate household AV ownership. An underlying assumption of this treatment is that the consumer preference and behavior of the Florida residents in the 2040 planning year would remain the same as those from 2015 to 2018. This assumption may not be true since there are many other factors also have a role to play regarding user preferences in AVs, such as technology interest, environmental concern, enjoy driving, and so on (Haboucha et al., 2017). This problem can be viewed as a problem of selectivity, with transportation outcome being an outcome of a selection process (to be in a subgroup for whatever reasons) that results in a nonrandom sample of data (Washington et al., 2020). Statistical modeling approaches could be applied to address this issue. For example, one can adopt a two-step approach that uses a probabilistic model to capture the selection process and incorporates the selection probability outcome to correct the bias in the transportation outcome model. Besides, a joint approach that models the selection process and transportation outcome jointly could also be applied. Please refer to Chapter 15 in Washington et al. (2020) for details regarding these statistical approaches.
Additionally, we only consider two AV scenarios where AVs operate as private vehicles. However, the future of AVs is uncertain, and thus there are many other theoretically feasible AV business models, such as shared AV services, niche services limited to designated environments, and dedicated AV lanes (Cohn et al., 2019). These business models may produce different equity impacts than the private AV scenarios investigated in this study. Additional simulation runs are necessary to understand the equity impacts of these models.

Finally, the population groups and transportation outcome metrics considered are limited. We only consider population subgroups defined by income level, gender, age, land use property, while racial equity, is not considered. There has been evidence that different racial groups do not benefit equally from transportation systems (Karner and Niemeier, 2013; Gurram et al., 2019). We select metrics in terms of accessibility, mobility, and affordability to quantify the impacts of the AV systems on travelers. Although these metrics have been popularly used in transportation planning practice, no single metric can identify all aspects of a transportation outcome. For example, although the person miles traveled has been popularly used as a mobility metric, it has also been criticized for not capturing the impact of travel speed on mobility. Thus, in practice, the planning agencies may need to consider population groups defined by other sociodemographic and multiple metrics for one outcome to paint a more complete picture of the system impacts.

5.4 Chapter Summary

This chapter proposes a methodological framework for studying the inequality impacts of AV systems. It consists of an agent-based simulation model, a series of transportation outcome metrics, and inequality analysis methods. The proposed methodological framework is applied to study future private AV scenario with low market penetration rates in the Tampa Bay Region. The main findings are:
(i) An AV system with private AVs and low market penetration rates will decrease the accessibility to jobs, schools, activity locations, the person miles traveled, the travel cost, and the travel cost to income ratio. Further, the increase in the AV market penetration rate from 10% to 30% slightly increases the magnitude of the reduction in these mean outcome values.

(ii) In the AV scenarios, individual differences in the accessibility to jobs, schools, and social activity opportunities increases, leading to a more uneven accessibility distribution from the horizontal perspective. However, the distributions of the person miles traveled, the travel cost, and the travel cost to income ratio are not substantially changed.

(iii) Incorporating private AVs with low market penetration rates may not change the disparity direction of the transportation outcome distributions. However, these systems can change the magnitude of the between group disparity. Further, whether the disparity is widened or bridged depends on the specific outcomes and the groups being analyzed.

The above findings offer managerial insights for transportation authorities in the Tampa Bay Region regarding the AV technologies. The case study also validates the proposed method’s capability in analyzing the equity performance of AV systems. However, only two AV scenarios is studied in this chapter. The future of AV technologies is uncertain and thus there are various possibilities of what an AV system will be like in the future. To draw a complete picture of the equity impacts of AV technologies, we will apply the proposed method to study other possible AV scenarios by incorporating other factors such as different market penetration rates, shared mobility services, and managed AV lanes, and so on. Additionally, incorporating the modular operations of MAVs into the simulation model to study the equity impacts of MAV transportation systems in a city scale would also be an interesting future research direction.
Chapter 6: The Road Ahead

MAVs are being tested in industry and may bring profound impacts to society in the future. This dissertation is the first step in answering the many scientific questions associated with MAV transportation systems. We propose the concept of elastic capacity transportation systems with MAVs as well as develop scalable computational and analytical tools for designing and analyzing these systems. Extensive empirical studies are conducted to test the validity and the performance of these tools and to assess the potential impacts (i.e., quality of service, energy implications, and inequality impacts) of the MAV/AV systems. We now close this dissertation with a summary of the main findings and a discussion of the possible avenues to extend this work.

6.1 Summary of Results

6.1.1 Computational and Analytical Tools

In Chapters 3 and 4, we develop a suite of computational and analytical tools to design the next-generation MAV systems. Numerical experiments are conducted to evaluate computational performance of these tools. Results reveal that existing commercial solvers (e.g., Gurobi) cannot solve discrete formulations of the system design problem to optimality in most cases due to the limited computational resources and the typical large problem scale in the real world. The proposed customized exact solution algorithms (i.e., the dynamic programming algorithm for shuttle system and the branch and bound algorithm for the corridor system design) solve the discrete models for medium-size problem instances to optimality within an acceptable amount of time. Thus, these algorithms can be used by researchers as benchmarks to evaluate future research efforts that aim to develop better solution approaches for MAV system design. They can also be used by future
MAV system planners to solve planning problems that has a relatively low requirement on the computational speed. Instead, the continuous approximation model solves highly accurate near-optimal solutions within one second. These continuous methods can be used to solve extremely large-scale problems and real-time operational design problems that demands a fast computational speed. The discrete and continuous methods complement each other in terms of solution quality and computational speed.

In Chapter 5, we develop a methodological framework for studying the inequality impacts of AV systems. This framework consists of a sophisticated agent-based simulation model, a series of transportation outcome metrics, and inequality analysis methods. The case study application verifies its applicability in studying the inequality impacts of AV systems in a metropolitan area. It can be used by transportation planners in the Tampa Bay Region and beyond for future equity studies on AV and AV-related emerging vehicle technologies.

These results answer the first scientific question listed in Chapter 1.

6.1.2 Properties of Optimal MAV System Design

We mathematically prove elegant theoretical properties of the optimal MAV system design and verify them with numerical experiments in Chapters 3 and 4.

For MAV shuttle systems, it is optimal to dispatch vehicles with the maximum capacity and the minimum headway (or the maximum transportation capacity) during oversaturated periods. With this, the original system design problem with oversaturated demand can be converted to a revised unsaturated problem with only unsaturated traffic. In the revised unsaturated problem, the optimal queue length right after each dispatch is less than the capacity of a single modular pod.

For MAV corridor systems, the optimal vehicle formation will be shorter than or equal to the upper-bound vehicle formation. This is a general property of the investigated problem and can
be applied at any parameter settings. Under certain (yet not uncommon) conditions, the optimal vehicle formation will be greater than or equal to the lower-bound vehicle formation. Further, if the passenger queue at each station along the corridor at the time point right before a dispatch is greater than the remaining capacity of the dispatched MAV at this station, then dispatching this MAV earlier will always lead to a better solution. We also identify a dominance rule to compare two feasible solutions.

These properties imply structural insights into the (near-) optimal MAV system design. They serve as the theoretical foundation for decomposing the spatiotemporal correlation between vehicle dispatch decisions in the MAV system design problem, which makes the problem regain the “local-impact” property required by the formulation of a continuum approximation model. Also, they are used to develop valid inequalities that can greatly reduce the solution space of the MAV system design problem, which greatly expedite the solution speed of the customized solution algorithms. Finally, following these properties, transportation practitioners can manually design operational plans for MAV system when the required computational resources are not available.

These results answer the second scientific question listed in Chapter 1.

6.1.3 Performance of MAV Systems

The empirical case studies show the performance of the MAV transportation systems by using existing fixed-capacity systems as benchmarks. Results reveal that it is necessary to flexibly adjust the dispatched vehicles’ capacities at different stations (only at the origin station for shuttle systems) across different dispatches. Indeed, the capacity of the vehicle currently adopted by the Batong Line in the Beijing Subway System is not needed most time of the day; the maximum capacity is only needed during the oversaturated period.
It is found that flexibly adjusting the vehicle capacity with the modular vehicle operations consistently increase the average load percentage (or vehicle occupancy) of the dispatched vehicles in MAV shuttle systems. Consequently, the operational cost (mainly comprised of energy cost) to move the vehicles from the origin to the destination is also consistently reduced. The passenger waiting cost is reduced or remains the same as the status quo in most instances. Only in several instances the passenger waiting cost is slightly increased. However, the slight increase in the passenger waiting cost is completely offset by the decrease in the vehicle operational cost. As a result, the total system cost (i.e., the sum of the vehicle operational cost and the passenger waiting cost) of the MAV shuttle systems is the same as that in the benchmark system.

Sensitivity analysis shows how variations in the input parameters affect the effectiveness of the flexible capacity operations of MAVs in shuttle systems. We find that the effectiveness of the flexible capacity operations increases as the increase in the maximum vehicle type, and the added dispatch cost per unit capacity. Adopting a longer minimum dispatch headway, instead, weakens the effectiveness of the flexible capacity operations, but there are still improvements in certain ranges. Additionally, the unitless parameter in the vehicle operational cost function and the cost of changing vehicle lengths are vital parameters that should be analyzed before introducing MAVs into transportation systems since system performances may be degraded if these parameters are not within a certain range.

Similar findings are found on the performance of the MAV corridor systems. Overall, these findings offer empirical evidence of the potential impacts of the MAV technology on urban mass transportation systems in terms of the quality of service (measured by the passenger waiting cost) and the energy efficiency (measured by the vehicle operational cost). They also reveal important managerial insights into UMT system operations for MAV transportation operators in the future.
These results answer the third to fifth scientific questions listed in Chapter 1.

6.1.4 Inequality Impacts of AV Systems

In Chapter 5, the proposed methodological framework is applied to study two AV scenarios and a benchmark scenario in the Tampa Bay Region in the 2040 planning year. The AV scenarios consider AVs as private vehicles with a market penetration rate of 10% and 30%, respectively. Six metrics are adopted to quantify the outcomes of the transportation system under different scenarios, including accessibility to job locations, accessibility to schools, accessibility to social activities, person miles traveled, travel cost, and travel cost to income ratio.

Compared with the base scenario, the accessibility to jobs, schools, activity locations, the person miles traveled, the travel cost, and the travel cost to income ratio are decreased in both AV scenarios. This decline can be attributed to the low AV market penetration rates and the fact that AVs operate as private vehicles, which incur many empty vehicle trips and thus an increase in the travel time. Further, the increase in the AV market penetration rate from 10% to 30% slightly increases the magnitude of the reduction in these mean outcome values.

In terms of the inequality impacts, it is found that AVs increase individual differences in the accessibility to jobs, schools, and social activity opportunities. However, the distributions of the person miles traveled, the travel cost, and the travel cost to income ratio are not substantially changed. Additionally, the disparity direction of the transportation outcome distributions between different population groups may not be changed by incorporating private AVs with low market penetration rates. Yet, the magnitude of the between group disparity is expected to change, but the direction of the change depends on the specific outcomes and the groups being analyzed.

These results inform transportation authorities how AVs will affect the performance of the transportation system and associated inequality impacts in the Tampa Bay Region. To determine
whether and how these inequality impacts can be interpreted as equity outcomes needs further studies (with possible collaborations with policy makers).

These results answer the sixth scientific question listed in Chapter 1.

6.2 Future Research Directions

6.2.1 Transmodal Systems

We have so far focused on the demand-supply asymmetry within one mode of transportation. Such an asymmetry also exists across different modes. Thus, vehicles with different capacities are dispatched to serve passengers in different modes. One may be curious about whether the multi-modal transportation service mitigates this asymmetry because different modes seem to complement one another in vehicle occupancy, spatiotemporal coverage and costs. Nevertheless, the complementarity of different modes primarily manifests in separately serving different trips without proper coordination. This lack of coordination incurs great hassle to the additional transfers between different modes, e.g., customers failing to ride to their trip destinations due to the unavailability of shared bicycles at a transit station. As a result, the effectiveness of the existing multimodal service is limited in alleviating the demand-supply asymmetry. Instead, an ideal complementarity shall take effect across different phases within one trip seamlessly served by vehicles with different capacities regardless of the modes. The MAV technologies makes such an operation possible by providing transmodal services. Specifically, in terminal neighborhoods (e.g., residence areas) or during off-peak hours with abundant transportation capacity yet sparsely distributed demand, MPs will spread apart to serve travelers individually (similar to private cars) or in small groups (similar to carpooling or microtransit) with minimum waiting time and detoured distance. Along corridors with dense traffic demand or during peak hours, the MPs will aggregate into longer vehicles (similar to mass transit buses) to increase the corridor throughput and to reduce
associated energy costs. Designing a transmodal system will be an interesting but very challenging future research direction simply because there haven’t been such systems in the real world.

6.2.2 Stochastic Demand

UMT operators typically adopt historical or predicted passenger arrival counts for system design. We follow this practice and investigate a static problem where the passenger arrival demand is known a priori. However, the passenger arrival process is affected by various factors (e.g., weather, unexpected events, and uncertain travel time) and thus exhibits great uncertainty in the real world. As a result, operational plans generated with the historical or predicted passenger arrival counts usually could not be exactly executed during operations. This phenomenon raises a few interesting scientific questions. Will the stochastic demand drift the optimal design away from the one generated with the methods proposed in this dissertation? If yes, how large is the difference between the true optimal solution and the one generated with the methods proposed here? Will traditional stochastic programming method or the emerging learning-based methods (e.g., deep reinforcement learning) be a solution to demand uncertainty? Will the theoretical properties identified in this dissertation still hold when the passenger arrival process is stochastic? How can these properties be integrated with stochastic programming or learning-based methods to design MAV systems under stochastic demand? Searching answers to these questions are the key to extend this work to incorporate stochastic demand.

6.2.3 Simulating MAVs in a Metropolitan Area

Chapter 5 in this dissertation develops an agent-based simulation model for studying the inequality impacts of AV systems. By modifying the model’s vehicle module to allow vehicle docking and undocking, the model can be extended to simulate MAV transportation systems. However, a few challenges need to be addressed to realize this functionality. First, at the system
management level, the system design models and algorithms in Chapters 2 and 3 can be integrated to determine the vehicle dispatch frequency and capacity in the simulation model. Second, at the vehicle operation level, efficient methods are needed to design how different MAVs dock and undock with each other at terminal stations or during operations in open city streets. This is essentially the vehicle trajectory control problem that have been studied in the AV literature. Third, the agent-based simulation model consumes substantial computational resources due to various factors such as the large number of individual agents, the interactions between different modes of transportation, and the complicated operational details in each mode of transportation. It would be a very important and challenging task to improve the speed of the simulation model. The use of parallel computing and distributed computing would be helpful. Despite these challenges, the development of such a model is very meaningful. It would allow us to study a range of potential impacts of the MAV technologies such as energy costs, congestion effects, and equity impacts in metropolitan areas. It could also be the foundation of developing simulation-based algorithms for citywide MAV system design where other methods (e.g., dynamic programming, continuous approximation) are not feasible.

6.2.4 Equity-aware System Design

Previous studies on transportation system design have predominantly aimed at maximizing system benefits or minimizing system costs. However, there has been evidence that emerging transportation technologies do not necessarily benefit different population groups in society equally. It is possible that the MAV transportation systems would follow this track and bring skewed benefits or costs to different population groups. A solution to this issue is to incorporate equity considerations at the system design stage so that the benefit/cost distribution of the resultant system design can be controlled in a relatively equal manner. However, there is little knowledge
on how an equity-aware system design can be achieved. First, there have been various justice theories and inequality indicators proposed in transportation equity literature. There are also several approaches to incorporation the equity considerations into system design models. Thus, from which perspective the equity performance of the proposed MAV-based elastic capacity transportation systems should be studied? Among the many inequality indicators which one(s) should be selected to quantify equity levels? These are two fundamental questions that need to be answered. Further, most of the equity indicators in the literature are formulated as non-linear functions and to mathematically formulate them in the investigated problems likely introduce extra decision variables in the models. For example, when modeling efficiency-oriented design for the shuttle systems, the passenger boarding order needs not be explicitly modelled. This is simply due to the fact that passengers’ values of time are assumed to be homogeneous, meaning that for the same queue length, it does not matter who are in the queue. However, in the equity-aware design, the boarding order has to be considered since the waiting cost of each passenger need to be known. This requirement would further complicate the system design problems. Thus, it remains an open question what computational and/or analytical tools are proper for equity-aware system design. To answer these questions, one should turn to an interdisciplinary approach involving transportation engineering, social science, operations research, and probably computer science.

We hope that the above discussion starts a dialogue for transportation researchers and practitioners to think about the many scientific and engineering questions associated with the MAV transportation systems. MAVs are promising in revolutionizing the operational paradigm of future transportation systems. A new science and engineering of MAV systems to address these questions, however, is needed before society moves forward into the MAV era.
References

AAA Association Communication (2019). Your driving costs. Retrieved from:

[Accessed 08-April 2018].


Ceder, A. A. (2001). Bus timetables with even passenger loads as opposed to even headways. Transportation Research Record: Journal of the Transportation Research Board, 3-9.


Rail, H.-S. (2017). Flexible grouped high-speed rail will be tested this year so train lengths are no longer limited to 8 or 16 carriages [Online]. Finance. [Accessed 08-April 2018].


Appendix A: Copyright Permissions

The permission from ELSEVIER for material in Chapters 1, 2, 3, 4, and 5 is below.

Copyright

Describes the rights related to the publication and distribution of research. It governs how authors (as well as their employers or funders), publishers and the wider general public can use, publish and distribute articles or books.

Journal author rights

In order for Elsevier to publish and disseminate research articles, we need publishing rights. This is determined by a publishing agreement between the author and Elsevier. This agreement deals with the transfer or license of the copyright to Elsevier and authors retain significant rights to use and share their own published articles. Elsevier supports the need for authors to share, disseminate and maximize the impact of their research and these rights, in Elsevier proprietary journals* are defined below:

<table>
<thead>
<tr>
<th>For subscription articles</th>
<th>For open access articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors transfer copyright to the publisher as part of a journal publishing agreement, but have the right to:</td>
<td>Authors sign an exclusive license agreement, where authors have copyright but license exclusive rights in their article to the publisher**. In this case authors have the right to:</td>
</tr>
<tr>
<td>• Share their article for Personal Use, Internal Institutional Use and Scholarly Sharing purposes, with a DOI link to the version of record on ScienceDirect (and with the Creative Commons CC-BY-NC-ND license for author manuscript versions)</td>
<td>• Share their article in the same ways permitted to third parties under the relevant user license (together with Personal Use rights) so long as it contains a CrossMark logo, the end user license, and a DOI link to the version of record on ScienceDirect.</td>
</tr>
<tr>
<td>• Retain patent, trademark and other intellectual property rights (including research data).</td>
<td>• Retain patent, trademark and other intellectual property rights (including research data).</td>
</tr>
<tr>
<td>• Proper attribution and credit for the published work.</td>
<td>• Proper attribution and credit for the published work.</td>
</tr>
</tbody>
</table>
Quick definitions

Personal use
Authors can use their articles, in full or in part, for a wide range of scholarly, non-commercial purposes as outlined below:

- Use by an author in the author's classroom teaching (including distribution of copies, paper or electronic)
- Distribution of copies (including through e-mail) to known research colleagues for their personal use (but not for Commercial Use)
- Inclusion in a thesis or dissertation (provided that this is not to be published commercially)
- Use in a subsequent compilation of the author's works
- Extending the Article to book-length form
- Preparation of other derivative works (but not for Commercial Use)
- Otherwise using or re-using portions or excerpts in other works

These rights apply for all Elsevier authors who publish their article as either a subscription article or an open access article. In all cases we require that all Elsevier authors always include a full acknowledgement and, if appropriate, a link to the final published version hosted on Science Direct.
Appendix B: Proofs of Lemmas and Propositions

This appendix includes the proofs for the lemmas and theorems.

**Lemma 3.1.** For $f_i$ satisfying concave property (3.1), we obtain $f_{i_1} + f_{i_2} + \cdots + f_{i_n} \geq f_{i_1 + i_2 + \cdots + i_n}, \forall i_1, i_2, \cdots, i_n, i_1 + i_2 + \cdots + i_n \in \mathcal{I}, n \in \mathbb{Z}^+.$

**Proof:** Property (3.1) yields $\frac{i}{j} f_i + \frac{j-l}{j} f_{i+l} \leq f_j, \forall i \leq j, i + j \in \mathcal{I},$ which yields $f_i + f_j \geq \frac{i+j}{j} f_i + \frac{j-i}{j} f_j.$ Further, $\frac{i}{i+j} f_{i+j} = \frac{i}{i+j} f_{i+j} + \frac{j}{i+j} f_0 \leq f_i,$ which together with the previous equation yields $f_i + f_j \geq f_{i+j}.$ Then this lemma can be easily proved by iteratively applying this relationship; i.e., $f_{i_1} + f_{i_2} + \cdots + f_{i_n} \geq f_{i_1 + i_2} + f_{i_3} + \cdots + f_{i_n} \geq \cdots \geq f_{i_1 + i_2 + \cdots + i_n}.$ This completes the proof. □

**Lemma 3.2.** For $f_i$ satisfying concave property (3.1), we obtain $f_{i_1} + f_{i_2} + f_{i_3}, \forall i_1 \leq i_2 \leq i_3 \leq i_4 \in \mathcal{I}$ and $i_2 + i_3 = i_1 + i_4.$

**Proof:** Based on property (3.1), $\frac{i_4-i_2}{i_4-i_1} f_{i_1} + \frac{i_2-i_1}{i_4-i_1} f_{i_4} \leq f_{i_2}$ and $\frac{i_4-i_3}{i_4-i_1} f_{i_1} + \frac{i_3-i_1}{i_4-i_1} f_{i_4} \leq f_{i_3}.$ This yields $\frac{2i_4-i_2-i_3}{i_4-i_1} f_{i_1} + \frac{i_2-i_3-2i_1}{i_4-i_1} f_{i_4} \leq f_{i_2} + f_{i_3}.$ Since $i_2 + i_3 = i_1 + i_4,$ the previous equation yields $f_{i_1} + f_{i_4} \leq f_{i_2} + f_{i_3}.$ This completes the proof. □

**Proposition 3.1.** In an optimal dispatch solution $\{t_k, i_k\}_{\forall k \in \mathcal{K}}$ to problem (3.2) ~ (3.8) with arrival curve $A(t)$ and departure curve $D(t)$, if $A(t_k) - D(t_{k-1}) \geq Ic$, then $i_k = I, \forall k \in \mathcal{K}.$

**Proof:** This theorem will be proven by contradiction. If the condition in this theorem does not hold, then in this optimal solution, there exists a $\bar{k} \in \mathcal{K}$ with $i_{\bar{k}} < I$ and $A(t_{\bar{k}}) - D(t_{\bar{k}-1}) \geq Ic.$ Then find $m$ such that $\sum_{m'=0}^m i_{\bar{k}+m'} < I$ and $\sum_{m'=0}^m i_{\bar{k}+m'} \geq I$, as Case 1 in Figure B.1 shows. Note that

179
since $i_k < I$, then $m \geq 1$. Further, as Case 2 in Figure B.1 shows, construct an alternate solution

$$\{t_k, i'_k\} \forall k \in \mathbb{K} \setminus \{\bar{k}+1, \ldots, \bar{k}+m-1\}$$

where $i'_k = I$, $i'_k + m = \sum_{m'=0}^m i_{k+m'} - I$ (i.e., $I + i'_k + m = \sum_{m'=0}^m i_{k+m'}$), and $i'_k = i_k$ for all other $k$ indexes. Note that if $\sum_{m'=0}^m i_{k+m'} = I$, then $i'_k + m = 0$, which means no vehicles are dispatched at time $t_{k+m}$ in the alternate solution. Denote the cumulative departure curves of Case 2 as $D'(t)$. Then the difference in the waiting cost between the optimal and alternate solutions is proportional to the shaded area with vertical lines in Figure B.1, formulated as

$$\int_{t_k}^{t_k+I+C} (A(t) - D(t)) w dt - \int_{t_k}^{t_k+I+C} (A(t) - D'(t)) w dt = \int_{t_k}^{t_k+I+C} (D'(t) - D(t)) w dt$$

$$= w \sum_{m'=0}^{m-1} \left( (t_{k+m'}+1 - t_{k+m}) \left( Ic - \sum_{m''=0}^{m'} i_{k+m''} \right) \right) > 0.$$  

Figure B.1: An illustrative example to prove Propositions 3.1 and 3.2.

Then we investigate the difference in the energy cost between the optimal and the alternate solution $\sum_{m'=0}^m f_{i_{k+m'}} - f_{I} - f'_{i_{k+m}}$. Lemma 3.1 indicates that $\sum_{m'=0}^{m-1} f_{i_{k+m'}} \geq f_{\bar{i}_k}$ where $\bar{i}_k :=$
\[
\sum_{m' = 0}^{m-1} i_{k+m'} \]. Next, Lemma 3.2 indicates that \( f_{i_k} + f_{i_{k+m}} \geq f_{i} + f_{i'_{k+m}} \). This indicates \( \sum_{m' = 0}^{m} f_{i_{k+m'}} - f_{i} - f_{i'_{k+m}} \geq 0 \). With this, the objective value of the optimal solution is always strictly greater than that of the alternate solution, which is a contradiction. This completes the proof. ∎

**Proposition 3.2.** In an optimal dispatch solution \( \{t_k, i_k\}_{\forall k \in \mathcal{K}} \) to problem (3.2) ~ (3.8) with arrival curve \( A(t) \) and departure curve \( D(t) \), if \( A(t_k) - D(t_{k-1}) > I_c \), then \( t_k - t_{k-1} = h, \forall k \in \mathcal{K} \).

**Proof.** This theorem will be proven by contradiction. If the condition in this theorem does not hold, then in this optimal solution, there exists a \( \overline{k} \in \mathcal{K} \) with \( t_{\overline{k}} - t_{\overline{k}-1} > h \) and \( A(t_{\overline{k}}) - D(t_{\overline{k}-1}) > I_c \), as Case 2 in Figure B.1 shows. With this, as Case 3 in Figure B.1 shows, we can construct an alternate feasible solution \( \{t'_{\overline{k}}, i'_{\overline{k}}\}_{\forall k \in \mathcal{K} \setminus \{\overline{k} + 1, \ldots, \overline{k} + m - 1\}} \) such that \( t'_{\overline{k}} < t_{\overline{k}}, t'_{\overline{k}} = t_k \) for all other \( k \) indexes and \( i'_{\overline{k}} = i_k \) for all \( k \) indexes. Obviously, the energy cost remains the same since the number of dispatches are exactly the same. As to the waiting cost, we only need to make a comparison in the interval \( [t_{\overline{k}-1}, t_{\overline{k}}] \). Denote the cumulative departure curve of Case 3 as \( D''(t) \), and we can obtain

\[
\int_{t_{\overline{k}-1}}^{t_{\overline{k}}} (A(t) - D'(t)) - \int_{t_{\overline{k}-1}}^{t_{\overline{k}}} (A(t) - D''(t)) \, wt = (t_{\overline{k}} - t_{\overline{k}}') I_c > 0.
\]

Thus, the objective value of the optimal solution is greater than that of the alternate solution, which is a contradiction. This completes the proof. ∎

**Lemma 3.3.** Denote problem (3.2) ~ (3.8) as the original problem (OP), and the same problem where the original arrival curve \( A(t) \) is replaced with preferred virtual arrival curve \( B(t) \) as the revised unsaturated problem (RUP). Then the feasible regions of the OP and RUP are the same. For any feasible solution \( s := \{t_k, i_k\}_{\forall k \in \mathcal{K}} \), OP and RUP have the same dispatch curve, and the objective values of OP and RUP, respectively denoted as \( OP(s) \) and \( RUP(s) \), are always
separated by a constant difference:

\[ \text{OP}(s) - \text{RUP}(s) = W(A) := w \int_0^T (A(t) - B(t)) \, dt. \]

**Proof.** First, given a feasible dispatch solution \( s := \{ t_k, i_k \}_{k \in \mathcal{K}} \) to OP, we will use induction to show that the OP and RUP have the same departure curve. Let \( D(t) \) and \( D^B(t) \) denote the departure curves corresponding to arrival curves \( A(t) \) and \( B(t) \), respectively.

We first prove the base case. When \( z = 1 \), \( A(t) = B(t) \), thus \( D(t) = D^B(t), \forall t \in [\mathcal{E}(0), \mathcal{O}(1)] \) for the same dispatch solution.

Then we conduct the induction step. Assume \( D(t) = D^B(t), \forall t \in [\mathcal{E}(0), \mathcal{O}(z)) \) for a \( 0 \leq z \leq |\mathcal{O}| \).

Then for the oversaturated period \([\mathcal{O}(z), \mathcal{E}(z))\), the facts that \( D^B(t) \leq B(t) \) and \( B'(t) = l_c/h \) yield that at any dispatch point \( t_k \in [\mathcal{O}(z), \mathcal{E}(z)), \) if exists, then \( B(t_k) - D^B(t) \geq l_c \). Thus \( D^B(t) \) always grows \( i_k c \) at \( t_k \). Since \( A(t) \geq B(t) \), similarly, we obtain that \( D(t) \) also grows to \( i_k c \) at \( t_k \).

In this way, \( D(t) = D^B(t), \forall t \in [\mathcal{E}(0), \mathcal{E}(z)) \). Further, since \( A(t) = B(t) \) during unsaturated period \([\mathcal{E}(z), \mathcal{O}(z + 1)) \) (or \([\mathcal{E}(|\mathcal{O}|), \mathcal{O}(|\mathcal{O}|]) \) for the last unsaturated period), then we obtain \( D(t) = D^B(t), \forall t \in [\mathcal{E}(0), \mathcal{O}(z + 1)] \).

The above induction proves that a feasible solution to OP is feasible to RUP and yields the same departure curve. With the same induction, we can show that a feasible solution to RUP is feasible to OP and yields the same departure curve as well. With this, the formulation of objective function (3.2) yields \( \text{OP}(s) - \text{RUP}(s) = w \int_0^T (A(t) - B(t)) \, dt. \)  \( \Box \)

**Proposition 3.4.** An optimal dispatch solution \( \{ t_k, i_k \}_{k \in \mathcal{K}} \) to RUP satisfies: (i) if \( t_k - t_{k-1} > h \), then \( B(t_k) - D(t_k) = 0, \forall k \in \mathcal{K}; \) and (ii) if \( t_k - t_{k-1} = h \), then \( B(t_k) - D(t_k) \in [0, c), \forall k \in \mathcal{K}, \) where \( D(t) \) is the corresponding departure curve.

**Proof.** We first prove \( B(t_k) - D(t_k) \in [0, c), \forall k \in \mathcal{K} \) holds for both cases by induction.
We first prove the base case. First investigate the base case with $k = 1$. When $B(t_1) - D(t_1) > 0$, if $i_1 = I$, then Proposition 3.2 indicates that $t_1 \leq h$, which contradicts to $B(t)$ being unsaturated. Therefore, $B(t_1) - D(t_1) = 0$ if $i_1 = I$. Otherwise, if $i_1 < I$ and $B(t_1) - D(t_1) \geq c$, then we can raise the first dispatch to $i_1 + 1$ and drop the second dispatch to $i_2 - 1$. Then with similar analysis in Proposition 3.1 we can show that this will always strictly reduce the objective value, which obviously contradicts to this solution being optimal. Therefore $B(t_1) - D(t_1) \in [0, c)$ holds for the base case.

![Diagram](image.png)

Figure B.2: An illustrative example to prove Proposition 3.4.

Then we conduct the induction step. Assume that for a $k \in \mathcal{K}\{K\}$, $B(t_k) - D(t_k) \in [0, c)$. Then at time $t_{k+1}$, when $B(t_{k+1}) - D(t_{k+1}) > 0$, if $i_k = I$, Proposition 3.2 indicates that $t_{k+1} - t_k = h$, and thus $B(t_{k+1}) - D(t_{k+1}) = [B(t_k) - D(t_k)] + [B(t_{k+1}) - B(t_k)] - Ic < c + [B(t_{k+1}) - B(t_k)] - Ic$ (with the induction assumption) $< c$ (since $B(t)$ is unsaturated any time). Next, consider the case when $i_{k+1} < I$. If $B(t_{k+1}) - D(t_{k+1}) \geq c$, then $k + 1 < |\mathcal{K}|$ due to constraints (3.6). Then with similar analysis in Proposition 3.1, we can show that raising $i_{k+1}$ to $i_{k+1} + 1$ and dropping $i_{k+2}$ to $i_{k+2} - 1$ will always strictly reduce the objective value, which obviously contradicts to this solution being optimal. Therefore, $B(t_{k+1}) - D(t_{k+1}) \in [0, c)$ also holds.
The above induction shows that \( B(t_k) - D(t_k) \in [0, c) \) holds for both cases in an optimal solution.

Next we further prove that \( B(t_k) - D(t_k) = 0, \forall k \in \mathcal{K} \) holds if \( t_k - t_{k-1} > h \) by contradiction.

If the condition does not hold, then in this optimal solution, there exists a \( \tilde{k} \in \mathcal{K} \) with \( t_{\tilde{k}} - t_{\tilde{k}-1} > h \) and \( B(t_{\tilde{k}}) - D(t_{\tilde{k}}) > 0 \). Then, we can construct an alternate solution \( \{t'_k, i'_k\}_{\forall k \in \mathcal{K}} \) such that \( t'_k < t_{\tilde{k}}, t'_k = t_k \) for all other \( k \) indexes and \( i'_k = i_k \) for all \( k \) indexes, and there are two cases. If \( t'_k - t'_{k-1} > h \) can be satisfied, as Figure B.2 (a) shows, we can always find a \( t'_k \) such that \( B(t'_k) - D(t'_k) = 0 \). With similar analysis in Theorem 2, we can show that this will always strictly reduce the objective value, which obviously contradicts to this solution being optimal. Otherwise, if \( t'_k - t'_{k-1} = h \), as Figure B.2 (b) shows, this reduces to the first case and we have shown that \( B(t_k) - D(t_k) \in [0, c), \forall k \in \mathcal{K} \) holds if \( t_k - t_{k-1} = h \).

This completes the proof. \( \square \)

**Proposition 4.1.** An optimal solution \( \{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\} \) to problem (4.3) ~ (4.11) with time-dependent arrival rate curves \( A_s(t', t), \forall s \in S \) and time-dependent departure rate curves \( D_s(t', t), \forall s \in S \) must satisfy one of the following conditions for each dispatch \( k \in \mathcal{K} \): (i) \( t_k - t_{k-1} = h \); or (ii) there exists at least one \( s \in S \) such that \( q_s(t_k) = 0 \).

**Proof.** This proposition will be proven by contradiction. If the condition in this proposition does not hold, then in this optimal solution, there exists a \( \tilde{k} \in \mathcal{K} \) with \( t_{\tilde{k}} - t_{\tilde{k}-1} > h \) and \( q_s(t_{\tilde{k}}) > 0, \forall s \in S \). Then we can construct an alternate solution \( \{\tilde{K}, \tilde{t}_k, \tilde{i}_{ks}, \tilde{e}_{ks}, \tilde{d}_{ks}(t'), \tilde{D}_s(t', t)\} \) with the following steps: (i) keep the number of dispatches and vehicle formations the same as those in the optimal solution, i.e., \( \tilde{K} = K, \tilde{i}_{ks} = i_{ks}, \forall k \in \mathcal{K}, s \in S \); (ii) push the dispatch time of \( \tilde{k} \) forward until either \( t_{\tilde{k}} - t_{\tilde{k}-1} = h \) or we find an \( \tilde{s} \in S \) such that \( \tilde{q}_{\tilde{s}}(t_{\tilde{k}}) = 0 \) while keep \( \tilde{t}_k = t_k \) for all

184
other $k$ indexes. Note that with this construction process, the resultant alternate solution obviously satisfies Constraints (4.4).

Since all decisions before dispatch $\bar{k}$ in the alternate solution are the same as those in the optimal solution, we have $\hat{e}_{ks} = e_{ks}, \forall k < \bar{k}, s \in S, \hat{d}_{ks}(t') = d_{ks}(t'), \forall k < \bar{k}, s \in S, t' \in [0, T],$

$$\bar{D}_s(t', t) = D_s(t', t), \forall s \in S, t' \in [0, T], t \in [0, \hat{t}_\bar{k}).$$

Next, we solve $\hat{e}_{ks}, \hat{d}_{ks}(t'v)$, and $\bar{D}_s(t', t)$ for time indexes since dispatch $\bar{k}$ in the alternate solution. Based on how we construct the alternate solution, we have

$$\int_{t'=0}^{T} (A_1(t', \hat{t}_\bar{k}) - D_1(t', \hat{t}_{\bar{k}-1}))dt' = \int_{t'=0}^{\hat{t}_\bar{k}} (A_1(t', \hat{t}_\bar{k}) - D_1(t', \hat{t}_{\bar{k}-1}))dt' \geq i_{\bar{k}1}c$$

This inequality together with Eqs.(4.8) and (4.12) ($k$ replaced by $\bar{k}$) yield $\hat{e}_{\bar{k}1} = i_{\bar{k}1}c$. We apply this result into Eq.(4.7) and then solve it with $\bar{D}_s(t', t) = D_s(t', t), \forall s \in S, t' \in [0, T], t \in [0, \hat{t}_\bar{k})$ (proved above), Eqs.(4.9), and (4.10) (with $k = \bar{k}$), which results in $\hat{d}_{\bar{k}1}(t') = d_{\bar{k}1}(t'), \forall t' \in [0, T]$. Applying this equation into Eq.(4.5) with $k = \bar{k}$, we further obtain $\bar{D}_1(t', \hat{t}_\bar{k}) = D_1(t', \hat{t}_{\bar{k}-1}) + d_{\bar{k}1}(t'), \forall t' \in [0, T]$. Next, at Station 2, again, based on the way we construct the alternate solution, we obtain

$$\hat{e}_{\bar{k}1} - \int_{t'=0}^{\hat{t}_\bar{k}} \hat{d}_{\bar{k}1}(t')p_{12}(t')dt' + \int_{t'=0}^{\hat{t}_\bar{k}} (A_2(t', \hat{t}_\bar{k}) - D_2(t', \hat{t}_{\bar{k}-1}))dt' \geq i_{\bar{k}2}c.$$

With the same analysis used above we obtain $\hat{e}_{\bar{k}2} = i_{\bar{k}2}c, \hat{d}_{\bar{k}2}(t') = d_{\bar{k}2}(t'), \forall t' \in [0, T]$, and $\bar{D}_2(t', \hat{t}_\bar{k}) = D_2(t', \hat{t}_{\bar{k}-1}) + d_{\bar{k}2}(t'), \forall t' \in [0, T]$. Applying this analysis over Stations 3 to $S$, we obtain $\hat{e}_{ks} = i_{ks}c, \forall s \in S$, $\hat{d}_{ks}(t') = d_{ks}(t'), \forall t' \in [0, T]$, and $\bar{D}_s(t', \hat{t}_\bar{k}) = D_s(t', \hat{t}_{\bar{k}-1}) + d_{ks}(t'), \forall t' \in [0, T], \forall s \in S$. Further, since no MAVs are dispatched until $\hat{t}_{\bar{k}+1}$, we obtain from Constraints (4.5) that

$$\bar{D}_s(t', t) = \bar{D}_s(t', \hat{t}_\bar{k}) = D_s(t', \hat{t}_{\bar{k}-1}) + d_{ks}(t'), \forall s \in S, t' \in [0, T], t \in [\hat{t}_\bar{k}, t_{\bar{k}}].$$
Then, because $\hat{t}_k = t_k, \hat{i}_{ks} = i_{ks}, \forall k > \bar{k}, s \in S$, we obtain $\hat{e}_{ks} = e_{ks}, \forall k > \bar{k}, s \in S, \hat{d}_{ks}(t') = d_{ks}(t'), \forall k > \bar{k}, s \in S, t' \in [0, T]$, and
\[
\hat{D}_s(t', t) = D_s(t', t), \forall s \in S, t' \in [0, T], t \in (t_k, T).
\]
These results indicate that the alternate solution also satisfies Constraints (4.5) - (4.11). Therefore, the alternate solution is feasible to the investigated problem.

Regarding the change in the objective value, the operational cost of the alternate solution is obviously the same as that of the optimal solution since $\hat{i}_{ks} = i_{ks}, \forall k \in \mathcal{K}, s \in S$. Further, the difference between the passenger waiting cost of the optimal solution and that of the alternate solution is
\[
w \left[ \sum_{s \in S} \int_{t'=0}^{T} \int_{t=0}^{T} (A_s(t', t) - D_s(t', t)) dt \ dt' - \sum_{s \in S} \int_{t'=0}^{T} \int_{t=0}^{T} (A_s(t', t) - \hat{D}_s(t', t)) dt \ dt' \right] 
\]
\[
= w \left[ \sum_{s \in S} \int_{t'=0}^{T} \int_{t=0}^{T} (\hat{D}_s(t', t) - D_s(t', t)) dt \ dt' \right] 
= w \left[ \sum_{s \in S} \int_{t'=0}^{t_{k'}} \int_{t=0}^{T} (d_{ks}(t')) dt \ dt' \right] 
= w(t_k - \hat{t}_k) \left[ \sum_{s \in S} \int_{t'=0}^{T} (d_{ks}(t')) dt' \right] > 0.
\]
Thus, the objective value of the alternate solution is always strictly smaller than that of the optimal solution, which is a contradiction. This completes the proof. \[\square\]

**Proposition 4.2.** Any dispatch $k$ in an optimal solution $\{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\}$ to problem (4.3) ~ (4.11) with time-dependent arrival rate curves $A_s(t', t), \forall s \in S$ and departure rate curves $D_s(t', t), \forall s \in S$ must satisfy $\tilde{q}_s(t_k) < c, \forall s \in S$ if the following conditions hold:

(i) $S(f_2 - f_1) < cwh_1$.

(ii) $\sum_{u \in S_{s+1}} \sum_{v \in S_s} \left( \int_{t' = t_{k-1}}^{t_k} (A_u(t', t')) p_{uv}(t') dt' \right) \leq (l - 1)c, \forall s \in S$.

**Proof.** This proposition will be proven with induction.

186
We first prove the base case. First investigate the case with \( k = 1 \). If there exists a \( \overline{s} \in S \) such that \( \overline{q}_{\overline{s}}(t_1) \geq c \), we define the first origin station where there are passengers in cross-sectional queue \( \overline{q}_{\overline{s}}(t_1) \) as

\[
s_1 := \arg\min_{u < \overline{s} \in S} \left\{ \sum_{v \in S_{u+1}} \left( \int_{t' = 0}^{t_1} (A_u(t', t_1) - D_u(t', t_1)) p_{uv}(t') dt' \right) > 0 \right\}
\]

and the last destination station which passengers in cross-sectional queue \( \overline{q}_{\overline{s}}(t_1) \) travel to as

\[
s_2 := \arg\min_{v > \overline{s} \in S} \left\{ \sum_{u \in S_{v-1}} \left( \int_{t' = 0}^{t_1} (A_u(t', t_1) - D_u(t', t_1)) p_{uv}(t') dt' \right) > 0 \right\}
\]

Then, we construct an alternate solution \( \{ \hat{R}, \hat{t}_k, \hat{i}_{ks}, \hat{d}_{ks}(t'), \hat{D}_s(t', t) \} \) where \( \hat{R} = K, \hat{t}_k = t_k, \forall k \in K, \hat{i}_{ks} = i_{ks}, \forall k > 1, s \in S \) and

\[
\hat{i}_{1s} = \begin{cases} i_{1s} + 1 & \text{if } s \in [s_1, s_2] \\ i_{1s} & \text{otherwise} \end{cases}, \forall s \in S.
\]

That is, at the first dispatch, we increase the number of modular pods by 1 between \( s_1 \) and \( s_2 \) (inclusive) while keep the vehicle formations at other stations the same as those in the optimal solution. With condition (ii), we obtain \( \hat{i}_{1s} \leq I - 1, \forall s \in S \). Further, since the optimal solution satisfies Constraints (4.6) (i.e., all passengers are transported at \( T \)), the alternate solution increases the vehicle’s capacity at the first dispatch and thus obviously also satisfies Constraints (4.6). Thus, the constructed vehicle formations are always feasible to the investigated problem.

To analyze the change in the objective value, we first solve variables \( \hat{e}_{1s}, \forall s \in S, \hat{d}_{1s}(t'), \forall s \in S, t' \in [0, T], \) and \( D_s(t', t), \forall s \in S, t' \in [0, T], t \in [0, t_2] \). Based on the way we construct the alternate solution, these variables can be computed for three types of stations separately as follows.
(i) $s \in [1, s_1) \cup (s_2, S]$. Because $\hat{t}_k = t_k, i_{ks} = i_{ks}, \forall k \in K, s \in [1, s_1) \cup (s_2, S]$ and the cross-sectional queue at station $\bar{s}$ does not affect these stations, it is obvious that $\hat{e}_{1s} = e_{1s}, \forall s \in [1, s_1) \cup (s_2, S], \hat{d}_{1s}(t') = d_{1s}(t'), \forall s \in [1, s_1) \cup (s_2, S], t' \in [0, T]$, and

$$
\hat{D}_s(t', t) = D_s(t', t), \forall s \in [1, s_1) \cup (s_2, S], t' \in [0, T], t \in [0, t_2).
$$

(ii) $s \in [s_1, \bar{s}]$. Since $\hat{q}_s(t_1) < c, \forall s \in [s_1, \bar{s})$ in the optimal solution, all passengers in the cross-sectional queue at station $s \in [s_1, \bar{s})$ can be accommodated by increasing the number of modular pods at this station by 1, i.e., $\hat{e}_{1s} = e_{1s} + \hat{q}_s(t_1), \forall s \in [s_1, \bar{s})$ (Constraints (4.8) and (4.12)). With similar analyses we obtain $\hat{e}_{1\bar{s}} = e_{1\bar{s}} + c$. To further compute $d_{ks}(t')$ and $D_s(t', t)$, we define

$$
t^1_{1s} := \arg\inf_{t' \in [0, T]} \{ A_s(t', t_1) - D_s(t', t_1) > 0 \}, \forall s \in S
$$

to represent the smallest arrival time index of the passengers at station $s$ who are not served by the first dispatch in the optimal solution (also the arrival time index of the first passenger boarding the additional modular pods at dispatch 1 in the alternate solution based on Constraints (4.10)). With this, we solve the arrival time index of the last passenger who board the additional modular pods at dispatch 1 at station $s$ in the alternate solution as

$$
t^2_{1s} := \arg\inf_{t'' \in [t^1_{1s}], T} \left\{ \int_{t' = t^1_{1s}}^{t''} (A_s(t', t_1) - D_s(t', t_1)) dt' \geq o_{1s} \right\}, \forall s \in S
$$

where

$$
o_{1s} := \hat{e}_{1s} - \sum_{u \in \delta_s^+} \sum_{v \in \delta_s^+} \left( \int_{t' = 0}^{t_1} (\hat{d}_{1u}(t') p_{uv}(t')) dt' \right)
$$

$$
- \left( e_{1s} - \sum_{u \in \delta_s} \sum_{v \in \delta_s^+} \left( \int_{t' = 0}^{t_1} (d_{1u}(t') p_{uv}(t')) dt' \right) \right), \forall s \in S
$$
represents the number of passengers boarding the additional modular pods at dispatch 1 at \( s \). Then with similar analysis in Proposition 4.1, we obtain

\[
\hat{d}_{1s}(t') = \begin{cases} 
\frac{d_{2s}(t')}{d_{1s}(t')} & \text{if } t' \in [t_{1s}^1, t_{1s}^2], \forall s \in [s_1, \overline{s}], t' \in [0, T]. \\
\text{otherwise} & \end{cases}
\]

\[
\hat{D}_s(t', t) = \begin{cases} 
\frac{D_s(t', t_2)}{D_s(t', t)} & \text{if } t' \in [t_{1s}^1, t_{1s}^2], t \in [t_1, t_2), \forall s \in [s_1, \overline{s}], t' \in [0, T], t \in [0, t_2). \\
\text{otherwise} & \end{cases}
\]

(iii) \( s \in (\overline{s}, s_2] \). Similar to type-2 stations, the cross-sectional queue at this type of stations can also be cleared in the alternate solution by a single modular pod. Thus, we have

\[
\hat{e}_{1s} = e_{1s} + \sum_{u \in S} \sum_{v \in S} \left( \int_{t'}^{t} \left( \hat{d}_{1u}(t') p_{uv}(t') \right) dt' \right) + q_s(t_1), s \in (\overline{s}, s_2].
\]

Then with the same analysis at type-2 stations, we obtain that the analytical formulas of \( \hat{d}_{1s}(t') \) and \( \hat{D}_s(t', t) \) for \( s \in [s_1, \overline{s}] \) also hold for \( s \in (\overline{s}, s_2] \).

From the above discussion, we have that \( \hat{D}_s(t', t_1) \geq D_s(t', t_1), \forall s \in S, t' \in [0, T] \). Since \( \hat{k} = t_k, \hat{s}_k = i_{ks}, \forall k > 1, s \in S \), we obtain \( \hat{D}_s(t', t) \geq D_s(t', t), \forall s \in S, t' \in [0, T], t \in (t_1, T] \) based on Constraints (4.7) - (4.9) (i.e., the same amount of passengers will be transported at each dispatch, if not more, in the alternate solution due to the increased capacity at dispatch 1).

With this, based on condition (i) in this proposition, we obtain that the difference between the objective function of the optimal and alternate solution is

\[
\Delta C_f + \Delta C_w \geq S(f_1 - f_2) + cwh > 0.
\]

This forms a contradiction. We repeat this process until \( \bar{q}_s(t_1) < c \) is satisfied. Thus \( \bar{q}_s(t_1) < c, \forall s \in S \) must hold for the base case.

Then we conduct the induction step. Assume that for a \( k \in K \setminus \{K\}, \bar{q}_s(t_k) < c, \forall s \in S \). At time \( t_{k+1} \), if there exists a \( \overline{s} \in S \) such that \( \bar{q}_{\overline{s}}(t_{k+1}) \geq c \), then similar to the base case, we can construct an alternate solution \( \{R, \hat{t}_k, i_{ks}, \hat{e}_{ks}, d_{ks}(t'), \hat{D}_s(t', t)\} \) by raising the number of modular
Proof. Proposition 4.4. \( \sum_{\mathcal{S}} q_s(t_{k+1}) < c, \forall s \in \mathcal{S} \) also holds. This completes the proof. \( \square \)

**Proposition 4.3.** Any dispatch \( k \) in an optimal solution \( \{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\} \) to problem (4.3) \( \sim \) (4.11) with time-dependent arrival rate curves \( A_s(t', t), \forall s \in \mathcal{S} \) and departure rate curves \( D_s(t', t), \forall s \in \mathcal{S} \) must satisfy \( q_s(t_k) < c, \forall s \in \mathcal{S} \) if the following conditions hold:

(i) \( S(f_2 - f_1) < c w_h \)

(ii) \( \sum_{u \in \mathcal{S}^{s+1}} \sum_{v \in \mathcal{S}^s} \left( \int_{t' = t_k}^{t_k} (A_u(t', t')) p_{uv}(t') dt' \right) \leq (l - 1)c, \forall s \in \mathcal{S} \).

**Proof.** Based on Proposition 4.2, \( q_1(t_k) = \bar{q}_1(t_k) < c \). Further, with the definitions of \( \bar{q}_s(t_k) \) and \( q_s(t_k) \), we obtain

\[
\bar{q}_2(t_k) = \int_{t' = 0}^{t_k} (A_1(t', t_k) - D_1(t', t_k)) p_{uv}(t') dt' + q_2(t_k) < q_1(t_k) + q_2(t_k) < c,
\]

which together with \( 0 \leq q_1(t_k) < c \) yields \( q_2(t_k) < c \). Then this proposition can be easily proved by iteratively applying this relationship; i.e., \( \bar{q}_s(t_k) < q_1(t_k) + q_2(t_k) + \ldots + q_s(t_k) < c \). This completes the proof. \( \square \)

**Proposition 4.4.** Any dispatch \( k \) in an optimal solution \( \{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\} \) to problem (4.3) \( \sim \) (4.11) with time-dependent arrival rate curves \( A_s(t', t), \forall s \in \mathcal{S} \) and departure rate curves \( D_s(t', t), \forall s \in \mathcal{S} \) must satisfy \( i_{ks} \leq i_{ks}, \forall s \in \mathcal{S} \) if the following conditions hold:

(i) \( S(f_2 - f_1) < c w_h \)

(ii) \( \sum_{u \in \mathcal{S}^{s+1}} \sum_{v \in \mathcal{S}^s} \left( \int_{t' = t_k}^{t_k} (A_u(t', t')) p_{uv}(t') dt' \right) \leq (l - 1)c, \forall s \in \mathcal{S} \).
Proof. This proposition must hold since otherwise Proposition 4.3 will be violated. This completes the proof. □

Proposition 4.5. An optimal solution \( \{K, t_k, i_{ks}, e_{ks}, d_{ks}(t'), D_s(t', t)\} \) to problem (4.3) \~ (4.11) must satisfy \( i_{ks} \leq \bar{i}_{ks}, \forall k \in \mathcal{K}, s \in \mathcal{S} \).

Proof. This proposition will also be proven by contradiction. If the condition in this proposition does not hold, then there exist \( \bar{k} \in \mathcal{K} \) and \( \bar{s} \in \mathcal{S} \) such that \( i_{ks} > \bar{i}_{ks} \) in this optimal solution. Note that \( \bar{i}_{ks} < 1 \) (i.e., \( \bar{i}_{ks} = \left\lceil \frac{\bar{a}_{ks}}{c} \right\rceil \)) since otherwise this optimal solution will not be feasible. Then we can construct an alternate solution \( \{\bar{K}, \bar{t}_k, \bar{i}_{ks}, \bar{e}_{ks}, \bar{d}_{ks}(t'), \bar{D}_s(t', t)\} \) where \( \bar{K} = K, \bar{t}_k = t_k, \forall k \in \mathcal{K}, \bar{i}_{ks} = \bar{i}_{ks}, \) and \( \bar{i}_{ks} = i_{ks} \) for all other \( k \) and \( s \) indexes. Since \( \bar{i}_{ks} = \left\lceil \frac{\bar{a}_{ks}}{c} \right\rceil \), we obtain \( i_{ks} c > \bar{i}_{ks} c \geq \bar{a}_{ks} \), meaning that both \( \bar{i}_{ks} \) and \( i_{ks} \) can transport all passengers who want to board dispatch \( \bar{k} \) at station \( \bar{s} \). Therefore, \( \bar{e}_{ks} = e_{ks}, \forall k \in \mathcal{K}, s \in \mathcal{S}, \bar{d}_{ks}(t') = d_{ks}(t'), \forall k \in \mathcal{K}, s \in \mathcal{S}, t' \in [0, T], \bar{D}_s(t', t) = D_s(t', t), \forall k \in \mathcal{K}, s \in \mathcal{S}, t' \in [0, T], t \in [0, T]. \) Thus, the constructed alternate solution is feasible to the investigated problem. Further, this result also indicates that the passenger waiting costs are the same between the alternate solution and optimal solution. Thus, the difference between the objective value of the optimal solution and the alternate solution is

\[
\sum_{s \in \mathcal{S}, k \in \mathcal{K}} (f_{i_{ks}} - f_{i_{ks}}) = f_{(i_{ks})} - f_{(i_{ks})} > 0
\]

which forms a contradiction. This completes the proof. □

Proposition 4.8. A feasible solution \( \{x_{ij}, b_{j's}, y_{j's}, r_{js}\} \) to problem (4.22) \~ (4.36) is not optimal if there exists \( \tilde{j} < j \) such that: (i) \( \sum_{l \in J} x_{il} = 1, \sum_{l \in J} x_{ij} = 1, \sum_{j'' \in (j, J)} \sum_{l \in J} x_{ij''} = 0; \) (ii) \( j - \tilde{j} > j; \) (iii) \( a_{(j-1)s} \geq c \sum_{l \in J} i_{x_{ij}} - \sum_{j'=0}^j \sum_{u \in \bar{S}'} \sum_{v \in \bar{S}'} \left( a_{j'u} p_{j'u} (y_{j'u} - y_{j'_{(j-1)u}}) \right), \forall s \in \mathcal{S} \).

Proof. This proposition will be proven by contradiction. If there is such an optimal solution, then
we can construct an alternate solution \( \{ \hat{x}_{ijs}, \hat{b}'_{j's}, \hat{y}'_{j's}, \hat{r}_{js} \} \) such that \( \hat{x}_{i(j-1)s} = x_{ijs}, \hat{x}_{ijs} = 0, \forall i \in \mathcal{I}, s \in \mathcal{S} \), and \( \hat{x}_{ijs} = x_{ijs} \) for other combinations of \( i, j, s \) indexes. Since \( j - \tilde{j} \geq j \), we have \( j - 1 - \tilde{j} \geq j \) and thus Constraints (4.22) and (4.23) still hold. Further, since \( \hat{x}_{ijs} = x_{ijs}, \forall i \in \mathcal{I}, j < \tilde{j} - 1, s \in \mathcal{S} \), we obviously obtain

\[
\hat{b}'_{j's} = b'_{j's}, \hat{y}'_{j's} = y'_{j's}, \hat{r}_{js} = r_{js}, \forall j' \in \mathcal{J}, j \in \{j', \ldots, \tilde{j} - 2\}, s \in \mathcal{S}.
\]

Because \( q_{(j-1)s} \geq c \sum_{i \in \mathcal{I}} x_{ijs} - \sum_{j' = 0}^{j} \sum_{u \in \mathcal{S}^-_i} \sum_{v \in \mathcal{S}_u^+} (a_{j'u} p_{j'uv} (y'_{j'u} - y'_{(j-1)u})) \), with Constraints (4.24) – (4.30) we obtain that

\[
\hat{b}'_{j'(j-1)s} = b'_{j'(j-1)s}, \hat{y}'_{j'(j-1)s} = y'_{j'(j-1)s}, \hat{r}_{js} = r_{js}, \forall j' \in \mathcal{J}, s \in \mathcal{S}.
\]

Next, since \( \hat{x}_{ijs} = 0, \forall i \in \mathcal{I}, s \in \mathcal{S} \), and \( \hat{x}_{ijs} = x_{ijs}, \forall i \in \mathcal{I}, j > \tilde{j}, s \in \mathcal{S} \), we obtain

\[
\hat{b}'_{j's} = b'_{j's}, \hat{y}'_{j's} = y'_{j's}, \hat{r}_{js} = r_{js}, \forall j' \in \mathcal{J}, j \in \{j, \ldots, \tilde{j}\}, s \in \mathcal{S}.
\]

Thus, the constructed alternate solution is a feasible solution because it also satisfies Constraints (4.24) – (4.35). Furthermore, the difference between the objective values of the optimal solution and the alternate solution is

\[
\sum_{s \in \mathcal{S}} \left[ w_\delta \sum_{j' = 0}^{j-1} (\hat{y}'_{j'(j-1)s} - y'_{j'(j-1)s}) \right] = \sum_{s \in \mathcal{S}} \left[ w_\delta \sum_{j' = 0}^{j-1} (a_{s j'} (y'_{j's} - y'_{j'(j-1)s})) \right] > 0.
\]

Note that there must exist \( j' \in \mathcal{J} \) such that \( y'_{j's} - y'_{j'(j-1)s} > 0, s \in \mathcal{S} \) since otherwise Constraints (4.30) will be violated. Thus, the objective value of the alternate solution is strictly lower than that of the optimal solution, which is a contradiction. This completes the proof. \( \square \)

**Proposition 4.9.** Given two feasible solutions \( o := \{x_{ijs}, b'_{j's}, y'_{j's}, r_{js} \} \) and \( \hat{o} := \{\hat{x}_{ijs}, \hat{b}'_{j's}, \hat{y}'_{j's}, \hat{r}_{js} \} \) (all variables relevant to \( \hat{o} \) will be denoted with a hat accent in the following analysis) to problem (4.22) – (4.36), we say \( o \) dominates \( \hat{o} \) if there exists \( j' \in \mathcal{J} \) such that: (i) \( h_j \geq \hat{h}_{j'} \); (ii) \( \bar{a}_{js} \leq \bar{\hat{a}}_{js}, \forall s \in \mathcal{S} \); and (iii) \( z_j(o) < z_j(\hat{o}) \).
Proof. This proposition obviously holds for $j = J$. For $j < J$, we will prove that $\hat{o}$ cannot be optimal by contradiction. To this end, we construct another solution $\bar{o} := \{\bar{x}_{ijs}, \bar{b}_{j's}, \bar{y}_{j's}, \bar{r}_s\}$ (variables relevant to this solution will be denoted with a bar accent) where

$$\bar{x}_{ijs} = \begin{cases} x_{ijs} & \text{if } j \leq \bar{j}, \forall i, j, s \in S, \\ \bar{x}_{ijs} & \text{if } j > \bar{j}, \forall i, j, s \in S. \end{cases}$$

Based on the way solution $\bar{o}$ is constructed, we obviously have

$$\bar{b}_{j's} = b_{j's}, \bar{y}_{j's} = y_{j's}, \bar{r}_s = r_s, \forall j', j \in J, j \in \{j', \ldots, \bar{j}\}, s \in S.$$  

Next we solve these variables for $j > \bar{j}$. Let $j_1 := \arg\max_{j'' \in \{0, \ldots, \bar{j}\}} \{\sum_{i \in \mathcal{I}} x_{i j''} = 1\}$ (time index of the last dispatch before $\bar{j}$ in $\bar{o}$), $\bar{j}_1 := \arg\max_{j'' \in \{0, \ldots, \bar{j}\}} \{\sum_{i \in \mathcal{I}} \bar{x}_{i j''} = 1\}$ (time index of the last dispatch before $\bar{j}$ in $\bar{o}$), and $\hat{j}_1 := \arg\min_{j'' \in \{0, \ldots, \bar{j}\}} \{\sum_{i \in \mathcal{I}} \bar{x}_{i j''} = 1\}$ (time index of the first dispatch after $\bar{j}$ in $\bar{o}$). Since $\bar{o}$ is feasible (it must satisfy Constraints (4.22)), and thus we have

$$\hat{j}_2 \sum_{i \in \mathcal{I}} \hat{x}_{i j_2 - 1} - j_1 \sum_{i \in \mathcal{I}} \hat{x}_{i j_1 - 1} = \hat{j}_2 - \hat{j}_1 = (\hat{j}_2 - \bar{j}) + (\bar{j} - \hat{j}_1)$$

$$= \begin{cases} (\hat{j}_2 - \bar{j}) + \hat{h}_j & \text{if } \bar{j} - \hat{j}_1 \leq \bar{j}, \\ (\hat{j}_2 - \bar{j}) + \hat{h}_j + (\bar{j} - \bar{j}_1 - 1) & \text{if } \bar{j} - \hat{j}_1 > \bar{j}. \end{cases}$$

In the constructed solution $\bar{o}$, we have

$$\hat{j}_2 \sum_{i \in \mathcal{I}} \bar{x}_{i j_2 - 1} - j_1 \sum_{i \in \mathcal{I}} x_{i j_1 - 1} = \hat{j}_2 - j_1 = (\hat{j}_2 - \bar{j}) + (\bar{j} - j_1)$$

$$= \begin{cases} (\hat{j}_2 - \bar{j}) + h_j & \text{if } \bar{j} - j_1 \leq \bar{j}, \\ (\hat{j}_2 - \bar{j}) + h_j + (\bar{j} - j_1 - 1) & \text{if } \bar{j} - j_1 > \bar{j}. \end{cases}$$

If $\hat{j}_2 - j_1 = (\hat{j}_2 - \bar{j}) + h_j$ and $\hat{j}_2 - j_1 = (\hat{j}_2 - \bar{j}) + \bar{h}_j$, we obtain $\hat{j}_2 - j_1 = \bar{j}_2 - j_1 \geq j$ since $h_j \geq \bar{h}_j$(condition (i)). If $\hat{j}_2 - j_1 = (\hat{j}_2 - \bar{j}) + \bar{h}_j$, we obtain
\( j_2 - j \geq 0 \) and \( h_j \geq j \) (because \( \hat{h}_j = j \), and \( h_j \geq \hat{h}_j \)), and thus \( j_2 - j_1 \geq j \). If \( j_2 - j_1 = (j_2 - j) + h_j + (j - j_1 - j) \), we obtain \( j_2 - j \geq 0 \), \( h_j = j \) and \( j - j_1 - j > 0 \) (since \( j - j_1 > j \)), and thus \( j_2 - j_1 \geq j \). Thus, \( j_2 - j_1 \geq j \) always holds, indicating that the constructed solution \( \bar{o} \) satisfies Constraints (4.22) and (4.23). Further, since \( \bar{a}_{js} \leq \hat{a}_{js}, \forall s \in S \) (condition (ii)), we have \( \bar{a}_{js} \leq \hat{a}_{js} \), \( \forall s \in S \). Combining Constraints (4.24) – (4.26), (4.30), and Eq. (4.37), we obtain

\[
\hat{q}_{js} = \hat{a}_{js} - \hat{r}_{js}, \bar{q}_{js} = a_{js} - \bar{r}_{js}, \forall s \in S.
\]

If \( \bar{r}_{js} = \hat{a}_{js} \), then \( \hat{a}_{js} \leq c \sum_{i \in J} \bar{x}_{ij}a_{js}, \) which together with \( \bar{a}_{js} \leq \hat{a}_{js} \) yield \( \bar{r}_{js} = \hat{a}_{js} \). Thus,

\[
\bar{q}_{js} = \hat{q}_{js} = 0, \forall s \in S.
\]

If \( \bar{r}_{js} = c \sum_{i \in J} \bar{x}_{ij}a_{js} \), the value of \( \bar{r}_{js} \) depends on the relationship between \( \bar{a}_{js} \) and \( c \sum_{i \in J} i \bar{x}_{ij} a_{js} \). Specifically, if \( \bar{a}_{js} \leq c \sum_{i \in J} i \bar{x}_{ij}a_{js} \), we obtain \( \bar{r}_{js} = a_{js} \), and therefore \( \bar{q}_{js} = \hat{q}_{js}, \forall s \in S \) (since \( \bar{q}_{js} = 0, \hat{q}_{js} = \hat{a}_{js} - c \sum_{i \in J} i \bar{x}_{ij}a_{js} \)). Thus, \( \bar{q}_{js} \leq \hat{a}_{js}, \forall s \in S \) always hold. Then with Constraints (4.27) – (4.29) we obtain \( \bar{b}_{j's} \geq \hat{b}_{j's}, \bar{y}_{j's} \geq \hat{y}_{j's}, \forall j' \in J, j \in \{j', \ldots, j_2\}, s \in S \).

Applying the above analysis process over all time points from \( j_2 + 1 \) to \( j \), we obtain

\[
\bar{b}_{j'js} \geq \hat{b}_{j'js}, \bar{y}_{j'js} \geq \hat{y}_{j'js}, \forall j' \in J, j \in \{j', \ldots, j_2\}, s \in S.
\]

Then, since \( \hat{y}_{j'js} = 1, \forall j' \in J, s \in S \) and \( \bar{y}_{j'js} \in [0,1], \forall j' \in J, j \in \{j', \ldots, j\}, s \in S \), we obtain \( \bar{y}_{j'js} = 1, \forall j' \in J, s \in S \), meaning that the constructed solution also satisfies Constraints (4.31).

Thus, the alternate solution is feasible to the investigated problem. Further, the difference between the objective values of \( \hat{o} \) and \( \bar{o} \) is
\[ z_f(\hat{\delta}) - z_f(\hat{\delta}) = z_f(\hat{\delta}) - z_f(\hat{\delta}) + w_\delta \sum_{s \in S} \left[ \sum_{j \in \{J, \cdots, J\}} \left( \sum_{j'=0}^{j-1} \left( a_{j's} (\tilde{y}_{j'(j-1)s} - \tilde{y}_{j'(j-1)s}) \right) \right) \right] > 0. \]

This contradicts \( \hat{\delta} \) begin the optimal solution. Thus, this proposition holds for \( j < J \) as well. This completes the proof. \( \Box \)

**Proposition 4.10.** For a node \((l, n), l \in L, n \in N_l\) in the B&B tree with cross-sectional passenger queue \( q_{jn_s}, \forall s \in S\), passenger arrival demand \( a_{js}, \forall j \in J, s \in S\), passenger demand distribution \( p_{juv}, \forall j \in J, u \in S, v \in S_u^+, \) and cumulative percentage of boarding passengers \( y'_{j's}, \forall j' \in J, j \in \{j, \cdots, J\}, s \in S\), a lower bound to the future operational cost is

\[
\min \{ f_{ln}(x_{ij}) \} = \sum_{s \in S} \left\{ Q_{lns} := q_{jn_s} + \sum_{u \in S_{s+1}} \sum_{v \in S_u^+} \sum_{j'=j_{ln}} a_{j'u} p_{j'u} (1 - y'_{j'u}) \right\} f^{i}_{l}, \tag{4.53} \]

**Proof.** To prove this proposition, we will show that for a feasible solution \( \{x_{ij}, b'_{j's}, y'_{j's}, r_{j's}\} \) to problem (4.22) \( \sim (4.36) \) constructed based on the solution at node \((l, n)\), a lower-bound solution with a future operational cost of \( \min \{ f_{ln}(x_{ij}) \} \) can be found. To find such a solution, we first define \( \mathcal{K} := \{1, 2, \cdots K\} \) as the set of future dispatches (i.e., dispatches after \( j_{ln} \)) in this feasible solution, where \( K := \sum_{l \in L} \sum_{j \in \{j_{ln}+1, \cdots, J\}} x_{ij} \) is the number of future dispatches. Let \( i_{k_s}, \forall k \in \mathcal{K}, s \in S \) be the formation of the \( k \)-th future dispatched MAV at station \( s \). With this new notation, the future operational cost of this feasible solution can be rewritten as

\[
\sum_{l \in L} \sum_{j \in \{j_{ln}+1, \cdots, J\}} x_{ij} f_l = \sum_{k \in \mathcal{K}} f^{i}_{k_s}, \forall s \in S. \]

Next, we construct the lower-bound solution by concatenating the MAVs in this feasible solution into the longest vehicle formation at any \( s \in S\). Note that the following concentration procedure can be conducted for any station, but we omit \( \forall s \in S \) in the equations for the convenience of the
notation. First, define \( m = 1, k_{m_1} = 1 \). Second, let \( k_{m_2} = \arg\min_{k_{m_2} \in \{k_{m_1}, \ldots, K\}} \{\sum_{k \in \{k_{m_1}, \ldots, k_{m_2}\}} i_{ks} \geq I\} \) and \( \Delta i_{ms} = \sum_{k \in \{k_{m_1}, \ldots, k_{m_2}\}} i_{ks} - I \). Third, let \( m = m + 1, k_{m_1} = k_{(m-1)2}, \) and \( i_{km_2} = \Delta i_{m-1} \).

Check if \( k_{m_1} = K \). If not, go to the second step and otherwise, terminate this concatenation process.

Based on Constraints (4.31), we know that the minimum number of MPs needed to serve all these passengers is \( \lceil \frac{Q_{ins}}{c} \rceil \). Therefore, any feasible solution must satisfy \( \sum_{i \in I} \sum_{j \in \{j_{(m+1), \ldots, j\}} i_{ixjs} \geq \lceil \frac{Q_{ins}}{c} \rceil, \forall s \in S \). Further, we can prove that \( I \left\lfloor \frac{Q_{ins}}{Ic} \right\rfloor - \left\lceil \frac{Q_{ins}}{Ic} \right\rceil \leq I \) and \( \left\lceil \frac{Q_{ins}}{Ic} \right\rceil = \left\lfloor \frac{Q_{ins}}{Ic} \right\rfloor + 1 \), which results in

\[
\sum_{i \in I} \sum_{j \in \{j_{(m+1), \ldots, j\}}} i_{ixjs} \geq \left\lceil \frac{Q_{ins}}{c} \right\rceil \geq I \left\lfloor \frac{Q_{ins}}{Ic} \right\rfloor, \forall s \in S.
\]

Therefore, for each station, the above concatenation process will be repeated for \( \left\lceil \frac{Q_{ins}}{Ic} \right\rceil \) times before \( k_{m_1} = K \) is satisfied; i.e., the final value of \( m \) is \( \left\lfloor \frac{Q_{ins}}{Ic} \right\rfloor \). Further, since the operational cost is a concave function over \( i \) (Eq. (3.2)), we can prove that

\[
\sum_{k \in \{k_{m_1}, \ldots, k_{m_2}\}} f_{iks} \geq f_i + f_{\Delta i_{ms}}, \forall m \in \left\{1, \ldots, \left\lfloor \frac{Q_{ins}}{Ic} \right\rceil \right\}.
\]

Since \( \sum_{k \in \{k_{m_1}, \ldots, k_{m_2}\}} i_{ks} = \Delta i_{(m-1)s} + \sum_{k \in \{k_{m_1}, \ldots, k_{m_2}\}} i_{ks} \), summing the above inequality over \( m \) yields

\[
\sum_{k \in K} f_{iks} \geq \left\lceil \frac{Q_{ins}}{Ic} \right\rceil f_i + f_{\Delta i_{ms} \left\lfloor \frac{Q_{ins}}{Ic} \right\rceil} \geq \left\lceil \frac{Q_{ins}}{Ic} \right\rceil f_i.
\]

Finally, summing the above inequality over \( s \) yields

\[
f_{in}(x_{ij}) = \sum_{s \in S} \sum_{k \in K} f_{iks} \geq \sum_{s \in S} \left\lfloor \frac{Q_{ins}}{Ic} \right\rceil f_i.
\]

This completes the proof. □

**Proposition 4.11.** For a node \((l, n), l \in L, n \in N_l\) in the B&B tree with time index \( j_{in} \), bounded
number of time points elapsed from the previous dispatch $h_{ln}$, passenger queue $q_{lns}, \forall s \in S$, and passenger arrival demand $a_{js}, \forall j \in J, s \in S$, we construct an ordered set $E_{lns} := \{q_{lns}\} \cup \{a_{js}\}_{j \in \{ln+1, \cdots, J\}}, \forall s \in S$, where the elements are sorted in an ascending order. Then a lower bound to the future passenger waiting cost is

$$\min \{g_{ln}(y'_{js})\} = w \delta \sum_{s \in S} \sum_{e \in \{1, \cdots, e_{ln}\}} E_{lns},$$

(4.54)

where $e$ denotes the $e$-th element in $E_{lns}$ and $e_{ln} := \left\lfloor \frac{J - j_{ln} + (j - h_{ln} + 1)}{j} \right\rfloor (j - 1)$.

**Proof.** Based on Constraints (4.22) – (4.23), for any feasible solution $\{x_{ils}, b'_{js}, y'_{js}, r_{js}\}$ to problem (4.22) – (4.36), $j_{ln} + (j - h_{j}) + 1$ is the first time point after $j_{ln}$ when MAV can be dispatched. Thus,

$$\sum_{i \in I} \sum_{j \in \{j_{ln}+1, \cdots, J\}} x_{ij1} \leq \left\lfloor \frac{J - j_{ln} + (j - h_{ln} + 1)}{j} \right\rfloor,$$

which means that the number of dispatches one can dispatch in any feasible solution to problem (4.22) – (4.36) is bounded by the number of dispatches in a solution where we adopt the minimum dispatch headway between every two consecutive dispatches. Let $\bar{J} = J - j_{ln}$. Then the number of time points where no MAVs can be dispatched after $j_{ln}$ in any feasible solution satisfies

$$\bar{J} - \sum_{i \in I} \sum_{j \in \{j_{ln}+1, \cdots, J\}} x_{ij1} \geq J - \left\lfloor \frac{J - j_{ln} + (j - h_{ln} + 1)}{j} \right\rfloor = \left\lfloor \frac{\bar{J} - j_{ln} + (\bar{J} - h_{ln} + 1)}{\bar{J}} \right\rfloor (\bar{J} - 1) = e_{ln}.$$

Since no MAVs are dispatched at these time points, passengers arriving at any station at these time
points must wait for at least one time point before being able to board a MAV in any feasible solution. Define the set of time points where no MAVs are dispatch in any feasible solution as $\mathcal{J}_1 := \{ j \in \{ j_{ln}, \ldots, J \} | \sum_i x_{ij} = 0 \}$. With this, the above inequality can be equivalently written as $|\mathcal{J}_1| \geq e_{ln}$. Further, we obtain

$$g_{ln}(y'_{j's}) \geq w\delta \sum_{s \in S} \sum_{j \in \{ j_{ln+1}, \ldots, J \}} \left( \sum_{j=0}^{j-1} (a'_{js}(1 - y'_{j'(j-1)s})) \right) \geq w\delta \sum_{s \in S} \left[ q_{j_{ln}s} + \sum_{j \in \mathcal{J}_1 \setminus \{ j_{ln} \}} (a_{js}) \right].$$

Based on the definition of $\mathcal{E}_{\text{ins}}$, $\sum_{e \in \{ 1, \ldots, e_{ln} \}} \mathcal{E}_{\text{ins}}$ represents the minimum between the summations of any $e_{ln}$ elements in $\mathcal{E}_{\text{ins}}$. Thus, if $|\mathcal{J}_1| = e_{ln}$, we obtain $q_{j_{ln}s} + \sum_{j \in \mathcal{J}_1 \setminus \{ j_{ln} \}} (a_{js}) \geq \sum_{e \in \{ 1, \ldots, e_{ln} \}} \mathcal{E}_{\text{ins}}$. If $|\mathcal{J}_1| > e_{ln}$, $q_{j_{ln}s} + \sum_{j \in \mathcal{J}_1 \setminus \{ j_{ln} \}} (a_{js}) \geq \sum_{e \in \{ 1, \ldots, e_{ln} \}} \mathcal{E}_{\text{ins}}$ also holds because $q_{j_{ln}s} \geq 0$ and $a_{js} > 0, \forall j \in \mathcal{J}, s \in S$. Summing this inequality over $s \in S$ and multiply it by $w\delta$, we obtain

$$g_{ln}(y'_{j's}) \geq w\delta \sum_{s \in S} \left[ q_{j_{ln}s} + \sum_{j \in \mathcal{J}_1 \setminus \{ j_{ln} \}} (a_{js}) \right] \geq w\delta \sum_{s \in S} \sum_{e \in \{ 1, \ldots, e_{ln} \}} \mathcal{E}_{\text{ins}}.$$

This completes the proof. $\square$