Towards A Robust Integrated Urban Mobility System: Public Transit and Ride-Sharing Systems

by

Xiaotong Guo

M.S.T., Massachusetts Institute of Technology, 2023
M.S., Cornell University, 2019
B.Eng., Tongji University, 2017

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN TRANSPORTATION

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Xiaotong Guo Department of Civil and Environmental Engineering May 10, 2024
Certified by:	Jinhua Zhao Professor of City and Transportation Planning, Thesis Supervisor
Accepted by:	Heidi Nepf Donald and Martha Harleman Professor of Civil and Environmental Engineering Chair, Graduate Program Committee

Towards A Robust Integrated Urban Mobility System: Public Transit and Ride-Sharing Systems

by

Xiaotong Guo

Submitted to the Department of Civil and Environmental Engineering
on May 10, 2024, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Transportation

Abstract

The global pandemic has fundamentally changed lifestyles, impacting how, when, and where people travel within cities. In this post-pandemic world, urban mobility demand patterns are experiencing significant shifts. To manage the growing uncertainty in urban mobility, there is a growing need to develop a robust urban mobility system. This system must be adaptable to evolving demand patterns while ensuring efficiency and environmental sustainability in transporting large populations. Additionally, the increasing popularity of shared mobility and rapid advancements in autonomous driving technologies are creating new opportunities for innovative approaches to urban transportation systems.

This dissertation delves into the development of a robust and integrated urban mobility system for the future, with a focus on the public transit and ride-sharing systems. While the advent of shared mobility platforms such as Uber and Lyft, along with Autonomous Mobility-on-Demand (AMoD) services like Waymo and Cruise, have revolutionized urban travel, public transit systems remain the backbone of urban mobility. This is attributed to their capacity to move large numbers of people over long distances at a relatively low cost and an environmentally friendly way. Thus, this study aims to enhance the robustness of both public transit and ride-sharing systems and explore ways to seamlessly integrate these two components. The dissertation presents five distinct studies to elaborate on these objectives.

The first three studies focus on the vehicle rebalancing problem, which is one of the most critical strategies in ride-sharing operations. An effective rebalancing strategy can significantly reduce empty miles traveled and reduce customer wait times by better matching supply and demand. While the supply (vehicles) is usually known to the system, future passenger demand is uncertain. The first study proposes a novel approach to better immunize rebalancing decisions against demand uncertainty. This approach, namely the matching-integrated vehicle rebalancing (MIVR) model, incorporates driver-customer matching into vehicle rebalancing problems to produce better rebalancing strategies. For further protection against uncertainty, robust optimization (RO) techniques are introduced to construct a robust version of the MIVR

model. Problem-specific uncertainty sets are designed for the robust MIVR model. The second study further explores different approaches for handling demand uncertainty in the vehicle rebalancing problem. There are two ways to handle uncertainty. First, the point-prediction-driven optimization framework involves predicting the future demand and then producing rebalancing decisions based on the predicted demand. Second, data-driven optimization approaches directly prescribe rebalancing decisions from data. In this study, a predictive prescription framework is introduced to this problem, where the benefits of predictive and data-driven optimization models are combined.

Although vehicle rebalancing algorithms could improve system efficiency, there exists a detrimental feedback loop where underserved communities with low demand density are unintentionally discriminated. To resolve this fairness issue, the third study develops algorithms for vehicle rebalancing that aim to minimize disparity within the system. Grasping the concept of disparity is a foundation for understanding fairness in the ride-hailing system. The vehicle rebalancing encompasses two critical aspects: upstream demand forecasting and downstream vehicle repositioning. The issues of disparities within both these components are addressed. To reduce disparity in demand prediction, we implement a strategy utilizing a Socio-Aware Spatial-Temporal Graph Convolutional Network (SA-STGCN), aimed at improving demand forecast accuracy while reducing discrepancies in prediction errors across diverse regions. For equitable repositioning of the supply side vehicles, we introduce a disparity-reducing MIVR system. This system is designed to facilitate a balanced vehicle distribution, ensuring that ride-hailing services are accessible equitably across different areas.

The fourth study focuses on the robustness of public transit systems. Limited studies have considered demand uncertainties when designing transit schedules. To better address demand uncertainty issues inherent in public transit systems, this study utilizes the RO framework to generate robust transit schedules against demand uncertainty. A nominal (non-robust) optimization model for the transit frequency setting problem (TFSP) under a single transit line setting is first proposed. The model is then extended to the RO-based formulation to incorporate demand uncertainty. The large-scale origin-destination (OD) matrices for real-world transit problems present computational challenges to solve the optimization problem. To efficiently generate robust transit schedules, a Transit Downsizing (TD) approach is proposed to reduce the dimensionality of the problem.

The last study focuses on the integration of emerging AMoD systems with existing public transit networks. We propose a novel optimization framework to generate the system design of the Transit-Centric Multimodal Urban Mobility with Autonomous Mobility-on-Demand (TCMUM-AMoD) at scale. The system operator (public transit agency) determines the network design and frequency settings of the PT network, fleet sizing and allocations of the AMoD system, and the pricing for using the multimodal system with the goal of minimizing passenger disutility. Passengers' mode and route choice behaviors are modeled explicitly using discrete choice models. A first-order approximation algorithm is introduced to solve the problem at scale. Using a case study in Chicago, we show the potential to generate integrated urban mobility systems

in different demand scenarios.

The final chapter summarizes the whole dissertation and outlines potential avenues for future research directions.

Thesis Supervisor: Jinhua Zhao
Title: Professor of City and Transportation Planning

Acknowledgments

Thank you to everyone who has supported me throughout my Ph.D. studies. It has truly been an honor to learn from and grow alongside each one of you. First and foremost, I extend my heartfelt gratitude to my Ph.D. advisor, Prof. Jinhua Zhao, for being an exceptional mentor and advisor. Throughout our numerous discussions, both within the realm of academia and beyond, I have consistently gained new and invaluable insights from Jinhua. His passion for research is not only inspiring but has profoundly influenced my approach to scholarly inquiry and my broader perspective on life.

Next, I extend my sincere appreciation to my Ph.D. committee members, Prof. Patrick Jaillet and Prof. Daniel Freund, for their valuable comments and feedback on my dissertation. Their unique approaches to problem-solving have not only inspired me but have also elevated my thinking to new heights. Their insights have been instrumental in refining my research and broadening my academic perspective.

Third, I extend my deepest gratitude to my mentors and senior researchers whose invaluable feedback has significantly shaped my academic journey. I am particularly thankful to Hongmou Zhang for sparking my interest in shared mobility. I also appreciate Shenhao Wang and Zhao Zhan for their insightful perspectives that have enriched my research process. Additionally, I would like to express my sincere thanks to Prof. Haris Koutsopoulos and Prof. Nigel Wilson for their guidance and direction in my research endeavors. Their expertise and support have been pivotal in steering my work towards meaningful outcomes.

Next, I would like to express my heartfelt thanks to my colleagues at the MIT Transit Lab and MIT Urban Mobility Lab for their support and camaraderie throughout my Ph.D. studies. Special thanks go to Baichuan Mo, Qingyi Wang, and Yunhan Zheng, who helped me navigate the challenges of the pandemic and in making my time at MIT truly unforgettable. I'm also grateful to Nick Caros for his collaboration on various projects, which provided great inspiration to my work. Thanks to my cohort—John Moody, Andy Haupt, and Jonas Lehmann—for being such wonderful

company during our studies.

Additionally, I extend my appreciation to my collaborators in the lab: Joseph Rodriguez, Xinyi Wang, Daniela Shuman, Gabriel Barrett, Yen-Chu Wu, Xinling Li, Hanyong Xu, and Dingyi Zhuang for their contributions and teamwork.

Last but not least, I owe a deep gratitude to my friends and family, who have been my constant source of strength throughout my research journey. Thanks to Yuting Wang, Lei Huang, Shenghao Guo, and Jianting Guo for countless memorable nights. Thanks to Cheng Tang and Ced Zhang for being great gym and basketball buddies. I am also grateful to everyone on the MIT CSSA Men's Basketball Team for enriching my experience at MIT.

A special thanks to my girlfriend, Lenna Yang, for her unwavering support during my Ph.D. career. Thanks to my parents, Xiangyang Guo and Rong Hu, for their endless encouragement in pursuing my dreams. I would also like to extend my gratitude to my grandmother, Meilan Zhao, who has been a source of constant support and inspiration since my childhood, embodying perseverance, kindness, and love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

In recent years, the demand for urban mobility has undergone significant changes, largely due to the global pandemic. This COVID-19 has not only altered people's lifestyles, but has also increased the demand uncertainty surrounding urban transportation. There are three perspectives of demand uncertainty in urban mobility: i) demand changes, ii) uncertainty level, and iii) uncertainty changes.

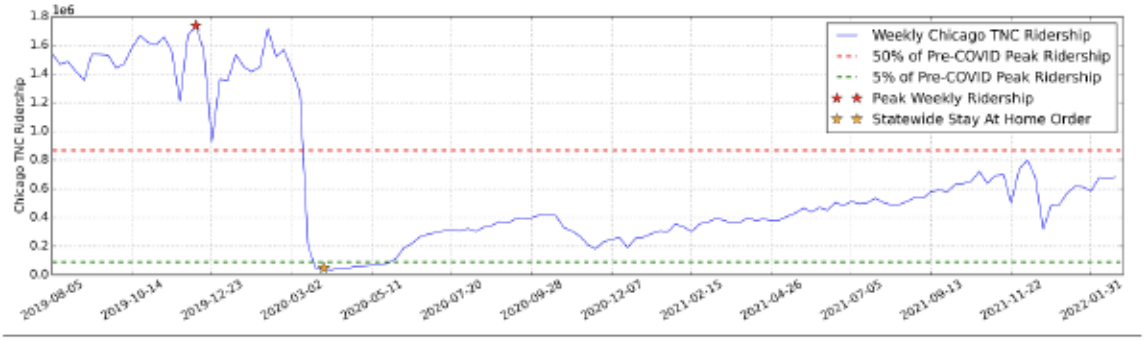


Figure 1-1: Weekly ride-hailing demand for city of Chicago (2019 to 2022).

First, demand changes indicate both spatial and temporal changes in travel demand. Figure 1-1 shows the weekly ride-hailing demand for the city of Chicago from 2019 to 2022. Chicago's ride-hailing demand only recovers to 50% of the pre-COVID level at 2022. Figure 1-2 shows the daily average demand distributions for sub-regions in Manhattan, New York City (NYC) from 2019 to 2022. Sub-regions are taxi zones

defined by the NYC Taxi and Limousine Commission[130]. In 2022 compared to 2019, a 36% demand increase is found. Different patterns for temporal demand changes are found in two major US cities.

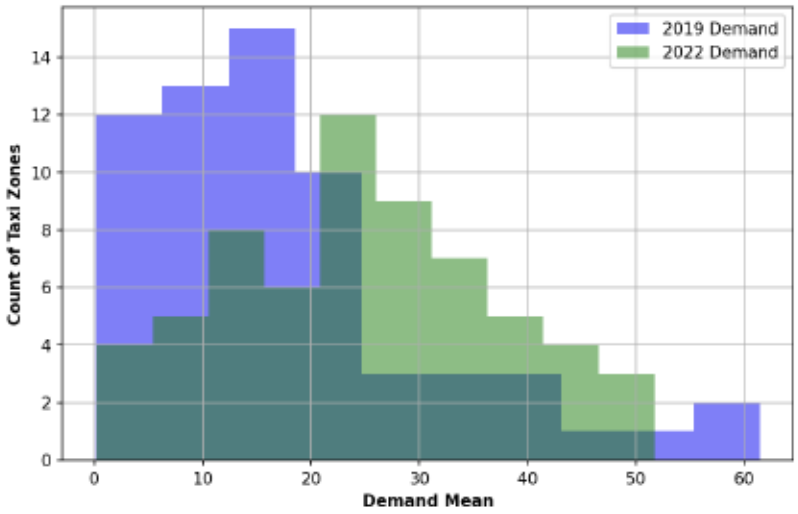
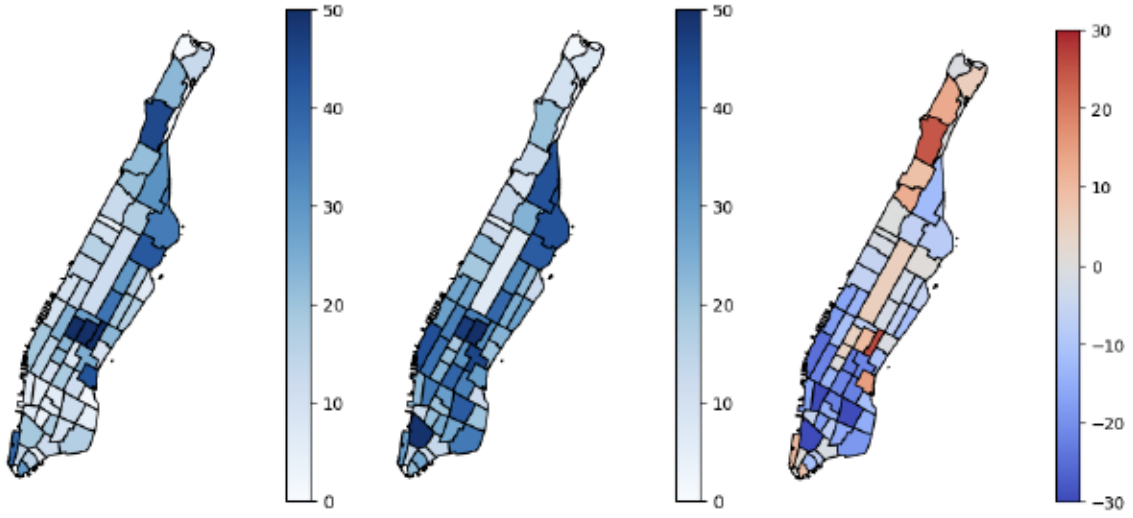


Figure 1-2: Daily average demand distributions for sub-regions in Manhattan (2019 to 2022).

Figure 1-3 illustrates the spatial demand distributions for ride-hailing in Manhattan for the years 2019 and 2022, as well as the differences between these two years.



(a) 2019 Demand (b) 2022 Demand (c) 2019 - 2022 Demand

Figure 1-3: Spatial ride-hailing demand distributions for Manhattan in 2019 and 2022.

Although the overall demand for ride-hailing services increased by 36% from 2019 to 2022, the spatial patterns of demand changed significantly. Notably, demand in lower Manhattan increased, whereas upper Manhattan experienced a decrease in demand.

Similar changes in the demand pattern have also been observed in transit systems. Figure 1-4 shows the changes in ridership by modes in the public transportation system from 2019 to 2022. Till the end of 2022, the transit ridership has only recovered to 68% of the pre-pandemic level. As agencies struggle to maintain service level, transit ridership might never return as people travel less to work.

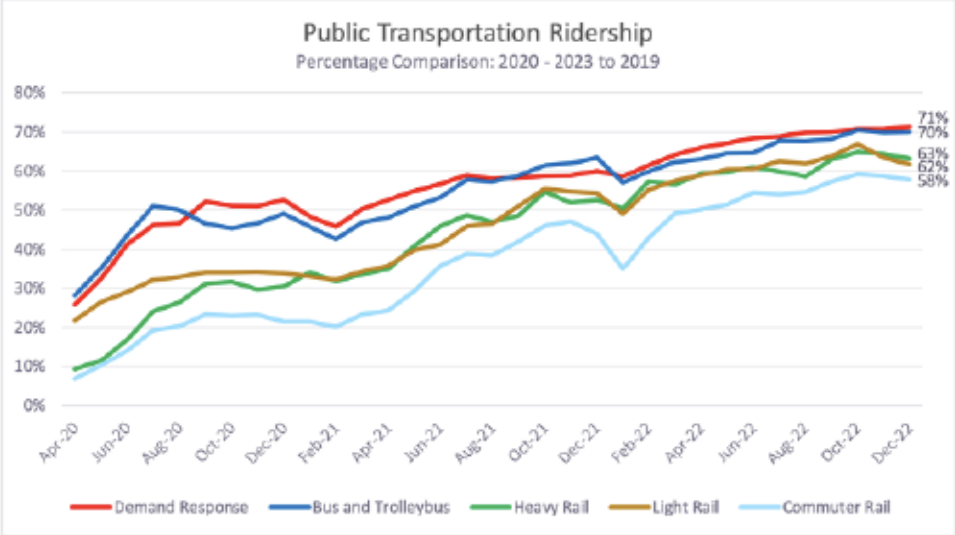


Figure 1-4: Public transportation ridership changes by modes (2019 to 2022).

Secondly, the level of demand uncertainty provides a direct measure of the fluctuations in urban mobility demand. The term "uncertainty level" is defined in this dissertation as the standard deviation of daily demand changes. Figures 1-5 and 1-6 depict the demand uncertainty levels for both ride-hailing and transit systems. Throughout both the pre-COVID and post-COVID periods, these uncertainty levels have been substantial and cannot be overlooked. Specifically, ride-hailing demand uncertainty increased by 15% from 2019 to 2022, while transit demand uncertainty saw a significant reduction of 70% between 2019 and 2020.

Thirdly, demand levels have changed over time, as illustrated in Figures 1-5 and 1-6. These changes in demand emphasize the critical importance of integrating demand

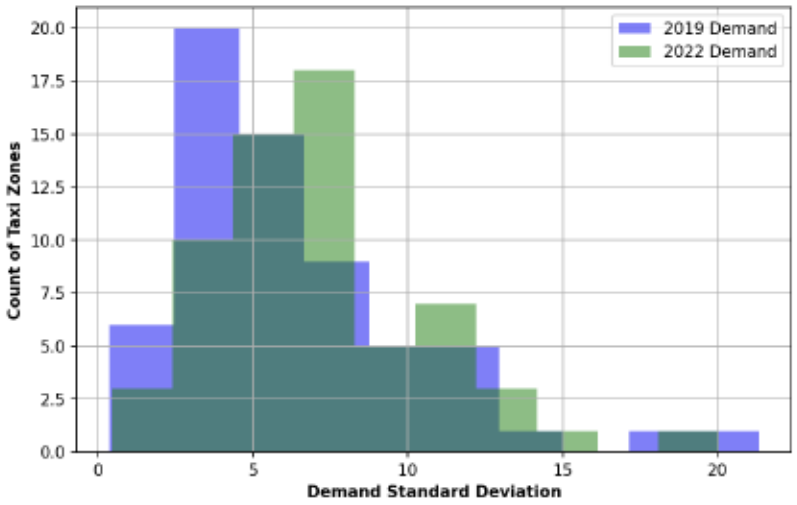


Figure 1-5: Ride-hailing demand uncertainty distributions for sub-regions in Manhattan (2019 to 2022).

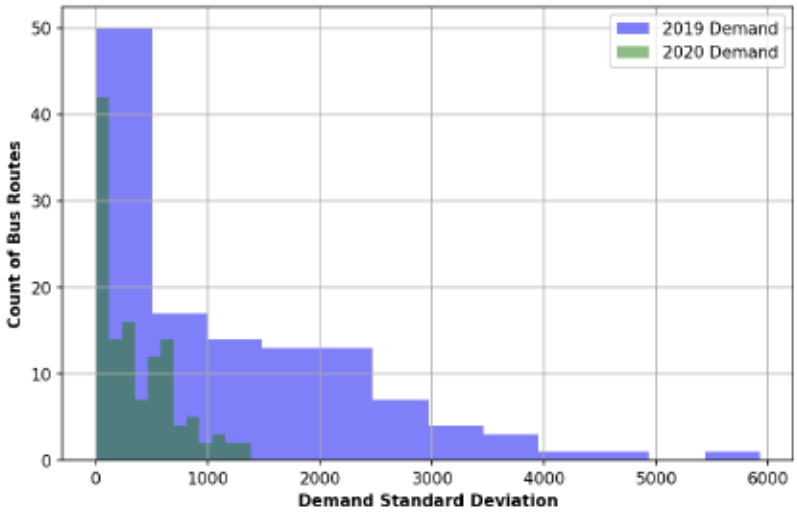


Figure 1-6: Transit demand uncertainty distributions for bus routes in Chicago (2019 to 2022).

uncertainty into the design and operation of transportation systems. However, both transit agencies and ride-hailing platforms tend to focus solely on adapting to demand changes, often neglecting the broader implications of demand uncertainty on system performance. In this dissertation, we propose a strategic framework to address the prevalent uncertainty issues in urban mobility systems.

The central focus of this dissertation is to address system-wise uncertainty through two approaches: i) system integration and ii) robust operation. We propose an inte-

grated and robust urban mobility system, with a particular emphasis on public transit and ride-sharing systems. Robust operation includes strategies that mitigate various sources of uncertainty. In terms of system integration, by integrating systems, such as public transit and ride sharing, passengers gain access to a wider range of mobility options, enhancing the likelihood of their transportation needs being met. This dissertation does not explore the exact dynamics on how the system integration addresses uncertainties. Instead, it focuses initially on the development of an integrated urban mobility system, setting the stage for future discussions and explorations of these concepts.

1.2 Uncertainty Definition

Uncertainty is formally defined as "epistemic situations involving imperfect or unknown information." It has been explored in various research fields, each adopting a unique perspective. Lo and Mueller [112] have contributed to the understanding of uncertainty by proposing a taxonomy that includes five distinct levels: 1) complete certainty, where everything is known; 2) risk without uncertainty, which involves known probabilities; 3) fully reducible uncertainty, which can be entirely eliminated through information gathering; 4) partially reducible uncertainty, where only some aspects can be clarified; and 5) irreducible uncertainty, which cannot be reduced regardless of the information gathered. In the context of transportation systems—a large-scale dynamic entity—uncertainty most commonly aligns with level 4, indicating that while uncertainty can be managed, it cannot be completely eliminated, and thus must be strategically addressed using modeling approaches.

Meanwhile, it is crucial to consider uncertainty within specific contexts. For example, model uncertainty arises when multiple models are consistent with observations. Data uncertainty refers to data compromised by noise, which causes deviations from true values. This dissertation focus on system uncertainty, specifically within the physical urban mobility system.

The urban mobility system is subject to three main sources of uncertainty, which

are crucial to its operation and planning:

1. **Supply Uncertainty**: This involves variations in the availability of transportation resources, such as the number of vehicles or operational status.

2. **Demand Uncertainty**: This pertains to the unpredictable fluctuations in user demand for transportation services, which can vary widely due to numerous factors including temporal patterns, economic conditions, and social events.

3. **Environment Uncertainty**: This encompasses external conditions that can impact transportation systems, such as weather changes, road conditions, or regulatory changes.

Addressing these types of uncertainties is crucial for creating more resilient and efficient urban mobility systems. This dissertation specifically explores how uncertainties can be integrated into the design and operations of two key urban mobility systems: public transit and ride-sharing. Each system-level uncertainty—supply, demand, and environment—presents unique challenges and opportunities for integration into these systems.

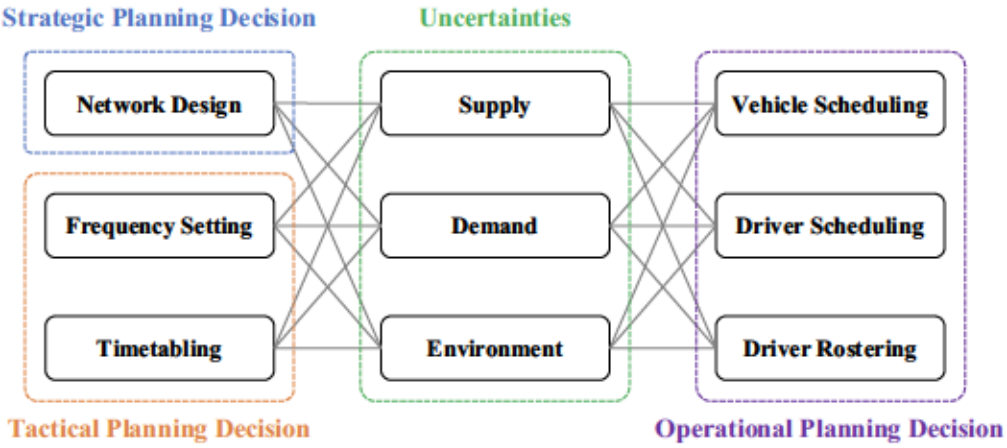


Figure 1-7: Uncertainty in transit systems.

Figures 1-7 and 1-8 illustrate how these uncertainties can be accounted for in the design and operational strategies of the transit and ride-sharing systems, respectively. By incorporating uncertainty directly into system planning and management, both

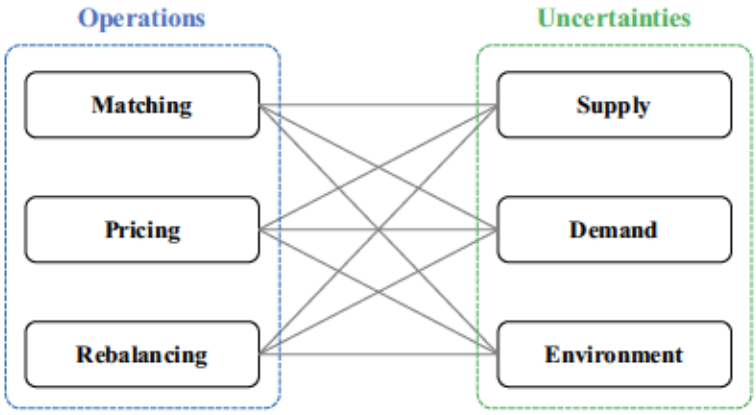


Figure 1-8: Uncertainty in ride-sharing system.

systems can be optimized to better handle the dynamic and often unpredictable conditions affecting urban mobility.

In this dissertation, the primary focus is on demand uncertainty, which poses significant challenges to urban mobility systems. Specifically, Chapters 2 and 3 are dedicated to addressing demand uncertainty within the context of the vehicle rebalancing problem. This involves developing strategies to ensure that ride-sharing vehicles are appropriately distributed across the city to meet fluctuating demands efficiently.

Chapter 5 shifts the focus to the transit frequency setting problem, where demand uncertainty can greatly impact the efficiency and effectiveness of public transit services. The chapter explores methodologies for adjusting the frequencies of transit services to better align with changing passenger volumes, thereby minimizing wait and in-vehicle travel times.

1.3 Research Questions

This dissertation focuses on addressing system-wise uncertainty by building a robust and integrated urban mobility system, with a focus on public transit and ride-sharing systems. Specifically, the dissertation

1. **Identify sources of uncertainty and propose robust operation strate-**

gies for both public transit and ride-sharing systems.

2. **Design an integrated urban mobility system with both public transit and ride-sharing systems.**

For the second research question concerning the design of an integrated urban mobility system that combines public transit and ride-sharing, previous studies have highlighted the potential benefits and proposed various frameworks for the operations and design of such systems. However, there is a lack of a comprehensive framework that simultaneously addresses the design of transit networks, fleet sizing and allocation for ride-sharing vehicles, and pricing strategies for the integrated system.

This dissertation aims to bridge this research gap by proposing a methodological framework that integrates these critical aspects. This comprehensive approach lays the groundwork for future research on integrated urban mobility systems, enabling consideration of diverse objectives and system design philosophies.

1.4 Data Sources

This dissertation draws on two types of datasets: ride-sharing and public transit datasets. The ride-sharing dataset utilized is the New York City (NYC) high-volume ride-hailing trip data collected by the NYC Taxi and Limousine Commission [130]. This dataset contains information on pickup times, trip origins and destinations at the level of the taxi zone, and group size. In addition, the historical average traffic speed data provided by Uber Movement is used to estimate the travel time within the city of New York.

For the public transit dataset, this dissertation utilizes transit data provided by the Chicago Transit Authority (CTA). The running times between two stops for different transit routes are estimated from the Automatic Vehicle Location (AVL) data. The current transit schedule information is from an open-source Generalized Transit Feed Specification (GTFS) dataset, which is published by CTA every month. The transit demand data are obtained from CTA's ODX database. The "ODX" stands for the

"origin, destination, and transfer inference algorithm", an algorithm developed by Gabriel et al. [148] and currently implemented within the CTA. The CTA transit network is equipped with a "tap-on" only fare collection system, indicating that the alighting information is not reported in the system. The ODX algorithm is utilized to infer the alighting information and details can be found in [35, 148, 198].

1.5 Dissertation Outline

1.5.1 Chapter 2: Robust Matching-Integrated Vehicle Rebalancing in Ride-Hailing System with Uncertain Demand

With the rapid growth of the mobility-on-demand (MoD) market in recent years, ride-hailing companies have become an important element of the urban mobility system. There are two critical components in the operations of ride-hailing companies: driver-customer matching and vehicle rebalancing. In most previous literature, each component is considered separately, and performances of vehicle rebalancing models rely on the accuracy of future demand predictions. To better immunize rebalancing decisions against demand uncertainty, a novel approach, the matching-integrated vehicle rebalancing (MIVR) model, is proposed in this paper to incorporate driver-customer matching into vehicle rebalancing problems to produce better rebalancing strategies. The MIVR model treats the driver-customer matching component at an aggregate level and minimizes a generalized cost including the total vehicle miles traveled (VMT) and the number of unsatisfied requests.

For further protection against uncertainty, robust optimization (RO) techniques are introduced to construct a robust version of the MIVR model. Problem-specific uncertainty sets are designed for the robust MIVR model to reflect the demand uncertainty in ride-hailing systems. The proposed MIVR model is tested against two benchmark vehicle rebalancing models using real ride-hailing demand and travel time data from New York City (NYC). The MIVR model is shown to have better per-

formances by reducing customer wait times compared to benchmark models under most scenarios. Sensitivity analyses have been conducted to better understand how the proposed MIVR model performs under different demand-supply scenarios. In addition, the robust MIVR model produces better solutions by planning for demand uncertainty compared to the non-robust (nominal) MIVR model.

1.5.2 Chapter 3: Data-Driven Vehicle Rebalancing with Predictive Prescriptions in the Ride-Hailing System

Although the robust optimization techniques can effectively protect vehicle rebalancing decision against demand uncertainty, it suffers from the computation complexity issue. In this chapter, we discuss additional approaches for handling uncertainty in demand making under uncertainty. There are two ways to handle uncertainty. First, the point-prediction-driven optimization framework involves predicting the future demand and then producing rebalancing decisions based on the predicted demand. Second, the data-driven optimization approaches directly prescribe rebalancing decisions from data.

In this study, a predictive prescription framework is introduced to this problem, where the benefits of predictive and data-driven optimization models are combined. The predictive prescription framework utilizes unsupervised machine learning algorithms to generate weights for each historical day with demand data according to similarity from auxiliary information. The weights are then used in a stochastic optimization framework to generate rebalancing decisions. Based on the matching-integrated vehicle rebalancing (MIVR) model, predictive prescriptions are introduced to handle demand uncertainty.

Model performances are evaluated using real-world simulations with New York City (NYC) ride-hailing data under four demand scenarios. When demand can be accurately predicted, a point-prediction-driven optimization framework should be adapted. The proposed predictive prescription models achieve shorter customer wait times over the point-prediction-driven optimization models when future demand pre-

dictions are not so accurate, and achieve a competitive performance with respect to the cutting-edge robust optimization models. The proposed approach has a computational edge over the robust models, while achieving similar performances.

1.5.3 Chapter 4:Disparity-Reducing Vehicle Rebalancing in the Ride-Hailing System

The previous two chapters discuss the vehicle rebalancing problem and propose approaches for handling demand uncertainty. However, vehicle rebalancing models might unintentionally lead to fairness issues. Vehicle rebalancing models redistribute more vacant vehicles to areas with higher anticipated demand. As a result, low-demand areas, which are typically underserved communities, will be discriminated in the vehicle rebalancing algorithm. This chapter develops vehicle rebalancing algorithms aimed at minimizing such disparities within the system. Grasping the concept of disparity is a foundation for understanding fairness in the ride-hailing system.

The vehicle rebalancing encompasses two critical aspects: upstream demand forecasting and downstream vehicle repositioning. The issues of disparities within both these components are tackled. To reduce disparity in demand prediction, we implement a strategy utilizing a Socio-Aware Spatial-Temporal Graph Convolutional Network (SA-STGCN), aimed at improving demand forecasting accuracy while reducing discrepancies in prediction errors across diverse regions. There are three components added to the STGCN framework to reduce disparity: i) socio-demographic enriched adjacency matrix, ii) Fairness-Enhanced Loss Regularization, and iii) decomposed fairness weights from the enriched adjacency matrix for the downstream vehicle rebalancing.

For equitable supply-side vehicle repositioning, we introduce a disparity-reducing Matching-Integrated Vehicle Rebalancing (MIVR) system. This system is tailored to facilitate balanced vehicle distribution, ensuring that ride-hailing services are accessible equitably across different areas. The proposed model utilizes the fairness weights generated from the SA-STGCN framework. Our methodology, tested through simu-

lations with the New York City (NYC) taxi dataset, enhances accuracy and reduces disparity in demand forecasting, leading to fairer vehicle distribution. It also reduces perceived service disparity among customers. Specifically, it alleviates disparity—indicated by a more uniform distribution of wait times across regions—by 6.5% while not diminishing system efficiency-measured by customer wait times.

1.5.4 Chapter 5: Robust Transit Frequency Setting Problem with Demand Uncertainty

Public transit systems are the backbone of urban mobility systems in the era of urbanization. The design of transit schedules is important for the efficient and sustainable operation of public transit. However, limited studies have considered demand uncertainties when designing transit schedules. To better address demand uncertainty issues inherent in public transit systems, this chapter utilizes the robust optimization (RO) framework to generate robust transit schedules against demand uncertainty. A nominal (non-robust) optimization model for the transit frequency setting problem (TFSP) under a single transit line setting is first proposed. Different transit service patterns are incorporated.

The model is then extended to the RO-based formulation to incorporate demand uncertainty, which has not been considered in the literature. The large-scale origin-destination (OD) matrices for real-world transit problems bring computational challenges in solving the optimization problem. To efficiently generate robust transit schedules, a Transit Downsizing (TD) approach is proposed to reduce the dimensionality of the problem. The TD approach consists of an optimality-preserved component and a heuristic-based component, targeting to shrink the size of possible OD pairs when generating transit schedules.

The proposed models are tested with real-world transit lines and data from the Chicago Transit Authority (CTA). Meanwhile, a stochastic programming (SP) framework is used to construct a benchmark stochastic TFSP model. Compared to the current transit schedule implemented by the CTA, the nominal TFSP model with-

out considering demand uncertainty reduces passengers' wait times while increasing in-vehicle travel times. After incorporating demand uncertainty, both stochastic and robust TFSP models reduce passengers' wait times and in-vehicle travel times simultaneously. The robust transit schedules outperform the benchmark stochastic transit schedules by reducing both wait and in-vehicle travel times when demand is significantly uncertain.

1.5.5 Chapter 6: Design of Transit-Centric Multimodal Urban Mobility System with Autonomous Mobility-on-Demand

The last chapter focuses on the integration of public transit and ride-sharing systems. It addresses the pressing challenge of urban mobility in the context of growing urban populations, changing demand patterns for urban mobility, and emerging technologies like Mobility-on-Demand (MoD) platforms and Autonomous Vehicle (AV). As urban areas swell and demand pattern changes, the integration of Autonomous Mobility-on-Demand (AMoD) systems with existing public transit (PT) networks presents great opportunities to enhancing urban mobility. We propose a novel optimization framework for solving the Transit-Centric Multimodal Urban Mobility with Autonomous Mobility-on-Demand (TCMUM-AMoD) at scale.

The system operator (public transit agency) determines the network design and frequency settings of the PT network, fleet sizing and allocations of AMoD system, and the pricing for using the multimodal system with the goal of minimizing passenger disutility. Passengers' mode and route choice behaviors are modeled explicitly using discrete choice models. A mixed integer non-linear program (MINLP) is proposed to solve this joint design problem. A first-order approximation algorithm is introduced to solve the MINLP formulation at scale. Using a case study in Chicago, we showcase the potential to optimize urban mobility across different demand scenarios. To our knowledge, ours is the first paper to jointly optimize transit network design, fleet sizing, and pricing for the multimodal mobility system while considering passengers'

mode and route choices.

1.6 Related Publications

This dissertation has resulted in five journal publications.

Chapter 2 is published as the paper "*Xiaotong Guo, Nicholas S. Caros, and Jinhua Zhao. Robust matching-integrated vehicle rebalancing in ride-hailing system with uncertain demand. Transportation Research Part B: Methodological, 150:161–189, 2021*" [67].

Chapter 3 is published as the paper "*Xiaotong Guo, Qingyi Wang, and Jinhua Zhao. Data-driven vehicle rebalancing with predictive prescriptions in the ride-hailing system. IEEE Open Journal of Intelligent Transportation Systems, 3:251–266, 2022*" [71].

Chapter 4 is under review at *Transportation Research Part A: Policy and Practice*. The preprint is "*Xiaotong Guo, Hanyong Xu, Dingyi Zhuang, Yunhan Zheng, and Jinhua Zhao. Fairness-enhancing vehicle rebalancing in the ride-hailing system, 2023*" [72].

Chapter 5 is under the final publication process at *IEEE Transactions on Intelligent Transportation Systems*. The preprint is "*Xiaotong Guo, Baichuan Mo, Haris N. Koutsopoulos, Shenhao Wang, and Jinhua Zhao. Transit frequency setting problem with demand uncertainty, 2022*" [69].

Chapter 6 is a working paper. The preprint is "*Xiaotong Guo and Jinhua Zhao. Design of transit-centric multimodal urban mobility system with autonomous mobility-on-demand, 2024*" [73].

Chapter 2

Robust Matching-Integrated Vehicle Rebalancing in Ride-Hailing System with Uncertain Demand

2.1 Introduction

Advanced wireless communication and cloud computing technologies coupled with the growing popularity of shared mobility have led to a fast-growing Mobility-on-Demand (MoD) market in recent years [140]. Ride-hailing companies, also known as Transportation Network Companies (TNCs), such as Uber and Lyft have become ubiquitous forms of MoD in most cities over the past decade. The number of world-wide active drivers for Uber grew from almost zero in 2010 to over 3 million in 2017, while Lyft, a relative latecomer to the market, had 1.4 million active drivers in the US and Toronto in 2017 [94]. Two of the primary innovations that allowed them to capture a significant market share from their established competitors, the taxi industry, were: 1) matching trip requests with drivers using a mobile app rather than curbside hailing or an in-advance booking system, and 2) responding to changes in

demand by incentivizing or actively dispatching drivers to high-demand areas. These innovations have been identified as two important ride-hailing operations problems in the literature: the driver-customer matching problem and the vehicle rebalancing problem [165].

One of the key technological competence requirements for efficient operation of ride-hailing platforms is the algorithmic approaches for optimally matching drivers and customers in real-time [1]. Given a list of available vehicles and trips requested by customers, the matching algorithm pairs drivers and customers according to specific objectives and feasibility constraints. Moreover, matching decisions need to be made quickly, typically within seconds. Researchers have been seeking solutions to improve the operational and computational performance of the on-demand driver-customer matching problem.

Because the spatial distributions of supply and demand in the ride-hailing system are often unbalanced, platforms can improve the operational performance by actively rebalancing idle vehicles to areas where the demand is expected to exceed supply based on estimates of future demand. Algorithms for rebalancing idle vehicles have been proposed for ride-hailing platforms to reduce wait times for customers [164, 172, 121, 144]. However, the performance of vehicle rebalancing algorithms depends on the accurate future demand estimations. Rebalancing decisions generated with inaccurate demand forecasts could have negative impacts on the system performance. Incorporating robustness into the vehicle rebalancing algorithm is one approach to protect solutions against demand uncertainty that arise from inaccurate estimates of future demand [121].

While rebalancing and matching are often treated as separate operations in the literature [165], both problems relate to dispatching idle vehicles, either to pick up customers or to increase supply in areas with high expected demand. A common objective for the driver-customer matching problem is minimizing the vehicle miles traveled (VMT) and unsatisfied requests [7, 3] while the primary objective for the vehicle rebalancing problem is minimizing the VMT and a functional term measuring the system-wide service availability for future demand [164, 172, 121].

In the vehicle rebalancing problem, the overall goal for improving the system-wide service availability for incoming customers is to minimize the number of unsatisfied requests, which coincides with the objective of the driver-customer matching problem. The functional term in the objective of the vehicle rebalancing problem can therefore be treated as an approximation to represent the number of unsatisfied requests. However, the maximum system-wide service availability for incoming customers does not necessarily lead to the minimum number of unsatisfied requests if there are inaccurate future demand estimates. To immunize vehicle rebalancing decisions against the inherent demand uncertainty, we introduce the driver-customer matching component into the vehicle rebalancing problem in order to explicitly model the number of unsatisfied requests.

Nonetheless, there is a methodological difference between driver-customer matching problems and vehicle rebalancing problems. The driver-customer matching problem is typically solved by an agent-based model, where each driver and customer are considered individually. For the vehicle rebalancing problem, most methods divide the study area into several sub-regions and the vehicle rebalancing problem is solved at an aggregate level, where vehicles are rebalanced between sub-regions.

To resolve this methodological difference, we propose the matching-integrated vehicle rebalancing (MIVR) model where the area partitioning method is retained and the matching component is modeled at an aggregate level. The objective of the MIVR model is to minimize the total VMT and the number of unsatisfied requests. The aggregate matching component of the MIVR model provides a satisfying approximation of the vehicle pickup distance and the number of unsatisfied requests when using small regions.

Figure 2-1 provides a toy example to illustrate the benefits of the MIVR model compared to an independent vehicle rebalancing (VR) model, where the service availability is represented by the absolute difference between estimated future demand and supply. There are 16 unit squares (sub-regions) and a trip request is equally likely to appear in any of the orange sub-regions in the next time interval. For the independent rebalancing scenario, the rebalancing distance is 2 and the expected pick-up

distance is 3 (four possible pick-up distances 0, 3, 4, 5 with $\frac{1}{4}$ probability on each case). For the matching-integrated rebalancing scenario, the rebalancing distance is 2 and the expected pick-up distance is 2.0 (four possible pick-up distances 1, 2, 2, 3 with $\frac{1}{4}$ probability on each case). Compared to the independent rebalancing scenario, the matching-integrated rebalancing scenario dispatches the idle vehicle to a location near sub-regions with estimated future demand. This "smart" rebalancing decision compensates for inaccurate future demand estimation by harmonizing vehicle pickup distance across different demand profiles.

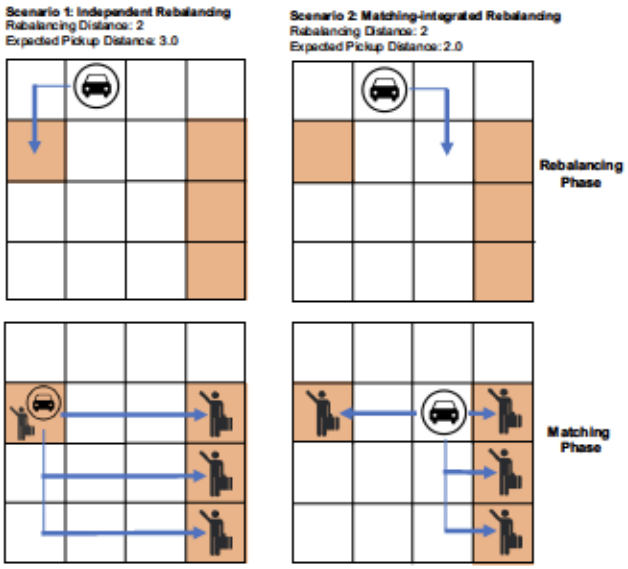


Figure 2-1: Example scenarios comparing regular VR decisions and the MIVR decisions.

To further protect the vehicle rebalancing decisions against demand uncertainty, we introduce robust optimization (RO) techniques to construct a robust MIVR model. Problem-specific uncertainty sets are established to better reflect the uncertainty within ride-hailing demand.

In short, the ride-hailing matching process and RO techniques can be incorporated into the rebalancing procedure to produce better vehicle rebalancing decisions for platforms when facing demand uncertainty. The contributions of this chapter can be summarized as follows:

- Proposing the MIVR model to incorporate driver-customer matching informa-

tion to improve vehicle rebalancing problems with explicit modeling of unsatisfied requests for the first time, to the best of authors' knowledge.

- Proposing the robust MIVR model to consider demand uncertainty and designing problem-specific uncertainty sets to better reflect the inherent demand uncertainty in the ride-hailing system.

- Using simulations to show performance improvements of the MIVR model compared to an independent VR model and a state-of-the-art empty-car routing policy with real demand data and travel times from New York City (NYC). In high supply scenarios, a *Pareto* improvement can be found for the MIVR model when compared to the VR model at aggregate level regarding the overall VMT, the average customer wait time and the number of unsatisfied requests.

- Comparing the nominal MIVR and the robust MIVR under multiple uncertain scenarios by solving a driver-customer matching problem with realized demand and vehicle distributions after rebalancing. The robust MIVR model is shown to perform better under demand uncertainty, especially in conditions of high supply relative to demand.

The remainder of the chapter is organized as follows. Section 2.2 reviews the relevant literature. Section 2.3 describes the nominal and robust MIVR models and the robust counterpart. Section 2.4 includes the empirical study design and descriptions for data used in this chapter. Benchmark comparisons, scenario testing results and robust solution performances are described in Section 2.5. Finally, Section 2.6 recaps the main contributions of this chapter, outlines the limitations and provides future research directions.

2.2 Existing Literature

2.2.1 Ride-hailing Matching and Rebalancing

Ride-hailing matching is a variant of the classical Dial a Ride Problem, where customer trips are matched with vehicles such that generalized costs are minimized. These costs can include VMT, customer wait time, and penalties for poor service quality. Development of new algorithms for this problem is a very active field of research and the methods have been used by platform operators in practice [176]. Agatz et al. [1] provided a comprehensive survey of literature related to optimization of driver-to-passenger for dynamic ride sharing between travelers with similar itineraries. In a more recent survey, Mourad et al. [127] reviews research related to optimization of shared mobility systems more broadly, which includes ride-hailing. The authors identify demand uncertainty as a critical issue in modeling shared mobility systems, and identify stochastic programming and multi-scenario optimization as two possible modeling techniques. Finally, Ho et al. [80] presents an overview of recent research relating to the general Dial a Ride Problem. While this survey is focused on applications such as paratransit and demand-responsive transit, the taxonomy and solution techniques are applicable to ride-hailing problems. Like Mourad et al. [127], the authors find that the development of models and solution methods that include stochastic demand is an important research direction.

Ride-hailing is one type of on-demand service platform, which is characterized by the waiting time sensitivity of customers and service providers without fixed work schedules. Other on-demand service platforms include food and goods delivery services such as DoorDash and Uber Eats, and ride-pooling platforms. Several recent papers have examined the dynamics of on-demand service platforms. It has been shown that customers' sensitivity to delay has a significant impact on optimial pricing and wage setting [151]. Another paper determines optimal prices and wages under different levels of demand, and calibrates parameters using actual ride-hailing data [6]. Cachon et al. [33] develops a model for dynamic pricing in on-demand platforms, demonstrating that such policies benefit stakeholders by expanding access to service

during periods of peak demand. Theoretical relationships between pricing, demand and detour policies within ride-pooling platforms, which are similar to ride-hailing but with the possibility of shared trips, have also been investigated [91].

Given the size and dynamic nature of the ride-hailing matching problem in large cities, many approaches involve metaheuristic methods to generate sub-optimal solutions [147, 53]. Recently, researchers have investigated the role of matching radii and matching time periods on the optimal solution [181]. Lyu et al. [116] develops an online matching algorithm that considers multiple objectives, and provides a theoretical optimality guarantee for the online solution. Xu et al. [176] proposed a dynamic programming approach to matching that seeks to optimize matching decisions over a long time horizon. Their method, which did not consider demand uncertainty, has been adopted by a leading ride-hailing platform.

Optimal rebalancing of idle ride-hailing vehicles has shown to substantially improve system performance. Typical considerations in designing a rebalancing algorithm are the duration of the decision period and the costs included in the objective function. Chen and Levin [45] proposed a simple linear programming (LP) model to select vehicle rebalancing flows that minimize travel cost for five minute periods. Zhang et al. [195] showed that a stable predictive control algorithm could be used for dispatching and rebalancing an autonomous ride-hailing fleet in a discrete time system. At each decision period, a mixed-integer linear programming (MILP) is solved to minimize rebalancing travel time. Their method produced significant reductions in peak wait times compared to the no rebalancing scenario. Similarly, Iglesias et al. [84] proposed a model predictive control algorithm for operating the ride-hailing system in real-time by leveraging short-term demand forecasts. They utilized the Long Short-Term Memory (LSTM) neural networks to forecast future customer demand for each origin and destination pair and their proposed algorithm outperformed a state-of-the-art rebalancing strategy by reducing up to 89.6% of the average customer wait time. Wallar et al. [164] developed an online vehicle rebalancing algorithm that discretized an area into optimal rebalancing sub-regions, resulting in an average wait time reduction of 37% compared to the scenario without rebalancing idle vehicles. Braverman

et al. [30] formulates a fluid-based optimization model for idle vehicle rebalancing in ride-hailing systems. The authors use a nine-region network and real-life ride-hailing data to show how the fluid-based model results in a higher fraction of passengers served compared to benchmark models. We include the Braverman et al. [30] model as a benchmark to test the results of our own model.

Al-Kanj et al. [2] combined the matching and vehicle rebalancing into a single dynamic programming method for autonomous electric vehicles. Their approach employs incentives rather than centralized control to rebalance vehicles, meaning that rebalancing decisions made by the platform are subject to some amount of non-responsiveness by the passenger or vehicle. Dandl et al. [48] also solves for matching and rebalancing decisions in a single optimization model to inform a simulation that tests how demand forecasting accuracy affects the system performance. The authors use an agent-based model, where the objective function is a combination of penalties and rewards for matching and for reducing demand-supply imbalances. Their simulation assumes that all requests are served and customers will wait indefinitely for pickup. In contrast, our method includes matching information and explicitly models customer wait time and unsatisfied requests in order to make rebalancing decisions.

In addition to optimization methods, machine learning (ML) approaches have been proposed to predict demand in rebalancing vehicles [172, 64]. There has also been considerable work on other practical methods, beyond explicit vehicle rebalancing, to achieve greater balance between supply and demand in ride-hailing systems. These methods include dynamic pricing [33, 179], providing more information to drivers [100], reward schemes [182], alternative market structures [189] and carpooling incentives [89].

2.2.2 Robust Optimization

RO is a common approach to handle data uncertainty in optimization problems. The general approach is to specify a range for an uncertain parameter (the "uncertainty set"), and optimize over the worst-case realizations within the bounded uncertainty set. The method is therefore well suited to applications where there is considerable

uncertainty related to the model input parameters, and when data uncertainties can lead to significant penalties or infeasibility in practice. The solution method for robust optimization problems involves generating a deterministic equivalent, called the robust counterpart. Computational tractability of the robust counterpart has been a major practical difficulty [12]. A variety of uncertainty sets have been identified for which the robust counterpart to a robust optimization problem is reasonably tractable [15].

The RO field has grown substantially over the past two decades. Seminal papers in the late 1990s [13, 14] and early 2000s [22] established the field. Comprehensive surveys on the early literature were done by Ben-Tal et al. [12] and Bertsimas et al. [15]. The development of the robust optimization technique has allowed researchers to tackle problems with data uncertainty in a range of fields. Examples can be found for renewable energy network design [174], supply chain operations [117] and health care logistics [170].

2.2.3 Applications of RO in Ride-hailing Operations

In recent years, robust optimization applications in transportation, and ride-hailing rebalancing more specifically, have attracted considerable research attention. Liu et al. [111] considered uncertain local demand in their matching algorithm for ridesharing operations. Miao et al. [122] proposed an RO model for the taxi dispatching problem and tested it using NYC taxi data. They also proposed a data-driven approach to construct the uncertainty set based on historical demand data with a probability guarantee, building on previous data-driven RO theory proposed by Bertsimas et al. [18]. He et al. [78] tackled the robust ride-hailing rebalancing problem using linear decision rules (LDR) to create a multi-period adaptive RO (ARO) model. Their ARO-based approach is heavily based upon theory developed by Bertsimas et al. [23]. To the best of authors' knowledge, no existing papers have incorporated matching component and robust optimization techniques into vehicle rebalancing problems. This research gap is important to address given the prominent role of ride-hailing in urban transportation. Demonstrating how robustness and matching-integrated

rebalancing can be combined in ride-hailing operations, and evaluating whether this combination of methods is advantageous, can help to improve future ride-hailing operations.

2.3 Methodology

2.3.1 Problem Description

Given an operation period $\mathcal{T}$, we first divide it into Ω identical time intervals indexed by $k = 1, 2, ..., \Omega$, where the length of each time interval is Δ[1]. Figure 2-2 displays the framework of the MIVR model. Grey intervals indicate past time intervals that have been optimized. The green interval represents the current decision time interval and red intervals stand for look-ahead time within the MIVR model. The MIVR model is solved in a rolling-horizon manner, where decision variables are determined repeatedly at the beginning of each time interval. At the beginning of time interval k, κ future time intervals are incorporated in the MIVR model, and only the vehicle rebalancing decisions of the current time interval k are implemented. When proceeding to the next time interval, vehicle locations are observed and updated as the input for the MIVR model. Let $(k, k+1, ..., k+\kappa-1)$ represent time intervals considered at time k, to simplify the notation, these time intervals are indexed by $k = 1, 2, ..., \kappa$. The study region is partitioned into n sub-regions, each sub-region i has an estimated demand $r_i^k \geq 0$ at time k. We define the following two sets: $N = \{1, ..., n\}$ representing the set of sub-regions and $K = \{1, ..., \kappa\}$ representing the set of time intervals considered in the problem.

The MIVR model introduces the driver-customer matching component into the vehicle rebalancing problem by considering interzonal matchings based on estimated demand. Within a time interval k, the vehicle rebalancing phase happens at the beginning of the interval and the driver-customer matching phase is conducted at the end of the interval. In the vehicle rebalancing phase, decision variables are represented

[1]The choice of Δ should depend on the size of sub-regions.

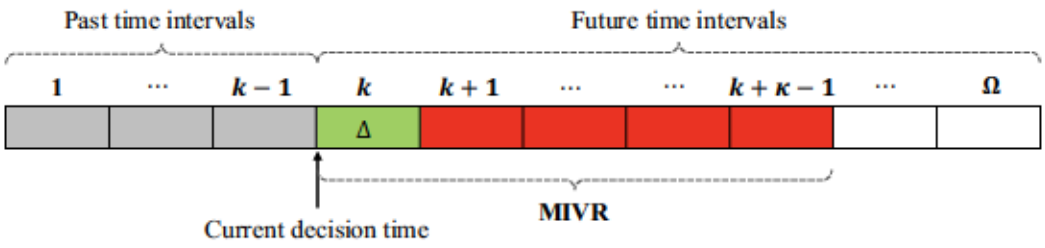


Figure 2-2: MIVR model framework. Each time interval has length Δ.

by $x_{ij}^k \in \mathbb{N}$ denoting the number of idle vehicles rebalanced from sub-region i to sub-region j at time k. Let $S_i^k \in \mathbb{N}$ indicate the number of available vehicles in sub-region i at time k for the matching phase. Let d_{ij}^k, w_{ij}^k denote the travel distance and time from sub-region i to sub-region j at time k, respectively, which can be approximated by the distance and travel time between the centroids of two sub-regions. We define a parameter $a_{ij}^k \in \{0, 1\}$ denoting whether an idle vehicle can be rebalanced from sub-region i to sub-region j at time k, where $a_{ij}^k = 0$ if rebalancing between sub-regions i, j is feasible at time k. The vehicle rebalancing from sub-region i to sub-region j at time k is feasible if $w_{ij}^k \leq \Delta$, stipulating that the vehicle can be rebalanced to the destination sub-region j within time interval k. Then the feasibility constraint of rebalancing between sub-regions is given by:

$$a_{ij}^k \cdot x_{ij}^k = 0 \quad \forall i, j \in N, \ \forall k \in K. \tag{2.1}$$

This constraint does not prevent long-distance rebalancing decisions that occur over several time periods, but rather limits the movement of rebalancing vehicles within a single time period to zones that are reachable within that time period.

In the driver-customer matching phase, matching is considered between sub-regions without considering actual demand and detailed locations of customers and vehicles. Let $y_{ij}^k \in \mathbb{N}$ denote the number of customers in sub-region i matched with vehicles in sub-region j at time k. It is worth mentioning that decision variables y_{ij}^k of the matching component only serve as auxiliary variables in the MIVR model, which focuses on computing the rebalancing decisions. When vehicle are rebalanced and requests are collected, the driver-customer matching problem can then be solved by

a separate driver-customer matching problem given the realized demand. Let $T_i^k \in \mathbb{N}$ denote the number of unsatisfied requests in sub-region i at time k. Then constraints related to the matching phase are:

$$\sum_{j=1}^{n} y_{ji}^k \leq S_i^k \quad \forall i \in N,\ \forall k \in K \tag{2.2a}$$

$$\sum_{j=1}^{n} y_{ij}^k \leq r_i^k \quad \forall i \in N,\ \forall k \in K \tag{2.2b}$$

$$T_i^k = r_i^k - \sum_{j=1}^{n} y_{ij}^k \quad \forall i \in N,\ \forall k \in K \tag{2.2c}$$

Constraints (2a) and (2b) restrict the interzonal matching decisions by the number of available vehicles S_i^k and estimated demand r_i^k. Constraints (2c) define the number of unsatisfied requests, which is equivalent to the number of customers who have not been assigned drivers within the current matching phase. When matching customers and drivers, a maximum pickup time constraint is imposed to guarantee that customers do not experience excessive wait times. Let $\bar{w}$ denote customers' maximum pickup time and parameter $b_{ij}^k \in \{0, 1\}$ denote whether customers in sub-region i can be matched with drivers in sub-region j at time k, where $b_{ij}^k = 0$ indicates a feasible interzonal matching. The matching between customers in sub-region i and drivers in sub-region j at time k is feasible if $w_{ji}^k \leq \bar{w}$, which enforces the maximum pickup time constraint. The matching feasibility constraint is then

$$b_{ij}^k \cdot y_{ij}^k = 0 \quad \forall i, j \in N,\ \forall k \in K. \tag{2.3}$$

Next, we establish the connection between the two phases. Let $V_i^k, O_i^k \in \mathbb{N}$ represent the number of vacant and occupied vehicles for sub-region i at the beginning of time interval k, respectively. The initial vehicle locations, $V_i^1, O_i^1, \forall i \in N$, are inputs for the MIVR model. Other inputs to the model are regional transition matrices P^k, Q^k, which describe the dynamics of occupied vehicles. The entry (i, j) for P^k, P_{ij}^k, denotes the probability that an occupied vehicle located in sub-region i at time

k will be in sub-region j and stay occupied at time $k+1$. The entry (i, j) for Q^k, Q^k_{ij}, indicates the probability that an occupied vehicle starting in sub-region i at time k will be in sub-region j and become vacant at time $k+1$.

In reality, the regional transition matrices depend on the spatio-temporal demand flows as well as the operator's dispatching and rebalancing strategies. The matching and rebalancing decisions in the MIVR model are defined at interzonal level, and the regional transition matrices formulated with interzonal level decision variables are approximations to the real matrices. To reduce the model complexity, we further approximate the real regional transition matrices with static matrices estimated from the historical data. The impact of utilizing static transition matrices will be elaborated in the results section. These matrices must satisfy the following constraints:

$$\sum_{j=1}^{n}(P^k_{ij} + Q^k_{ij}) = 1, \quad \forall i \in N,\ \forall k \in K.$$

Then, we specify the following relationships between S^k_i, V^k_i, O^k_i and decision variables x^k_{ij}, y^k_{ij}:

$$\sum_{j=1}^{n} x^k_{ij} \leq V^k_i \quad \forall i \in N,\ \forall k \in K \tag{2.4a}$$

$$S^k_i = V^k_i + \sum_{j=1}^{n} x^k_{ji} - \sum_{j=1}^{n} x^k_{ij} \quad \forall i \in N,\ \forall k \in K \tag{2.4b}$$

$$V^{k+1}_i = S^k_i - \sum_{j=1}^{n} y^k_{ji} + \sum_{j=1}^{n} Q^k_{ji} O^k_j \quad \forall i \in N,\ \forall k \in K \setminus \{\kappa\} \tag{2.4c}$$

$$O^{k+1}_i = \sum_{j=1}^{n} y^k_{ji} + \sum_{j=1}^{n} P^k_{ji} O^k_j \quad \forall i \in N,\ \forall k \in K \setminus \{\kappa\} \tag{2.4d}$$

Where constraints (4a) ensure that the number of vehicles in sub-region i that can be rebalanced to other sub-regions is bounded by the number of vacant vehicles. Constraints (4b) show that the available vehicles in sub-region i at time k consist of vacant and rebalanced vehicles. Similarly, constraints (4c) indicate that the set of vacant vehicles in sub-region i at time $k+1$ is comprised of currently vacant vehicles at

time k and currently occupied vehicles that become vacant in the next time interval. The number of unmatched vehicles at time k, denoted by $S_i^k - \sum_{j=1}^n y_{ji}^k$, is equal to the difference between the number of available vehicles and the number of vehicles dispatched for interzonal matching. The number of occupied vehicles at time k that become vacant at time $k+1$ in sub-region i is represented by $\sum_{j=1}^n Q_{ji}^k O_j^k$. Constraints (4d) state that occupied vehicles in sub-region i at time $k+1$ are comprised of currently vacant vehicles that become occupied in the next interval as well as currently occupied vehicles at time k. The number of vacant vehicles that become occupied in sub-region i at time $k+1$ because of interzonal matching at time k is indicated by $\sum_{j=1}^n y_{ji}^k$. The number of occupied vehicles at time k that stay occupied at time $k+1$ in sub-region i is enforced by $\sum_{j=1}^n P_{ji}^k O_j^k$.

The objective for the MIVR model is minimizing the number of unsatisfied requests and the total vehicle travel distance, which consists of vehicle rebalancing distance and vehicle pickup distance. To construct the objective function as the generalized VMT for ride-hailing operations, we assume γ to be a parameter indicating the penalty VMT induced by each unsatisfied request. Let β represent a parameter that defines the relative weighting of rebalancing distance and pickup distance. The parameter β controls the trade-off between the total non-occupied VMT (from the system perspective) and the service quality (from the customer perspective). A larger β indicates a higher priority on minimizing the vehicle pickup distance, which leads to better service quality with a smaller customer wait time. When $\beta = 1$, the MIVR model purely minimizes the total VMT and the number of unsatisfied requests without explicitly putting any weight on the customer wait times[2].

$$(MIVR) \quad \min \quad Z = \sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^k d_{ij}^k + \beta \cdot \sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}^k d_{ji}^k + \gamma \cdot \sum_{k=1}^{\kappa}\sum_{i=1}^{n} T_i^k \tag{2.5a}$$

$$\text{s.t.} \quad \text{Constraints } (1), (2a) - (2c), (3), (4a) - (4d)$$

[2]The MIVR model implicitly weights the customer wait times because of the correlation between the vehicle pickup distance and wait times.

$$x_{ij}^k, y_{ij}^k \in \mathbb{N} \quad \forall i, j \in N, \ \forall k \in K \tag{2.5b}$$

$$S_i^k, V_i^k, O_i^k, T_i^k \in \mathbb{N} \quad \forall i \in N, \ \forall k \in K \tag{2.5c}$$

The MIVR model is an integer linear programming (ILP) problem with integer variables x_{ij}^k, y_{ij}^k, S_i^k, V_i^k, O_i^k and T_i^k. ILP problems of this size and complexity can be difficult to solve in a reasonable time frame. To improve the computational performance of our model while producing satisfying results, we relax all integer variables in the problem to positive real numbers $\mathbb{R}^+$. The rebalancing decisions used for implementations can be generated by rounding down the solutions generated by the relaxed model. The approximated rebalancing decisions are guaranteed to be feasible regarding to constraints (4a), which impose an upper-bound on the number of vehicles that can be rebalanced.

By incorporating matching decisions within vehicle rebalancing problem, the model also considers future matching distances in addition to the rebalancing distance, leading to "smarter" rebalancing decisions. Essentially, the MIVR reduces the cost of inaccurate demand estimation when rebalancing idle vehicles. Meanwhile, the MIVR model is a forward-looking model by incorporating κ future time intervals into the model.

2.3.2 Robust Optimization Model Formulation

The estimation of the future demand r_i^k is crucial for vehicle rebalancing problems in ride-hailing systems. Previous studies have assumed the number of customers in any sub-region followed a Poisson distribution [164, 172]. However, in most applications we have limited knowledge about the "true" distribution for the future demand. The assumption that complex customer behaviour can be described by a simple probability distribution might be too strong. Instead of imposing a probability distribution on the future demand, we introduce the robust optimization technique where the uncertain demand parameters are described by uncertainty sets rather than specific probability distributions. The uncertainty sets specify a range for the uncertain demand where

the demand can lie anywhere in the range.

First, we define the uncertainty set for the robust MIVR model. For the uncertainty in the demand originating in sub-region i within time interval k, we construct an uncertainty set $\mathcal{U}$ from the intersection of two different sets: a box uncertainty set $\tilde{\mathcal{U}}_i^k$ and a polyhedral uncertainty set $\bar{\mathcal{U}}^k$ which constrains the total variation in demand across all sub-regions. The uncertainty set $\mathcal{U}$ was selected to reflect the actual range of demand variability across different sub-regions without producing solutions that are too conservative in practice.

The box uncertainty set imposes upper and lower bounds of ρ standard deviations between estimated regional demand and the mean regional demand at each time interval k. The parameter ρ is set according to the operator's level of risk tolerance, with a higher ρ representing a lower tolerance for risk. The mean μ_i^k and standard deviation σ_i^k of the demand in sub-region i during time k are estimated with the historical data. The box uncertainty set for estimated demand r_i^k is then

$$\tilde{\mathcal{U}}_i^k(\rho) = \left\{ r_i^k : \left| \frac{r_i^k - \mu_i^k}{\sigma_i^k} \right| \leq \rho \right\} \quad \forall i \in N,\ \forall k \in K.$$

The polyhedral uncertainty set limits the total offset in the sum of the demand during a time interval across all sub-regions. This second restriction is intuitive; within a given time interval, demand may be above or below the mean in one region, but the total demand across the entire service area could be expected to remain at a similar level compared to previous days under most scenarios. Sub-regions with unusually high demand should be offset by other nearby sub-regions of low demand. The polyhedral uncertainty set for estimated demand r_i^k is

$$\bar{\mathcal{U}}^k(\Gamma) = \left\{ (r_1^k, ..., r_n^k) : \left| \sum_{i=1}^{n} (r_i^k - \mu_i^k) \right| \leq \Gamma \right\} \quad \forall k \in K,$$

where Γ is the parameter to control the level of uncertainty for the polyhedral uncertainty set. It is worth noting that the construction of the uncertainty set indicates how much uncertainty the operator would like to tolerate in the operation. In reality, there exists scenarios where the total demand at certain time intervals

exceed the historical mean by far, for instance ride-hailing demand after concerts or large events. It is wise for the ride-hailing operator to not take such unusual demand scenarios into consideration.

The combined uncertainty set $\mathcal{U}$ for the estimated demand r_i^k is:

$$\mathcal{U} = \left[\bigcap_{i=1}^{n}\bigcap_{k=1}^{\kappa} \tilde{\mathcal{U}}_i^k(\rho)\right] \cap \left[\bigcap_{k=1}^{\kappa} \mathcal{U}^k(\Gamma)\right]$$

By defining an uncertain parameter $\zeta \in \mathbb{R}^{n\kappa}$ and letting $r_i^k = \mu_i^k + \zeta_i^k \sigma_i^k$, we can write $\mathcal{U}$ as follows:

$$\mathcal{U} = \left\{\zeta : \|\zeta\|_\infty \leq \rho;\ \left|e^T(\zeta^k \circ \sigma^k)\right| \leq \Gamma, \forall k \in K\right\}, \tag{2.6}$$

where $\boldsymbol{\zeta^k}, \sigma^k \in \mathbb{R}^n$ are vectors for a specific time interval k, $e \in \mathbb{R}^n$ is a vector with all entries equal to one, and $\zeta^k \circ \sigma^k$ indicates the element-wise product for vectors ζ^k and σ^k. The parameters ρ and Γ control the size of the uncertainty set for estimated demand, and can be adjusted based on the operators' risk tolerance or desired probability guarantee for constraints involving uncertain parameters. Increasing the value of ρ and Γ leads to more conservative rebalancing decisions for the robust model.

Combining the MIVR model with the uncertainty set described above, we propose a robust MIVR model:

$$(P) \quad \min_{x_{ij}^k, y_{ij}^k} \quad Z = \sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^k d_{ij}^k + \beta \cdot \sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}^k d_{ji}^k + \gamma \cdot \sum_{k=1}^{\kappa}\sum_{i=1}^{n} T_i^k \tag{2.7a}$$

$$\text{s.t.} \quad S_i^k = V_i^k + \sum_{j=1}^{n} x_{ji}^k - \sum_{j=1}^{n} x_{ij}^k \quad \forall i \in N,\ \forall k \in K \tag{2.7b}$$

$$V_i^{k+1} = S_i^k - \sum_{j=1}^{n} y_{ji}^k + \sum_{j=1}^{n} Q_{ji}^k O_j^k \quad \forall i \in N,\ \forall k \in K \setminus \{\kappa\} \tag{2.7c}$$

$$O_i^{k+1} = \sum_{j=1}^{n} y_{ji}^k + \sum_{j=1}^{n} P_{ji}^k O_j^k \quad \forall i \in N, \forall k \in K \setminus \{\kappa\} \tag{2.7d}$$

$$\sum_{j=1}^{n} x_{ij}^k \leq V_i^k \quad \forall i \in N,\ \forall k \in K \tag{2.7e}$$

$$\sum_{j=1}^{n} y_{ji}^k \leq S_i^k \quad \forall i \in N,\ \forall k \in K \tag{2.7f}$$

$$\sum_{j=1}^{n} y_{ij}^k \leq \mu_i^k + \zeta_i^k \sigma_i^k \quad \forall i \in N,\ \forall k \in K,\ \forall \boldsymbol{\zeta} \in \mathcal{U} \tag{2.7g}$$

$$T_i^k = \mu_i^k + \zeta_i^k \sigma_i^k - \sum_{j=1}^{n} y_{ij}^k \quad \forall i \in N,\ \forall k \in K,\ \forall \boldsymbol{\zeta} \in \mathcal{U} \tag{2.7h}$$

$$b_{ij}^k \cdot y_{ij}^k = 0 \quad \forall i \in N,\ \forall k \in K \tag{2.7i}$$

$$a_{ij}^k \cdot x_{ij}^k = 0 \quad \forall i \in N,\ \forall k \in K \tag{2.7j}$$

$$x_{ij}^k, y_{ij}^k \geq 0 \quad \forall i, j \in N,\ \forall k \in K \tag{2.7k}$$

$$S_i^k, V_i^k, O_i^k, T_i^k \geq 0 \quad \forall i \in N,\ \forall k \in K \tag{2.7l}$$

The problem (P) becomes infeasible even with a small value of ρ if the coefficient of variation[3] for uncertain demand is large for some sub-regions during certain time intervals. Particularly, the problem (P) is infeasible if $\exists i \in N, \exists k \in K$ and $\rho \geq \frac{\mu_i^k}{\sigma_i^k}$. Because when inequality $\rho \geq \frac{\mu_i^k}{\sigma_i^k}$ holds, the box uncertainty set $\tilde{\mathcal{U}}_i^k(\rho)$ allows ζ_i^k to take values smaller than $-\frac{\mu_i^k}{\sigma_i^k}$, which leads to a negative uncertain demand, i.e., $r_i^k = \mu_i^k + \zeta_i^k \sigma_i^k < 0$. The constraint (7g) is infeasible when the right-hand side is negative since the decision variable y_{ij}^k is non-negative. To prevent infeasibility that can results from demand uncertainty, we add restrictions on the uncertainty set in the problem (P) to guarantee that estimated demand is non-negative:

$$\mu_i^k + \zeta_i^k \sigma_i^k \geq 0 \quad \forall i \in N,\ \forall k \in K,\ \forall \boldsymbol{\zeta} \in \mathcal{U} \tag{2.8}$$

When modeling robust optimization problems, equality constraints with uncertain parameters should be avoided as much as possible since they dramatically shrink the feasible region and often lead to infeasibility [61]. For the problem (P) with uncertain parameter ζ, we must therefore reformulate equality constraints (7h). Equality constraints (7h) can be avoided by eliminating variable T_i^k through substitution. After this variable elimination step, objective function of problem (7a) becomes:

[3]Ratio of the standard deviation to the mean.

$$\min_{x_{ij}^k, y_{ij}^k} \left\{ \sum_{k=1}^{\kappa} \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij}^k d_{ij}^k + \beta \cdot \sum_{k=1}^{\kappa} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{ij}^k d_{ji}^k + \max_{\zeta \in \mathcal{U}} \left[\gamma \cdot \sum_{k=1}^{\kappa} \sum_{i=1}^{n} (\mu_i^k + \zeta_i^k \sigma_i^k - \sum_{j=1}^{n} y_{ij}^k) \right] \right\}. \tag{2.9}$$

The objective function (9) with min-max formulation can be reformulated by introducing an auxiliary variable ω:

$$\min \quad Z = \omega \tag{2.10a}$$

$$\text{s.t.} \quad \sum_{k=1}^{\kappa} \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij}^k d_{ij}^k + \sum_{k=1}^{\kappa} \sum_{i=1}^{n} \sum_{j=1}^{n} (\beta \cdot d_{ji}^k - \gamma) y_{ij}^k + \gamma \cdot \sum_{k=1}^{\kappa} \sum_{i=1}^{n} (\mu_i^k + \zeta_i^k \sigma_i^k) \leq \omega \quad \forall \zeta \in \mathcal{U} \tag{2.10b}$$

However, robust counterparts for equivalent formulations of the same problem are not necessarily equivalent [61]. To reformulate the problem while maintaining an identical robust counterpart, we make variables T_i^k *adaptive*, meaning that both variables are "wait-and-see"[4] variables relating to uncertain parameters ζ, i.e., $T_i^k = T_i^k(\zeta)$. Introducing adaptive variables turns the initial RO problem into an Adaptive Robust Optimization (ARO) problem. A commonly-used approximation method for solving ARO problems is the application of Linear Decision Rules (LDRs), which has been shown to perform well in practice [12, 16]. Also, if the coefficients for the variables to be eliminated in the equality constraint do not include uncertain parameters and the constraint is linear in the uncertain parameters, making such variables adaptive and applying LDRs is equivalent to directly eliminating them [61]. Substitutions with equality constraint (7h) satisfies both conditions, therefore we eliminate variables T_i^k in the problem (P) to ensure no uncertain parameters appear in equality constraints. The reformulation (P') is equivalent to an approximation for the original robust formulation (P) together with restriction (8) on the uncertainty set by applying LDRs:

[4]The value of "wait-and-see" variables are determined only after the future demand is revealed.

$$(P') \quad \min \quad Z = \omega \tag{2.11a}$$

$$\text{s.t.} \quad \sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^k d_{ij}^k + \sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n}(\beta \cdot d_{ji}^k - \gamma) y_{ij}^k + \gamma \cdot \sum_{k=1}^{\kappa}\sum_{i=1}^{n}(\mu_i^k + \zeta_i^k \sigma_i^k) \leq \omega \quad \forall \zeta \in \mathcal{U} \tag{2.11b}$$

$$\text{Constraints } (7b) - (7g), (7i) - (7l), (8)$$

After the reformulation, uncertain parameters only appear in the constraints. The next step is to derive the robust counterpart for the robust MIVR model. Constraints (7g), (11b) and Equation (8) with uncertain parameter ζ can be written as the following generic formulation:

$$L(\cdot) + v^T \zeta \leq c \quad \forall \zeta \in \mathcal{U}, \tag{2.12}$$

where $L(\cdot)$ indicates a function of decision variables in problem (P'), v is a vector in dimension $n\kappa$ and c is a scalar. The robust counterpart for the generic constraint (2.12) is

$$\begin{cases} L(\cdot) + \rho \left\| \theta_0 \right\|_1 + \Gamma \sum_{k=1}^{\kappa} (\eta_1^k + \eta_2^k) \leq c \\ (\eta_1^{k'} - \eta_2^{k'}) \sigma_i^{k'} = \theta_{k'}^{i,k} \quad \forall i \in N, \ \forall k = k' \in K \\ \theta_{k'}^{i,k} = 0 \quad \forall i \in N, \ \forall k \neq k' \in K \\ \eta_1^k, \eta_2^k \geq 0 \quad \forall k \in K \\ \sum_{k=0}^{\kappa} \theta_k = v \end{cases} \tag{2.13}$$

Where $\theta_k \in \mathbb{R}^{n\kappa}$ and $\theta_{k'}^{i,k}$ represents (ik)-th entry of vector $\theta_{k'}$, $\forall k' \in K$. The full derivation of the generic robust counterpart of (2.12) can be found in Appendix A.1. Then we derive the robust counterpart for problem (P'):

$$(RC) \quad \min \quad Z = \omega \tag{2.14a}$$

s.t. Constraints $(7b) - (7f), (7i) - (7l)$

$$\sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^{k} d_{ij}^{k} + \sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n}(\beta \cdot d_{ji}^{k} - \gamma) y_{ij}^{k} + \gamma \cdot \sum_{k=1}^{\kappa}\sum_{i=1}^{n} \mu_{i}^{k} + \rho \cdot \sum_{k=1}^{\kappa}\sum_{i=1}^{n} \tilde{\theta}_{0}^{i,k}$$

$$+ \Gamma \cdot \sum_{k=1}^{\kappa}(\eta_{1}^{k} + \eta_{2}^{k}) \leq \omega \tag{2.14b}$$

$$(\eta_{1}^{k'} - \eta_{2}^{k'})\sigma_{i}^{k'} = \theta_{k'}^{i,k} \quad \forall i \in N,\ \forall k = k' \in K \tag{2.14c}$$

$$\theta_{k'}^{i,k} = 0 \quad \forall i \in N,\ \forall k \neq k' \in K \tag{2.14d}$$

$$\sum_{k'=0}^{\kappa} \theta_{k'}^{i,k} = \gamma \cdot \sigma_{i}^{k} \quad \forall i \in N,\ \forall k \in K \tag{2.14e}$$

$$-\tilde{\theta}_{0}^{i,k} \leq \theta_{0}^{i,k} \leq \tilde{\theta}_{0}^{i,k} \quad \forall i \in N,\ \forall k \in K \tag{2.14f}$$

$$\eta_{1}^{k}, \eta_{2}^{k} \geq 0 \quad \forall k \in K \tag{2.14g}$$

$$\sum_{j=1}^{n} y_{ij}^{k} + \rho \sum_{k'=1}^{\kappa}\sum_{i'=1}^{n}(\tau_{1,i,k}^{i',k'} + \tau_{2,i,k}^{i',k'}) + \Gamma \sum_{k'=1}^{\kappa}(\tau_{3,i,k}^{k'} + \tau_{4,i,k}^{k'}) \leq \mu_{i}^{k} \quad \forall i \in N,\ \forall k \in K$$

$$\tag{2.14h}$$

$$\tau_{1,i,k}^{i',k'} - \tau_{2,i,k}^{i',k'} + \sigma_{i'}^{k'}(\tau_{3,i,k}^{k'} - \tau_{4,i,k}^{k'}) = 0 \quad \forall i', i \in N,\ \forall k', k \in K,\ (i',k') \neq (i,k)$$

$$\tag{2.14i}$$

$$\tau_{1,i,k}^{i',k'} - \tau_{2,i,k}^{i',k'} + \sigma_{i'}^{k'}(\tau_{3,i,k}^{k'} - \tau_{4,i,k}^{k'}) = -\sigma_{i}^{k} \quad \forall i' = i \in N,\ \forall k' = k \in K \tag{2.14j}$$

$$\tau_{1,i,k}^{i',k'}, \tau_{2,i,k}^{i',k'} \geq 0 \quad \forall i', i \in N,\ \forall k', k \in K \tag{2.14k}$$

$$\tau_{3,i,k}^{k'}, \tau_{4,i,k}^{k'} \geq 0 \quad \forall i \in N,\ \forall k', k \in K \tag{2.14l}$$

$$\rho \sum_{k'=1}^{\kappa}\sum_{i'=1}^{n}(\nu_{1,i,k}^{i',k'} + \nu_{2,i,k}^{i',k'}) + \Gamma \sum_{k'=1}^{\kappa}(\nu_{3,i,k}^{k'} + \nu_{4,i,k}^{k'}) \leq \mu_{i}^{k} \quad \forall i \in N,\ \forall k \in K$$

$$\tag{2.14m}$$

$$\nu_{1,i,k}^{i',k'} - \nu_{2,i,k}^{i',k'} + \sigma_{i'}^{k'}(\nu_{3,i,k}^{k'} - \nu_{4,i,k}^{k'}) = 0 \quad \forall i', i \in N,\ \forall k', k \in K,\ (i',k') \neq (i,k)$$

$$\tag{2.14n}$$

$$\nu_{1,i,k}^{i',k'} - \nu_{2,i,k}^{i',k'} + \sigma_{i'}^{k'}(\nu_{3,i,k}^{k'} - \nu_{4,i,k}^{k'}) = -\sigma_{i}^{k} \quad \forall i' = i \in N,\ \forall k' = k \in K \tag{2.14o}$$

$$\nu_{1,i,k}^{i',k'}, \nu_{2,i,k}^{i',k'} \geq 0 \quad \forall i', i \in N,\ \forall k', k \in K \tag{2.14p}$$

$$\nu_{3,i,k}^{k'}, \nu_{4,i,k}^{k'} \geq 0 \quad \forall i \in N,\ \forall k', k \in K \tag{2.14q}$$

The constraints (14b) - (14g) represent the robust counterpart of constraints (11b). Constraints (14h) - (14l) are the robust counterpart of constraints (7g) while con-

straints (14m) - (14q) are the robust counterpart of Equation (8). Compared to problem (P'), the robust counterpart (RC) introduces $(4n^2\kappa^2+5n\kappa^2+2n\kappa+2\kappa)$ new auxiliary continuous variables. Although the number of decision variables increases considerably in the robust counterpart, this LP problem can be solved efficiently even for large-scale instances.

2.4 Empirical Study Design

In this section, we describe a real-time ride-hailing simulator used to compare the MIVR model with an independent VR model. To justify the benefit of introducing the robust optimization technique into the vehicle rebalancing problem, a separate matching problem is solved over multiple demand scenarios to evaluate robust solutions and compare the nominal MIVR model with the robust MIVR model. We also describe the data used in the experiments.

2.4.1 Ride-hailing Simulator

The ride-hailing simulator is used to compare the nominal MIVR model with a benchmark VR model described in Appendix A.2. The results produced by this simulator allow us to evaluate the impact of the MIVR model independent of the robust optimization component. The simulation framework is shown in Figure 2-3.

Data Input. Data input for the ride-hailing simulator including the road network for the studied region with a shortest path distance matrix and a predecessor matrix, the set of n sub-regions N, a distance matrix d_{ij}^k and travel time matrix w_{ij}^k between centroids of sub-regions, the set of Ω time intervals with length Δ, a mean μ_i^k of demand for each sub-region during each time interval, a full day of ride-hailing demand, and regional probability transition matrices for occupied vehicles P, Q and vacant vehicles P_v, Q_v. Details of the P, Q matrix estimation methods are provided in Appendix A.4. Data sources are described in detail in Section 2.4.3.

Simulation Parameters. Table 2.1 presents and explains the simulation parameters. Rebalancing decisions are solved with a model considering κ look-ahead time

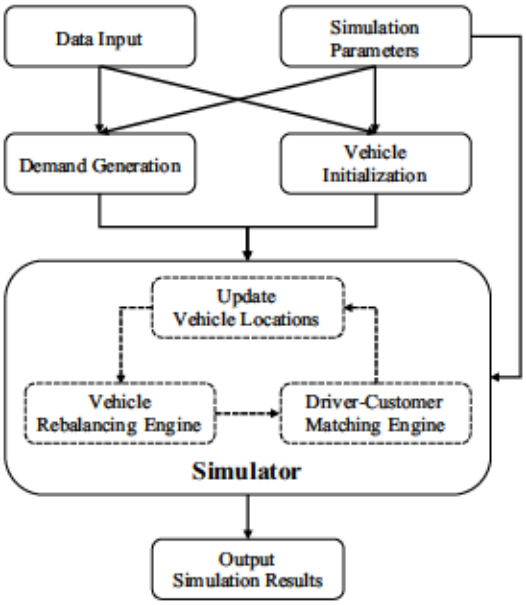


Figure 2-3: Ride-hailing simulation framework.

intervals.

Simulation Parameter	Explanation	Base Case Value
α	Cost parameter for regular rebalancing model	10^2
β	Weight parameter for pickup distance	1
γ	Cost parameter for unsatisfied requests	10^2
T_{start}	Start time of simulation	00:00
T_{end}	End time of simulation	24:00
Δ	Decision time interval length	300 (seconds)
δ	Matching batch size	30 (seconds)
κ	Number of time intervals considered in model	6
$\bar{w}$	Maximum pickup time	300 (seconds)
$\tilde{w}$	Maximum wait time	300 (seconds)
N_v	Number of vehicles	3000
$\bar{v}$	Average vehicle speed	20 (mph)

Table 2.1: Simulation parameters and base case value.

Demand Generation. Due to privacy concerns, historical TNC trip datasets typically do not provide exact addresses or coordinates for trip origins and destinations. Given the demand data at sub-regional level, we randomly assign road nodes within sub-regions as origins and destinations.

Vehicle Initialization. At the start of the simulation period, the N_v vehicles are

equally likely to be in any sub-region i. The initial location for a vehicle within a sub-region i is randomly assigned to a road node within i. All vehicles are considered to be available at the beginning of the simulation.

Simulator. There are two main components contained in the simulator: the vehicle rebalancing engine and driver-customer matching engine. Vehicle locations are updated at the beginning of each simulation iteration. The simulator works as follows: at the beginning of current simulation iteration, vacant and occupied vehicle locations are updated and used as the input for vehicle rebalancing engine; vacant vehicles are rebalanced based on rebalancing decision variables for the current iteration; within each simulation iteration, the driver-customer matching engine can be run multiple times depend on the matching batch size (e.g., 30 seconds); vehicles with assigned customers become occupied and start to pick up customers and finish their trips.

Driver-customer Matching Engine. The optimal assignment problem for matching drivers with customers in the simulator can be found in Appendix A.3. The objective of the optimal assignment problem is minimizing the number of unsatisfied requests while minimizing the pickup distance. The batch size of driver-customer matching engine is δ and customers will leave the ride-hailing system if they wait longer than the maximum wait time $\tilde{w}$.

Simulation Results. We evaluated the simulation with the following vehicle-related indicators: number of served customers, non-occupied VMT and number of rebalancing trips. Customer wait time is used as the customer-related indicator to evaluate the simulation. The customer wait time includes two components: the time for the vehicle to be assigned to the customer, and the time for the assigned vehicle to travel to the pickup location.

2.4.2 Robust Solution Evaluation

Evaluating the solutions from the robust model requires multiple different demand scenarios due to the stochastic inputs. We compare the average performance of the model across all demand scenarios in the study period for different uncertainty set sizes.

To evaluate the model performance under each demand scenario, we solve a separate driver-customer matching problem after the demand is realized and the (nominal or robust) rebalancing decision x_{ij}^k (generated with estimated demand) is executed. The driver-customer matching problem solved here is identical to the one solved in the simulator. The overall pickup time and the number of unsatisfied customers are used as outputs to evaluate robust solutions.

2.4.3 Data Description

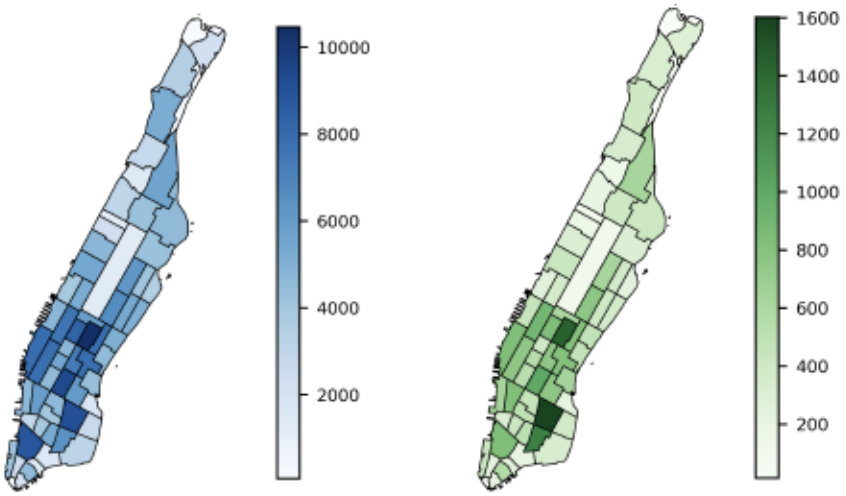
The study area used in the experiments is the island of Manhattan in NYC. We used the high-volume ride-hailing trip data collected by the NYC Taxi and Limousine Commission [130] as the demand data. The sub-regions used in the experiments are "taxi zones" defined within the high-volume ride-hailing trip dataset. There are 63 taxi zones on the island of Manhattan ($N = 63$).

For benchmark comparisons of the nominal MIVR model, weekdays in June 2019 were chosen as the analysis period. Only trips that began and ended on the island of Manhattan were included. The mean and standard deviation of daily trip count by zone are shown in Figure 4-4 to illustrate the overall demand pattern. Demand is generally concentrated around dense residential areas on the eastern and western sides of Manhattan. There was an average of 294,422 high-volume ride-hailing trips per weekday during the sample period.

The full day of ride-hailing demand used in the simulation is from June 10, 2019. We chose a non-holiday Wednesday as it represents a typical day of demand from the study period. Figure 2-5 shows the comparison between the real demand and estimated demand[5] aggregated into 5-minute time intervals. Based on the relationship between total real demand and total estimated demand, we identify four discrete demand scenarios over which the model can be tested:

I Low demand with accurate estimation (0 - 6): overall demand is relatively low and consistent with the historical average for this period.

[5]Mean demand μ_i^k is used as estimated demand.



(a) Mean (b) Standard deviation

Figure 2-4: Average daily demand by zone (trips).

II High demand with accurate estimation (6 - 10): overall demand is high and consistent with the historical average for this period.

III Demand underestimation (11 - 17): the total demand exceeds the historical average for this period.

IV Demand overestimation (20 - 24): the total demand is lower than the historical average for this period.

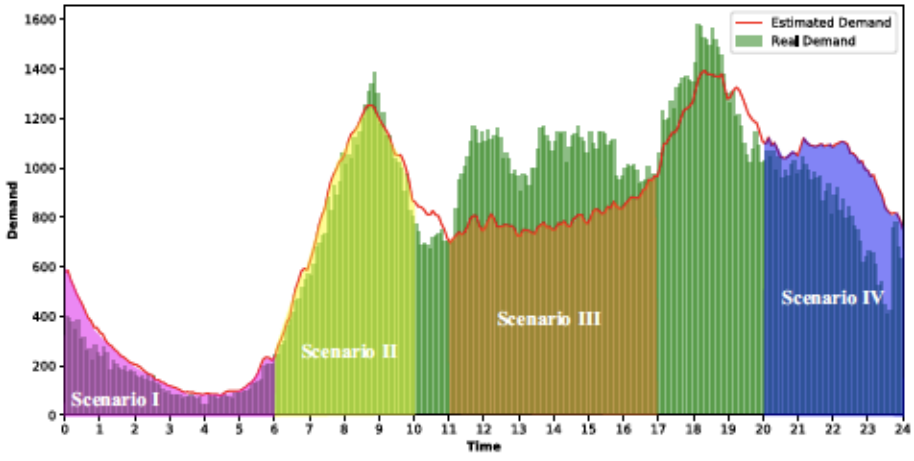


Figure 2-5: Estimated and real demand with four different types of demand scenarios.

It is worth mentioning that an accurate prediction of the total demand does not

lead to accurate sub-regional demand predictions. Demand uncertainties exist in every demand scenario and the overall level of uncertainty is higher in scenarios where demand is underestimated or overestimated. Simulation results for each demand scenario are shown in the next section in order to illustrate the difference in model performance across two dimensions: demand level and prediction accuracy.

For evaluations of the robust MIVR model, we utilized the actual demand data for the 65 week days from April to June 2019 to reflect the real demand uncertainty. Mean μ_i^k and standard deviation σ_i^k used in the robust MIVR model are generated from the same period.

The interzonal travel times for each time interval, w_{ij}^k, were collected from real travel speed data provided by the Uber Movement database for the study period of June 2019 [156]. Hourly link-level travel speed is available for every link with at least five unique trips during the hour. First, the average hourly speed across all days in the study period was determined. The average hourly link travel speed was then used as an input to find the shortest path travel time between each zone pair for each hour in the day. Dijkstra's algorithm [51] was used to determine the shortest path between zone centroids. The regional transition probability matrices for occupied and vacant vehicles, P, Q, P_v and Q_v, are generated based on the real travel time and demand data, and details are shown in Appendix A.4.

2.5 Results

All experiments in this chapter are conducted on a 3.0 GHz AMD Threadripper 2970WX Processor with 128 GB Memory. The integer linear program and linear progra m in the experiments are solved with Gurobi 9.0 [74].

Presentation and discussion of the results is organized into three subsections. Section 2.5.1 compares the MIVR model to two benchmark models: the VR model described earlier, and a recent state-of-the-art rebalancing model [30]. Section 2.5.2 explores the sensitivity of the MIVR results to variation in the model inputs. Section 2.5.3 discusses the impact of regional transition matrices to simulation results.

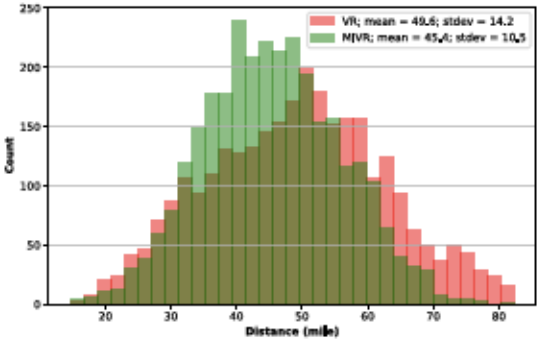
Finally, Section 2.5.4 provides the results for the robust MIVR model.

2.5.1 Benchmark Comparison

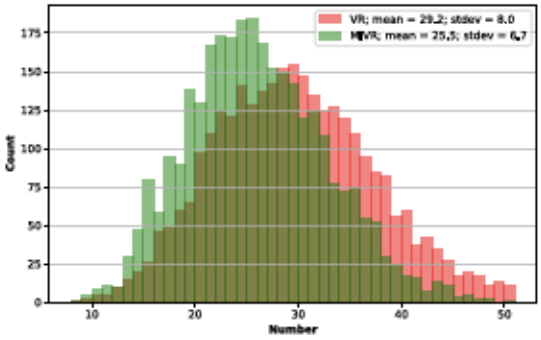
First, we compare the MIVR model with the benchmark VR model described in A.2 and a fluid-based empty-car routing policy (FERP) proposed by Braverman et al. [30]. The performance of each model is assessed with the ride-hailing simulator described in Section 2.4. To ensure a fair comparison, each vehicle rebalancing model uses the same demand profile and initial vehicle locations for each scenario.

Benchmark VR Model Comparison

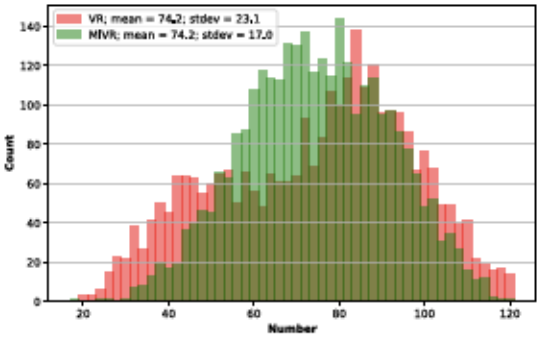
The base case scenario (full-day simulation) is tested with the simulation parameters shown in Table 2.1. The base case considers a scenario with 3000 vehicles, i.e.,



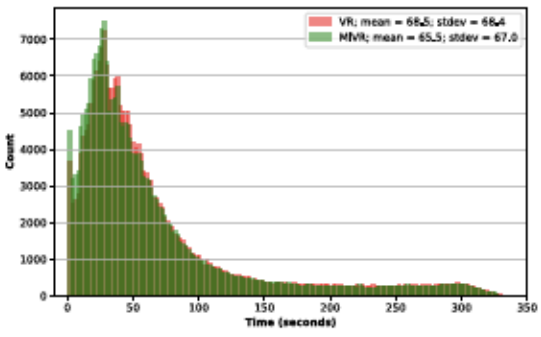
(a) Vehicle non-occupied travel distance distribution



(b) Vehicle rebalancing trip distribution



(c) Number of customers served (per vehicle) distribution



(d) Customer wait time distribution

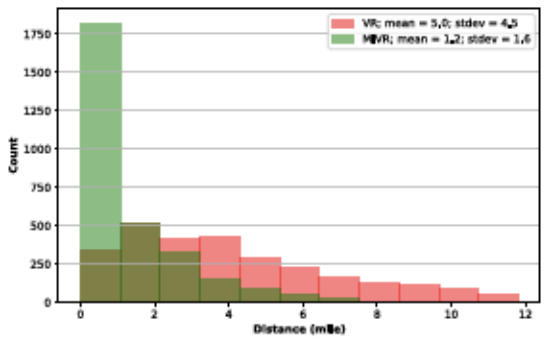
Figure 2-6: Vehicle- and customer-related metrics in the simulation for the base case.

$N_v = 3000$, and 6 future time intervals in the vehicle rebalancing model, i.e., $\kappa = 6$. The base case scenario purely minimizes the number of unsatisfied requests and the total non-occupied VMT, i.e., $\beta = 1$. Both vehicle- and customer-related metrics are presented in Figure 2-6, where each figure shows the distributions for vehicles or customers for both MIVR and VR model results.

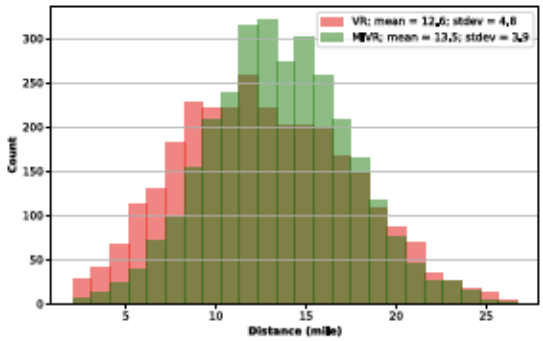
As shown in Figure 2-6a, the MIVR model reduces the non-occupied travel distance on average when compared to the VR model. Also, the number of vehicles with extremely long travel distance is reduced when utilizing the MIVR model. Figure 2-6b displays the rebalancing trip distributions, indicating that the MIVR dispatches fewer vacant vehicles for rebalancing purposes. The distribution of the number of served customers per vehicle is shown in Figure 2-6c. Although the average number of customers served by each vehicle is identical for two models, vehicles utilization is more evenly distributed under the MIVR model compared to the VR model. Figure 2-6d compares the wait time between the MIVR and VR models. The average wait times are 65.5 and 68.5 seconds for each model, respectively. This occurs because the MIVR model reduces the number of customers with longer wait times. The fraction of unsatisfied requests for both models is less than 0.1%. Under the base case scenario, the MIVR model reduces customer wait time by 4.4% on average and total non-occupied VMT by 8.5%.

To better understand the model performance relative to the magnitude of demand and the level of prediction accuracy, we compared the MIVR model with the VR model over the four demand scenarios described in Section 2.4.3. Figure 2-7 displays the non-occupied vehicle travel distance distributions and Figure 2-8 shows the customer wait time distribution over the four demand scenarios. For the low demand with accurate estimation (I) and demand underestimation (III) scenarios, the MIVR model outperforms the VR model by significantly reducing customer wait time while also reducing the average vehicle non-occupied travel distance. In the high demand with accurate estimation scenario (II), the MIVR model reduces customer wait time by proactively rebalancing vehicles more frequently than the VR model. In the demand overestimation scenario (IV), the MIVR model is outperformed by the VR model as

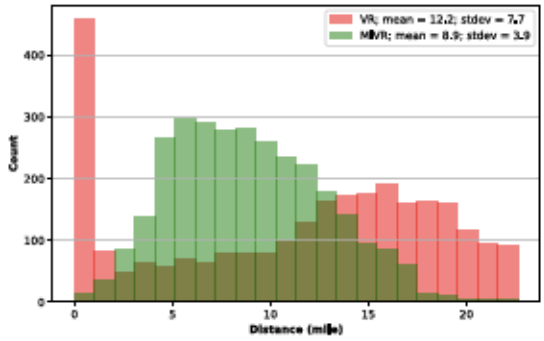
the VR model leads to lower average customer wait time and average vehicle non-occupied travel distance. The detailed simulation results for each demand scenario can be found in Appendix A.5.



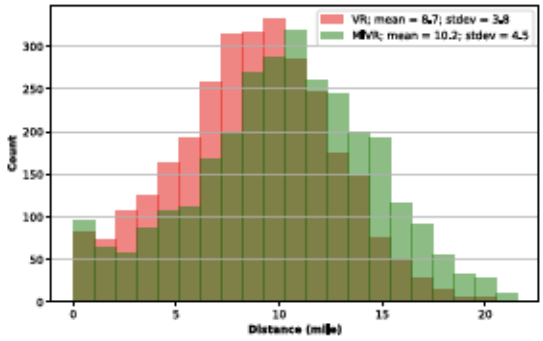
(a) Low demand with accurate estimation (0 - 6)



(b) High demand with accurate estimation (6 - 10)



(c) Demand underestimation (11 - 17)

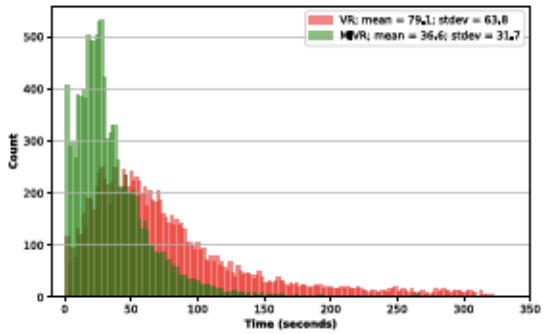


(d) Demand overestimation (20 - 24)

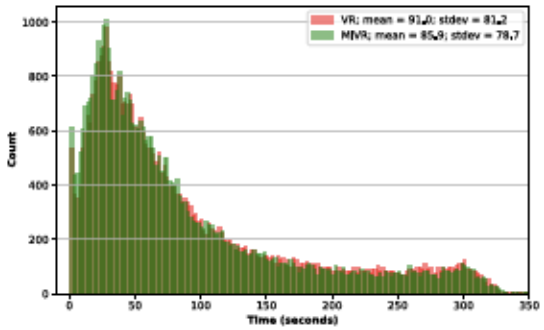
Figure 2-7: Vehicle non-occupied travel distance distributions for different demand scenarios.

To summarize, the MIVR model dispatches more vacant vehicles than the VR model when the level of estimated demand is high (given a specific fleet size N_v). On the other hand, fewer vehicles are dispatched by the MIVR model compared to the VR model when the level of estimated demand is low. This conclusion is further substantiated in Section 2.5.2, which discusses the results under different fleet sizes. We observe that the MIVR model is less proactive on dispatching vacant vehicles compared to the VR model when the fleet size is large relative to the level of demand.

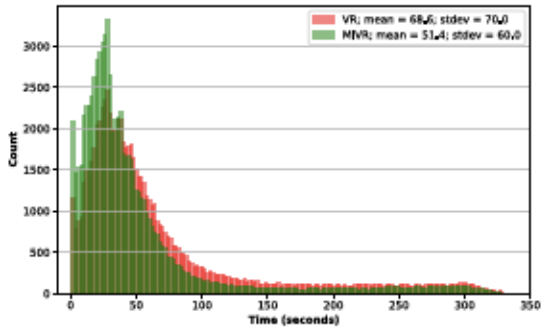
In this section, we have shown that the performance of rebalancing models, as measured by the average customer wait time, depends on the accuracy of demand



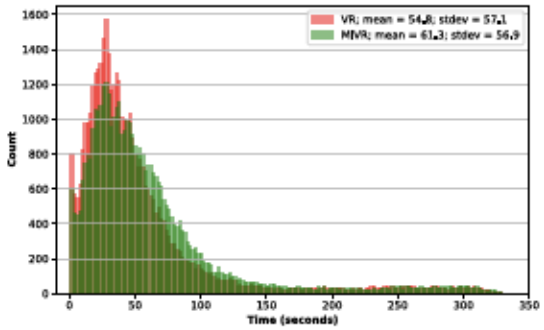
(a) Low demand with accurate estimation (0 - 6)



(b) High demand with accurate estimation (6 - 10)



(c) Demand underestimation (11 - 17)



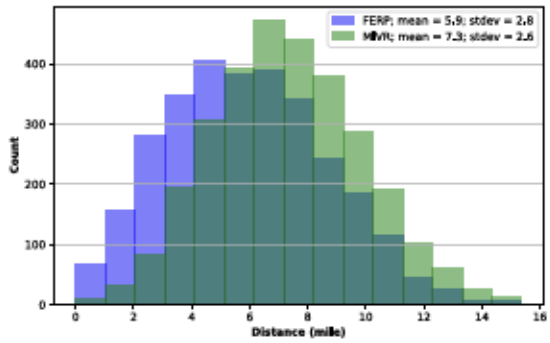
(d) Demand overestimation (20 - 24)

Figure 2-8: Customer wait time distributions for different demand scenarios.

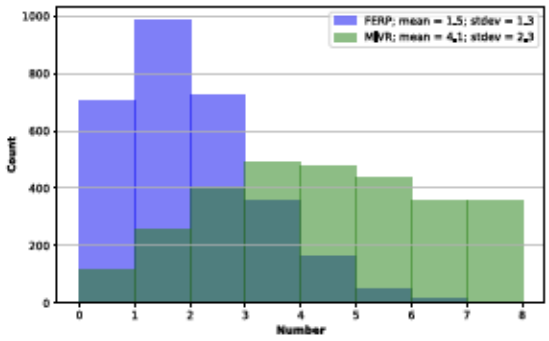
prediction and the level of demand. When the error in demand prediction is low, the MIVR model reduces the average customer wait time compared to the VR model. Model performance is penalized when the error in demand prediction is high (the total demand is underestimated or overestimated). Additionally, a rebalancing model which dispatches more vacant vehicles suffers higher penalties due to inaccurate demand estimation. In the demand scenario III, the level of predicted demand is low and the MIVR model dispatches fewer vacant vehicles than the VR model. Therefore, the MIVR model performs better than the VR model by reacting less often to inaccurate demand estimation. In the demand scenario IV, the level of predicted demand is high and the MIVR model dispatches more vacant vehicles than the VR model. The MIVR model experiences a higher penalty due to inaccurate demand estimations because of a proactive rebalancing strategy; hence, it performs worse than the VR model under these conditions. The demand scenario IV implies that the demand prediction serves

a critical role in the performance MIVR model. These results therefore demonstrate the value of a *robust* MIVR model that explicitly considers demand uncertainty.

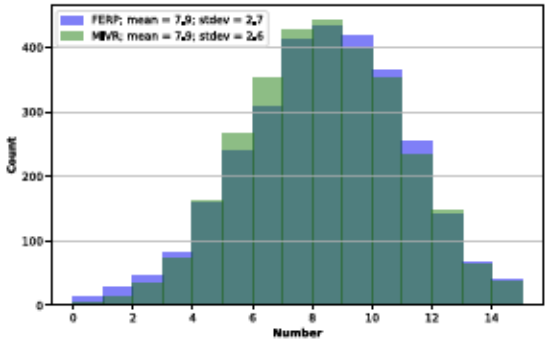
Benchmark FERP Comparison



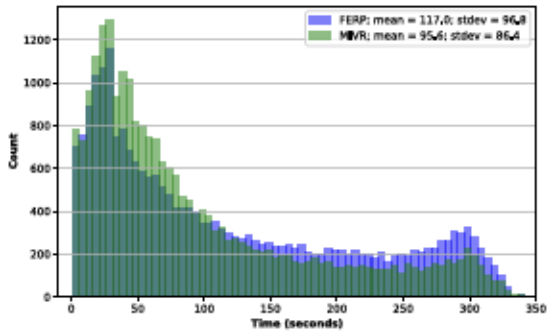
(a) Vehicle non-occupied travel distance distribution



(b) Vehicle rebalancing trip distribution



(c) Number of customers served (per vehicle) distribution



(d) Customer wait time distribution

Figure 2-9: Benchmark comparison results between MIVR and FERP models.

To further evaluate the performance of proposed MIVR model, we compared our approach with a state-of-the-art method for solving the vehicle rebalancing problem [30]. Braverman et al. [30] formulated a fluid-based optimization problem to generate a static empty-car routing policy. To guarantee a fair comparison, we chose a two-hour time period (7AM - 9AM) with historical demand and travel time data from June 2019 and 3000 vehicles to compute a static empty-car routing policy. We implemented the static routing policy in the simulator to dispatch vacant vehicles at each time interval instead of solving an optimization problem. Comparison results

are shown in Figure 2-9.

Figure 2-9a displays the distributions of non-occupied vehicle travel distance and Figure 2-9b shows the vehicle rebalancing trip distributions. The MIVR model dispatches vacant vehicles more proactively than the FERP. The distributions of number of customers served per vehicle are presented in Figure 2-9c, where vehicles are utilized slightly more evenly by the MIVR model than the FERP. Figure 2-9d displays the customer wait time distributions. The MIVR model reduces the average customer wait time by 18% while increasing total non-occupied VMT by 24%. The proportion of unsatisfied requests for both approaches is less than 0.1%, which is a result of the adequate supply of vehicles. The MIVR model optimizes rebalancing decisions during each time interval and the FERP maintains the same vehicle rebalancing policy throughout the simulation period. In general, the MIVR model provides better service quality for customers by producing a more proactive rebalancing strategy, but it also results in a somewhat higher non-occupied VMT.

2.5.2 Scenario Testing

Second, we test the sensitivity of the results when changing input parameters of the MIVR model, including the fleet size, N_v, the length of decision time interval, κ, the weight parameter for pickup distance, β, and the size of the sub-regions. To avoid the effect of inaccurate demand estimation when testing different scenarios[6], we tested different scenarios with N_v, κ and β over a four-hour time period (6 AM - 10 AM) assuming perfect future demand predictions. Alternative scenarios are generated by changing the simulation parameters for the base case.

Fleet size N_v

Results for scenarios with varying fleet sizes, represented by N_v in the simulation parameters, are shown in Figure 6-10. When there is a limited number of vehicles

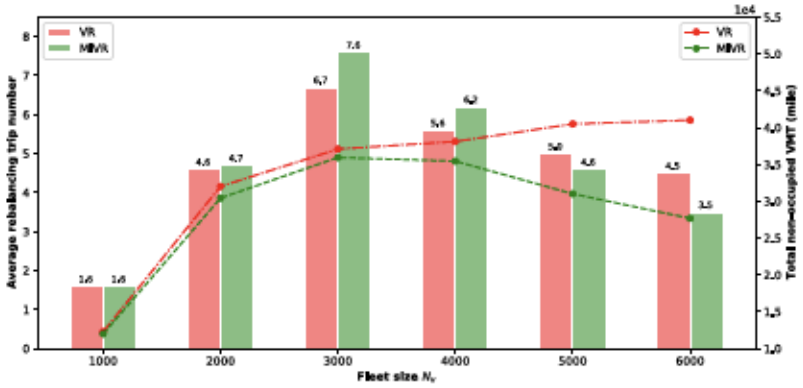
[6]The effect of input parameters on the simulation results can be overshadowed by the effect induced by inaccurate demand estimations when two have contradictory effects on certain performance metrics.

($N_v \leq 4000$) in the system, the MIVR model generates more rebalancing trips per vehicle compared to the VR model. When there are sufficient vehicles in the system ($N_v = 5000$ or 6000), the MIVR model dispatches fewer vacant vehicles and reduces the total non-occupied VMT compared to the VR model. This is intuitive; for the MIVR model, less rebalancing is needed when there is a higher concentration of idle vehicles since more passengers can be picked up (within the maximum wait time constraint) without significant rebalancing. Therefore, the MIVR model reduces the total non-occupied VMT. The MIVR model decreases the average customer wait time under all scenarios with different fleet sizes compared to the VR model. Customer wait time decreases significantly for the MIVR model when a larger fleet is available. Even though rebalancing is not as critical for a large fleet, the MIVR model continues to minimize pickup distance and therefore customer wait time. The proportion of unsatisfied requests is marginally decreased for the MIVR model compared to the VR model, regardless of fleet size.

The scenario testing with different fleet sizes implies the existence of the *Pareto* improvement at aggregate level for the MIVR model compared to the VR model. When a sufficient number of vehicles is available, the MIVR model reduces the total non-occupied VMT, average vehicle rebalancing trips and average customer wait time while satisfying more requests compared to the VR model. For instance, when there are 6000 vehicles in the system (with $\kappa = 6$), the MIVR model reduces the total non-occupied VMT by 33%, average vehicle rebalancing trips by 22% and average customer wait time by 36% when compared to the VR model. Under this scenario, the MIVR model clearly outperforms the VR model, indicating that the *Pareto* improvement exists.

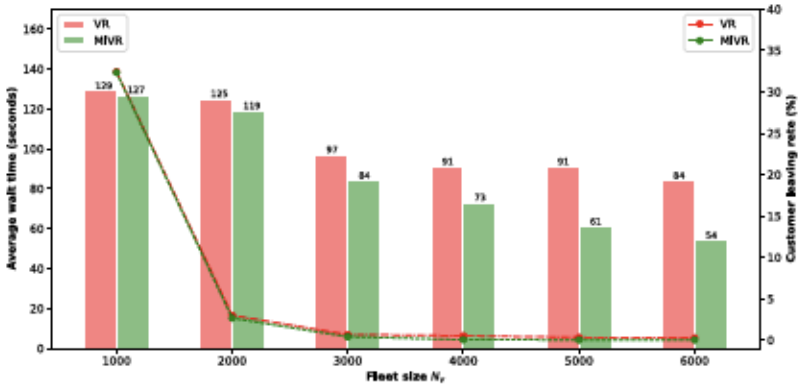
Decision time interval length κ

Figure 2-11 shows the results under scenarios with varying decision time intervals κ. Both models dispatch more vehicles when considering additional future time intervals (i.e. κ becomes large), and similar amount of vacant vehicles are dispatched by both models. Also, the total non-occupied VMT increases when considering more future



(a) Average number of rebalancing trips per vehicle and total non-occupied VMT[a].

[a]Bars indicate the average rebalancing trip number per vehicle and dashed lines show the total non-occupied VMT.



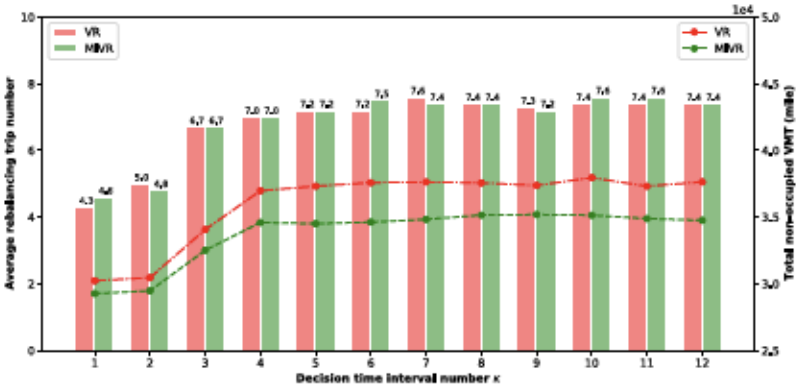
(b) Average wait time per customer and proportion of unsatisfied requests[a].

[a]Bars indicate the average wait time per customer and dashed lines show the proportion of unsatisfied requests.

Figure 2-10: Scenario testing results for different fleet size N_v.

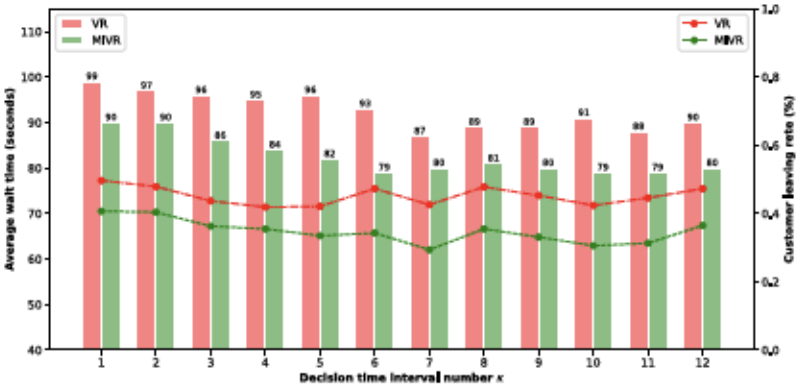
time intervals for both models, and the MIVR model leads to less non-occupied VMT compared to the VR model for all scenarios. With respect to customer wait time, considering additional time intervals benefits both models and the MIVR model reduces wait times for all scenarios compared to the VR model. The MIVR outperforms the VR model on the proportion of unsatisfied requests for all scenarios.

Note that selecting number of time intervals presents a trade-off between system performance and computation time. Increasing κ linearly increases the size of the problem, which may result in a solution time that is too long to use in practice. The



(a) Average number of rebalancing trips per vehicle and total non-occupied VMT[a].

[a]Bars indicate the average rebalancing trip number per vehicle and dashed lines show the total non-occupied VMT.



(b) Average wait time per customer and proportion of unsatisfied requests[a].

[a]Bars indicate the average wait time per customer and dashed lines show the proportion of unsatisfied requests.

Figure 2-11: Scenario testing results for different decision time interval length κ.

average computation time for solving the MIVR model with $\kappa = 6$ is 3.8 seconds and the average computation time for the MIVR model with $\kappa = 12$ is 7.5 seconds. Platform operators must therefore choose a look-ahead window that is suited to their system size and computational capacity.

Weight parameter for pickup distance β

The weight parameter β in the MIVR model controls the trade-off between the total non-occupied VMT and the service quality. In previous experiments, $\beta = 1$ was

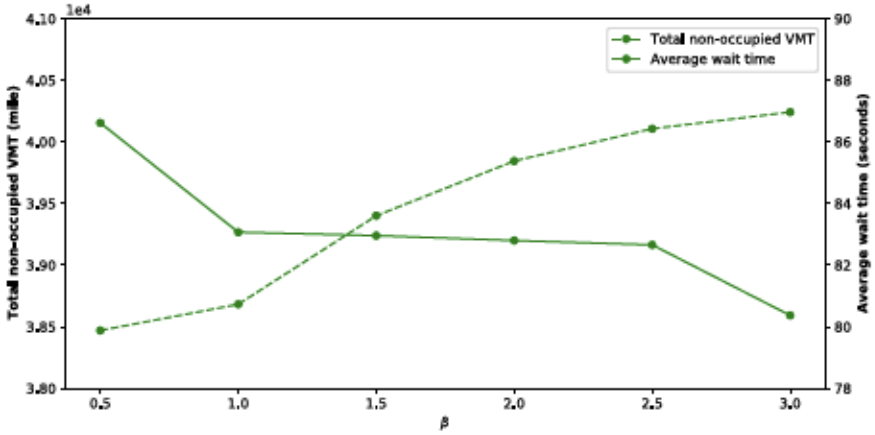


Figure 2-12: Sensitivity testing results for the weight parameter β in the MIVR model.

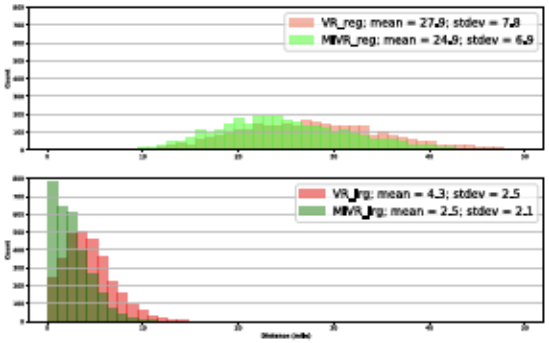
used as a base case, leading to a MIVR model which purely minimized the total non-occupied VMT and the number of unsatisfied requests. In this section, different values of β are tested based on the base case simulation setting assuming perfect future demand predictions, and the total non-occupied VMT and the average customer wait time are shown in Figure 2-12. Solid line indicates the average customer wait time and dashed line represents the total non-occupied VMT.

When β becomes larger, the MIVR model puts more weight on the service quality (customer wait times), and the total non-occupied VMT gets larger. The average customer wait time monotonically decreases when β increases. By increasing the value of β to 3, the average wait time is reduced by 3% while increasing the total non-occupied VMT by 4%. However, the MIVR model becomes more vulnerable to the demand uncertainty when the value of β is large. This is because more vacant vehicles are rebalanced when β is large, where a larger penalty is induced by the inaccurate demand estimations. Therefore, the service quality can be diminished if β is too large.

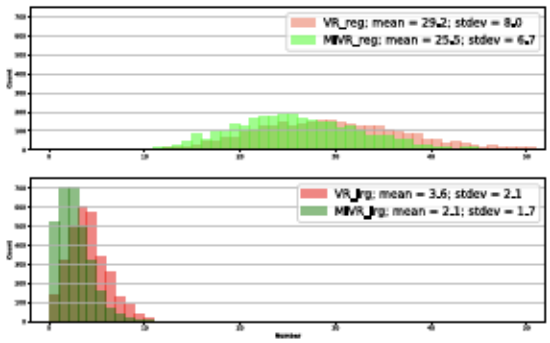
On the other hand, a negative weight is put on the service quality when $\beta < 1$, meaning that the service quality is sacrificed to reduce the total non-occupied VMT. For the scenario with $\beta = 0.5$, the total non-occupied VMT is reduced by 0.5% and the average wait time is increased by 4% compared to the base case. Since the vehicle

rebalancing distance is highly correlated with customer wait time, reducing β does not significantly decrease the total non-occupied VMT.

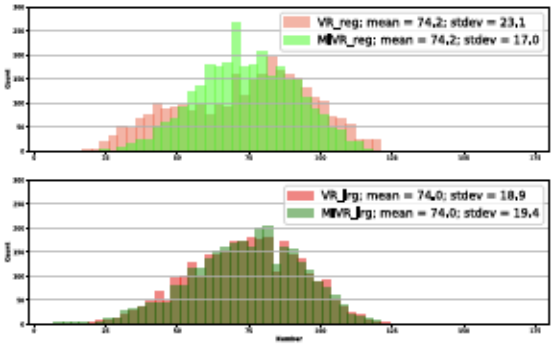
Sub-regional size



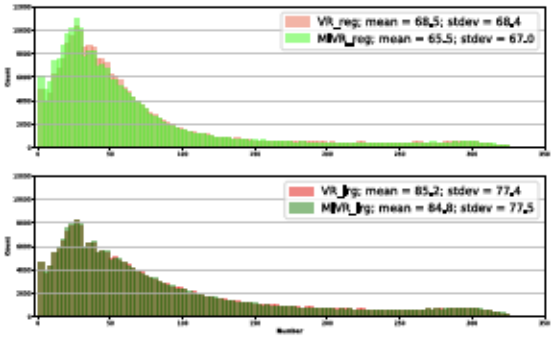
(a) Vehicle non-occupied travel distance distribution



(b) Vehicle rebalancing trip distribution



(c) Number of customers served (per vehicle) distribution



(d) Customer wait time distribution

Figure 2-13: Results comparison between simulations with 63 regular sub-regions and 13 large sub-regions.

The MIVR model performance relies on the size of sub-regions. Smaller sub-regions leads to more rebalancing options (decision variables) and a better overall model performance. However, the model complexity increases when considering smaller sub-regions. To quantify the effect of changing the size of sub-regions, we combined 63 taxi zones into 13 larger zones and ran simulations for the 13 large sub-regions. Comparison results are shown in Figure 2-13.

Figure 2-13a and 2-13b show the distributions of non-occupied vehicle travel dis-

tance and vehicle rebalancing trips. Fewer sub-regions with larger size reduces the opportunities for rebalancing vacant vehicles between sub-regions. Therefore, both the average vehicle non-occupied VMT and rebalancing trips are significantly decreased. The distribution of number of customers served per vehicle is shown in Figure 2-13c, where vehicles are more evenly utilized by the MIVR model under a smaller sub-region size. Figure 2-13d displays the customer wait time distributions for both scenarios. Compared to the scenario with larger sub-regions, the scenario with 63 sub-regions leads to 20% and 23% reductions on the average customer wait time for the VR and the MIVR, respectively. Differences between the MIVR and VR models hold regardless of the size of the sub-regions.

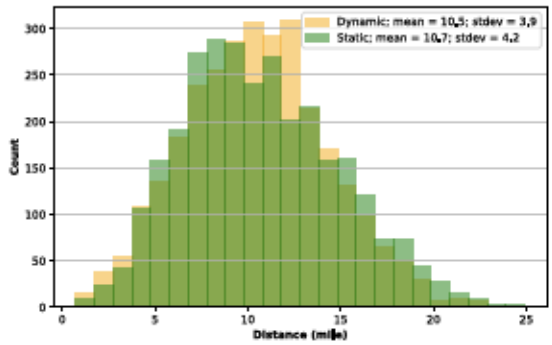
As for the computation complexity, the average running time for producing rebalancing decisions during each iteration by the MIVR model under a regular sub-region size is 3.95 seconds. The average running time for the MIVR model under a larger sub-region size is 0.18 seconds. Reducing the number of sub-regions from 63 to 13 saves approximately 95% of the computation time on generating rebalancing decisions. In general, the size of sub-regions should be chosen to balance computation complexity and model performance.

2.5.3 Impact of Regional Transition Matrices

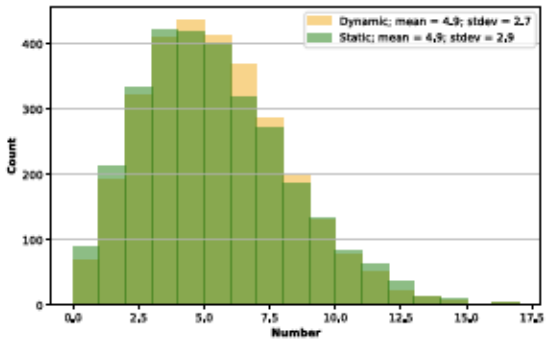
In the MIVR model, we utilized static regional transition matrices P and Q, which are estimated from the historical data, to reflect the movement of occupied vehicles. However, the true regional transition matrices depend spatio-temporal demand flows and operators' dispatching and rebalancing strategies. In this section, we will quantify the impact of approximating true regional transition matrices with the historical data.

To incorporate the true regional transition matrices in the model, we modified the simulator by estimating regional transition matrices for occupied vehicles based on preceding matching decisions at the beginning of each simulation period. By using the previous matching decisions in the simulation, only regional transition matrices between the current time period k and the next time period $k+1$ can be evaluated accurately. Therefore, we implemented a MIVR model with $\kappa = 2$ in the simulation,

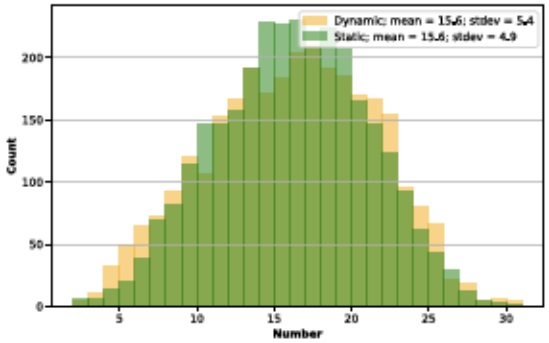
indicating that two time intervals were considered when making rebalancing decisions. Other simulation parameters are identical to the base case scenario. Such a modified simulator is able to produce rebalancing decisions based on the true regional transition matrices at each time interval.



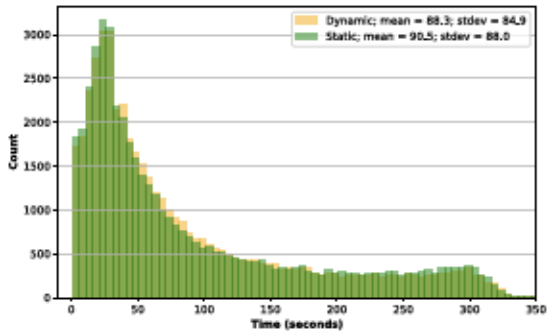
(a) Vehicle non-occupied travel distance distribution



(b) Vehicle rebalancing trip distribution



(c) Number of customers served (per vehicle) distribution



(d) Customer wait time distribution

Figure 2-14: Comparison results between simulators with dynamic and static regional transition matrices.

To quantify the impact of approximating regional transition matrices with the historical data, we compared results from the modified simulator to results from a standard simulator described in section 2.4.1 with $\kappa = 2$, which guarantees identical look-ahead windows in the MIVR model. Results are compared within a four-hour time period (8AM - 12PM) and detailed comparison results are shown in Figure 2-14. *Dynamic* indicates that regional transition matrices are estimated at the beginning of every simulation time interval. *Static* implies that regional transition matrices

estimated by the historical data are utilized.

Figure 2-14a shows the distributions of vehicle non-occupied travel distance. Utilizing true regional demand matrices reduces the total non-occupied VMT by 1.9%. Distributions of vehicle rebalancing trips and number of customers served are displayed in Figure 2-14b and 2-14c, where two simulators have identical performance on average. Figure 2-14d presents the distributions of customer wait time. Using true regional transition matrices reduces the average customer wait time by by 2.4%.

The comparison results imply that approximating the true regional transition matrices with static matrices estimated from the historical data has a marginal impact on model performance. This is intuitive; the regional transition matrices are used for constructing a forward-looking vehicle rebalancing model. In the simulation, only the rebalancing decisions for the first time interval will be implemented, although rebalancing decisions for κ time periods are generated. When moving to the next time period, real-time information (e.g., vehicle locations) is updated and a separate MIVR model considering κ time intervals is solved. Therefore, regional transition matrices have a limited impact on rebalancing decisions at the first time interval, which subsequently has a marginal impact on model performance.

2.5.4 Robust Model Results

To evaluate the robust optimization model, we tested multiple scenarios with different levels of uncertainty as defined by the uncertainty set size parameters ρ and Γ. Each robust solution was generated for the robust MIVR model considering 6 future time intervals, i.e., $\kappa = 6$. The model parameters were set as $\beta = 1$, $\gamma = 10^2$ and $\bar{w} = \tilde{w} = 300$. For the number of vehicles N_v, we considered the scenario with 3000 vehicles, indicating a sufficient supply (almost all customers can be served) given the demand profile, and 2000 vehicles, representing an insufficient supply. The initial vehicle distributions V_1 and O_1 are generated using the following process: each vehicle in the fleet with size N_v is either vacant or occupied with equal probability and is randomly assigned to a sub-region. To test the performance for different solutions, we utilized the real demand data from 9 AM - 9:30 AM for 65 work days from April

to June 2019 and solved a driver-customer matching problem with realized demand and vehicle distributions after rebalancing. The performance of each solution was evaluated based on the average values of the total pickup time and the number of unsatisfied requests over the 65 demand scenarios. The solution generated by the nominal MIVR model was used as the benchmark for evaluating robust solutions. The performance of each robust solution is displayed as the percentage reduction in performance measurements compared to the nominal solution.

ρ \ Γ	0	1	2	3	4	5	6	7	8	9	10
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.1	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
0.2	0.51	1.38	1.18	0.66	1.18	0.66	1.18	1.18	0.66	0.79	1.38
0.3	2.45	2.45	4.52	2.45	2.45	2.45	2.45	4.52	4.52	2.45	4.52
0.4	3.47	4.15	4.15	4.15	4.15	4.15	5.67	4.15	4.15	5.67	5.6
0.5	5.62	5.62	5.62	5.62	5.62	5.62	5.62	5.62	7.68	7.68	5.62
0.6	7.89	7.89	7.89	7.89	7.89	7.89	7.89	7.89	7.89	7.89	7.89
0.7	8.78	10.32	10.32	10.32	10.32	10.32	10.32	10.32	10.32	10.32	10.32
0.8	13.24	13.24	13.24	13.24	13.24	13.24	13.24	13.24	13.24	13.24	13.24
0.9	17.17	17.17	17.17	18.59	17.17	17.17	17.17	19.78	18.59	18.59	17.17
1.0	21.19	19.92	21.23	21.23	21.23	21.23	21.23	21.23	21.23	21.23	21.23

Table 2.2: Percentage reduction in the total pickup time compared to the nominal MIVR solution with insufficient supply ($N_v = 2000$), for different values of ρ and Γ.

For the scenario with insufficient supply ($N_v = 2000$ and a proportion of customers can not be served), Table 2.2 shows the results about the total pickup time[7] and and Table 2.3 displays the percentage reduction for the number of unsatisfied requests[8] over the nominal MIVR model. Introducing uncertainty into the model generates solutions that outperform the nominal solution for all values of ρ and Γ. The uncertain parameter ρ significantly affects the total pickup time and the number of unsatisfied requests while the uncertain parameter Γ has limited impact on them. When a high level of uncertainty is considered in development of the robust MIVR model, more customers can be served with less total pickup time.

[7]Gray cells indicate uncertain scenarios with the largest reduction in pickup time.
[8]Gray cells indicate uncertain scenarios with the largest reduction in unsatisfied requests.

ρ \ Γ	0	1	2	3	4	5	6	7	8	9	10
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
0.2	0.2	0.41	0.21	0.2	0.21	0.2	0.21	0.21	0.2	0.2	0.41
0.3	0.17	0.17	0.61	0.17	0.17	0.17	0.17	0.61	0.61	0.17	0.61
0.4	0.14	0.14	0.14	0.14	0.14	0.14	0.22	0.14	0.14	0.22	0.3
0.5	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.3	0.3	0.08
0.6	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
0.7	0.2	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22
0.8	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46
0.9	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
1.0	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54

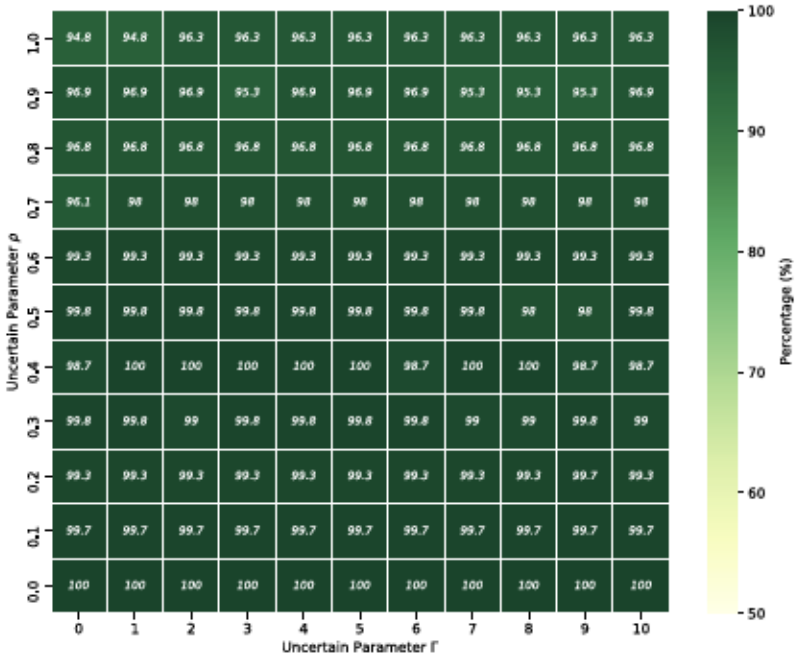
Table 2.3: Percentage reduction in the number of unsatisfied requests compared to the nominal MIVR solution with insufficient supply ($N_v = 2000$).

ρ \ Γ	0	1	2	3	4	5	6	7	8	9	10
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.1	4.2	4.2	4.2	4.2	4.2	4.2	4.2	4.2	4.2	5.23	4.19
0.2	6.4	6.4	6.4	7.44	7.01	7.27	6.67	6.45	6.4	6.4	6.4
0.3	12.51	12.51	12.51	12.0	12.51	12.0	12.0	12.0	12.0	12.51	12.51
0.4	16.23	15.33	15.33	15.33	15.33	15.33	15.33	15.33	15.33	15.33	15.33
0.5	18.19	18.32	18.32	17.57	18.32	18.32	18.32	18.32	18.32	18.32	18.32
0.6	24.14	22.96	22.96	22.98	22.98	22.96	22.98	22.96	22.96	22.96	22.98
0.7	25.62	25.18	25.18	25.18	25.18	25.18	25.18	25.18	25.18	25.18	25.18
0.8	30.89	29.39	29.22	29.13	31.44	29.39	30.98	29.82	29.4	29.73	29.39
0.9	39.82	36.55	38.02	38.92	36.6	36.44	37.16	36.44	38.1	36.44	36.44
1.0	38.22	39.1	39.49	39.01	39.01	40.41	39.49	39.41	41.03	40.47	40.93

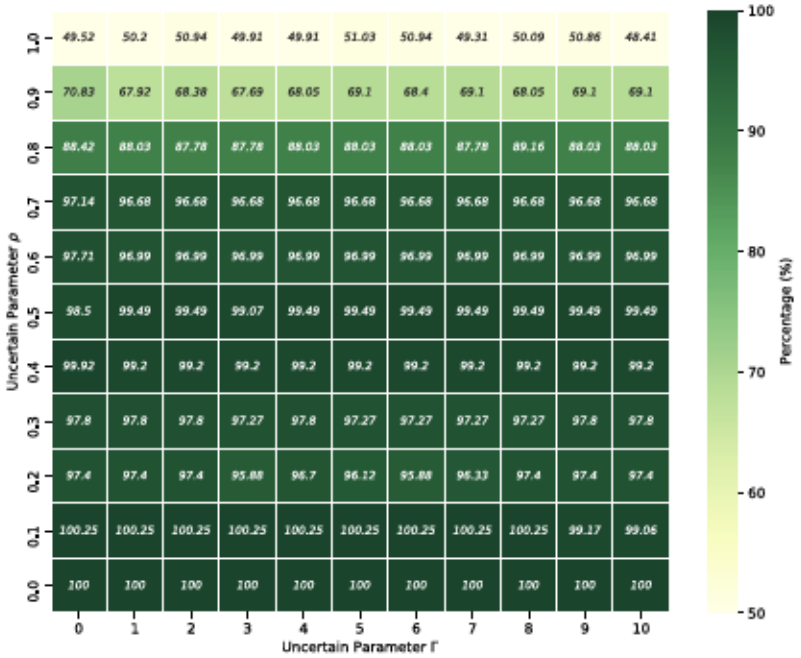
Table 2.4: Percentage reduction in the total pickup time compared to the nominal MIVR solution with sufficient supply ($N_v = 3000$), for different values of ρ and Γ.

For the scenario with sufficient number of vehicles ($N_v = 3000$ and almost all customers can be served), the percentage reduction of the total pickup time is shown in Table 2.4. The robust MIVR model benefits more when having a large fleet of vehicles in the system. The largest total pickup time reduction for the robust MIVR model with sufficient supply is 41.03% compared to 21.23% for the scenario with insufficient supply. Under the scenario with sufficient supply, all customers can be served and introducing uncertainty into the model generates solutions that outperform

the nominal solution for all values of ρ and Γ.



(a) Insufficient supply scenario with fleet size $N_v = 2000$ (the nominal MIVR model conducts 864 vehicle rebalancing trips)



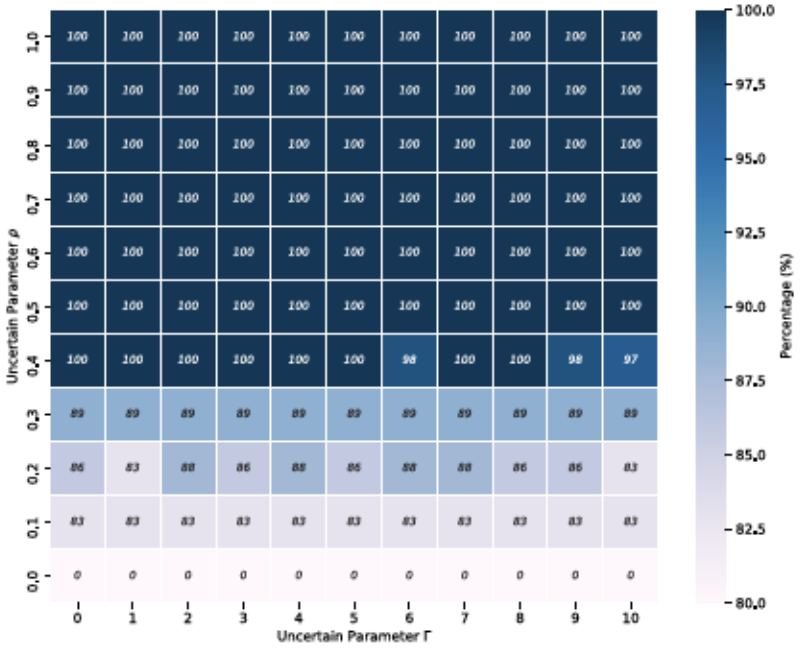
(b) Sufficient supply scenario with fleet size $N_v = 3000$ (the nominal MIVR model conducts 1228 vehicle rebalancing trips)

Figure 2-15: Rebalancing trips for the robust MIVR model under multiple uncertain scenarios.

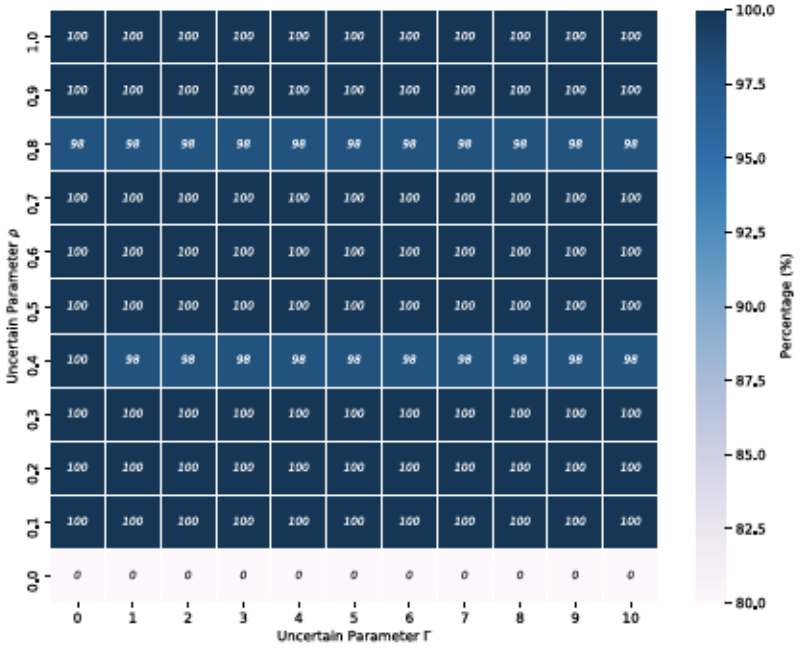
The robust MIVR model protects the rebalancing decisions against demand uncertainty by restricting the number of rebalancing trips compared to the nominal MIVR model, which is shown in Figure 2-15. Each cell represents the percentage of rebalancing trips under a specific level of uncertainty compared to the number rebalancing trips in the nominal MIVR model. When dispatching fewer vacant vehicles compared to the nominal case, the penalty incurred due to inaccurate demand estimations is decreased and the system becomes more robust against the demand uncertainty, hence has less total pickup time. The number of rebalancing trips is significantly restricted (less than 50% compared to the nominal MIVR model) when introducing a high level of uncertainty into the robust MIVR model under the sufficient supply scenario.

Figure 2-16 shows the daily performance of the robust MIVR model compared to the nominal MIVR model. Each cell represents the percentage of the 65 input days that the robust MIVR model performs strictly better than the nominal MIVR model under a given level of uncertainty. Under the insufficient supply scenario, even considering a low level of uncertainty ($\rho = 0.1$) can significantly improve the performance of the robust MIVR model (better performance than the nominal MIVR model for 83% of the 65 days tested). When incorporating a moderate level of uncertainty ($\rho \geq 0.5$) into the model, the robust MIVR model outperforms the nominal MIVR model for every day of demand tested. When a sufficient supply of vehicles is available, the robust MIVR model performs better than the nominal MIVR model for every weekday tested over most uncertain scenarios.

Overall, the robust MIVR model generates rebalancing decisions based on out-of-sample demand uncertainty defined by parameters ρ and Γ, and solutions are evaluated with real demand data reflecting in-sample demand uncertainty. The parameters ρ and Γ for uncertainty sets indicate the level of demand uncertainty that ride-hailing operators are willing to protect rebalancing decisions against. Based on experiment results, introducing robustness into the MIVR model and protecting rebalancing decisions against demand uncertainty improve the system performance effectively under insufficient and sufficient supply cases. The robust MIVR model performs even better when having sufficient number of vehicles in the system.



(a) Insufficient supply scenario with fleet size $N_v = 2000$



(b) Sufficient supply scenario with fleet size $N_v = 3000$

Figure 2-16: Daily robust MIVR model performance under multiple uncertain scenarios.

2.6 Conclusions and Future Work

In this chapter, we formulate the MIVR model, which incorporates the driver-customer matching component into the consideration of vehicle rebalancing decisions made by ride-hailing operators, to protect rebalancing decisions against future demand uncertainty induced by inaccurate demand estimates. We evaluate the performance of our model by comparing against a benchmark VR model and a state-of-the-art model, named fluid-based empty-car routing policy (FERP), using actual ride-hailing trip data. Comparing to the VR model, the MIVR model reduces the average customer wait time and the total non-occupied VMT under most scenarios. When a large fleet is available, a *Pareto* improvement can be found regarding the overall non-occupied VMT, the average vehicle rebalancing trips, the average customer wait time and the number of unsatisfied requests. Comparing to the FERP, the MIVR model reduces the average customer wait time by generating a more proactive rebalancing strategy. To further immunize solutions against demand uncertainty, we propose the robust MIVR model by introducing RO techniques. The robust MIVR is especially effective when the supply of ride-hailing vehicles is sufficient and most requests can be satisfied. Under both sufficient-supply and insufficient-supply cases, the robust MIVR model prevents rebalancing decisions from inaccurate demand estimation by rebalancing fewer vehicles. Additionally, introducing robustness into the MIVR model generates rebalancing decisions that performs better than decisions produced by the nominal MIVR model under most demand scenarios.

The main limitations of this study are a result of approximations embedded in the MIVR model. First, we are only able to model trips aggregated to the zonal level given the data availability. While we simulate actual pickups and drop-off locations within those zones, future work could incorporate disaggregate data to test rebalancing and matching at the individual address level. The model could be improved if these data were made available. We also assume static regional transition matrices estimated from the historical data. Though having limited impacts on model performances, matching and rebalancing decisions-based regional transition matrices can

be considered in the model to better reflect vehicle trajectories across multiple time periods.

This chapter shows how internalization of matching costs can be used to protect rebalancing decisions against demand uncertainty and improve the efficiency of ride-hailing operations regarding customers (satisfy more customers with shorter wait times), and under what conditions the proposed method is beneficial. Furthermore, it illustrates how robust optimization complements the MIVR model by further limiting the risk of increased cost due to incorrect demand estimations. Ride-hailing service operators should consider adopting the robust MIVR model for improved customer outcomes, such as wait time and unsatisfied requests, and reduced costs for operators.

There are several future research directions we identified in this chapter. First, the uncertainty set $\bar{\mathcal{U}}^k(\Gamma)$ has a limited impact on system performance. More effective and interpretable uncertainty sets could be designed to model the uncertainty in the ride-hailing system. Secondly, additional uncertainty variables could be considered besides the demand uncertainty, such as travel time. Thirdly, we used the historical average as the future demand estimates in this chapter. Advanced demand prediction algorithms can be incorporated within the robust MIVR model to further improve operational performances. Lastly, the MIVR model could be extended to solve the vehicle rebalancing problem in the shared MoD system.

Chapter 3

Data-driven Vehicle Rebalancing with Predictive Prescriptions in the Ride-Hailing System

3.1 Introduction

Ride-hailing platforms are one of the most essential components of the emerging Mobility-on-Demand (MoD) system, which provides passengers with improved mobility options through a traveler-centric multimodal urban transportation system [140]. With the rapid growth of ride-hailing platforms, such as Uber, Lyft, and DiDi, ride-hailing and ride-sharing services have become increasingly popular all over the world, especially in highly-urbanized regions. In New York City (NYC), ride-hailing platforms transported on average 15 million passengers per month in 2016, which was approximately the same number of trips served by NYC's 43,000 yellow cabs [139]. A recent survey indicates that 36% of American adults have used a ride-hailing platform (Uber or Lyft) in 2018, an increase from 15% in late 2015 [86].

However, ride-hailing platforms face significant challenges with respect to operational efficiency. Despite having algorithmic pricing and matching strategies currently in place, drivers from ride-hailing platforms spend an estimated 40% of the time cruis-

ing without passengers in major cities[32]. With technological advances in the field of autonomous driving in the past decade, Autonomous Mobility-on-Demand (AMoD) systems are becoming a reality. With a fleet of autonomous vehicles (AVs), centralized control and planning of vehicles become more vital to efficient operations [188].

One of the major operational decisions critical to the efficient operations of ride-hailing systems is vehicle rebalancing, where vacant vehicles are redistributed proactively to areas with anticipated high demand to reduce the discrepancy between spatial distributions of supply and demand during each time period, therefore reducing customer wait times [144, 164, 121, 172, 67].

Since the future demand in ride-hailing systems is unknown, performances of rebalancing decisions rely on both the prediction accuracy of future demand and the uncertainty considerations in subsequent optimization. Various machine learning approaches have been developed to produce a point prediction of future demand with high accuracy [184, 92, 90, 56, 185]. Subsequently, the decisions are made according to either the nominal predicted demand, which is named point-prediction-driven optimization, or an uncertainty set around the prediction, which is termed robust optimization. Robust optimization has been used widely for decision-making under uncertainty and has been applied to the vehicle rebalancing problem in [67]. However, a good demand prediction does not necessarily lead to a good rebalancing decision. In the demand prediction, all errors are considered the same, whereas in the rebalancing problem sending additional vehicles to remote regions due to overestimated future demand would incur a larger cost compared to if the additional vehicles were to be sent to more connected and central regions.

On the other hand, data-driven optimization directly prescribes decisions from data. For example, stochastic optimization has been commonly used for handling problems that require making decisions under uncertainty in Operations Research (OR) [28]. However, standard data-driven optimization approaches, such as Sample Average Approximation (SAA), do not utilize auxiliary information, which leads to an unacceptable waste of good data. To combine ideas from ML and OR while making use of all available observations and information, a data-driven predictive pre-

scriptions framework was proposed to prescribe optimal decisions in decision making under uncertainty [20].

In this chapter, a novel data-driven optimization approach, predictive prescription, is introduced into vehicle rebalancing problems to generate better rebalancing decisions against demand uncertainty for ride-hailing platforms. The predictive prescriptions are compared with the standard point-prediction-driven optimization framework, stochastic optimization methods, and robust optimization methods. The contributions of this chapter can be summarized as follows:

- Introducing the predictive prescription framework into solving the vehicle rebalancing problem in ride-hailing operations.

- Applying the graph convolutional Long Short-Term Memory (LSTM) and the station-based LSTM into predicting the future demand of ride-hailing systems. Simulations results indicate that prediction errors caused by demand underestimation in predictive models can benefit system performances.

- Using real-world simulations to compare model performances of predictive prescription models with point-prediction-driven optimization models under four different demand scenarios. When demand prediction accuracy is low, predictive prescriptions outperform point-prediction-driven optimization in terms of reducing average customer wait times. The edge of data-driven optimization over point-prediction-driven optimization increases when the supply to demand ratio increases. When demand can be predicted accurately, point-prediction-driven optimization is a better approach to adopt.

- Comparing predictive prescriptions with the robust matching-integrated vehicle rebalancing (MIVR) model proposed in [67]. Compared to the robust MIVR model, predictive prescriptions achieve competitive performances without relying on any additional information about the future demand.

The remainder of the chapter is structured as follows. Section 3.2 reviews the relevant literature in vehicle rebalancing operations, predictive models and data-driven

optimization approaches. Section 3.3 describes the basic MIVR model and approaches for improving model performances regarding demand uncertainty including predictive methods and data-driven optimization approaches. Data used in this chapter is discussed in Section 3.4. Real-world simulation settings and empirical results are shown in Section 3.5, including performance comparisons between point-prediction-driven optimization models, predictive prescription models, and robust models. Finally, Section 3.6 recaps the main contributions of this work and provides future research directions.

3.2 Literature Review

3.2.1 Vehicle rebalancing

Rebalancing vacant vehicles is a critical operational strategy for ride-hailing platforms in addition to matching customers with drivers [165]. Due to the spatial imbalance of demand and supply in ride-hailing systems, relocating idle vehicles to areas where estimated future demand exceeds vehicle supply could reduce empty miles traveled and customer wait times. An online vehicle rebalancing algorithm developed in [164] led to a 37% reduction in the average customer wait times compared to the scenario where no rebalancing took place.

The vehicle rebalancing problem is first studied in [59], where an adaptive dynamic programming algorithm is proposed for dynamic fleet management with single-period and multi-period travel times.

Since then, various approaches have been proposed to solve the vehicle rebalancing problem in ride-hailing systems. Typical vehicle rebalancing problems discretize the operating region into sub-regions and vacant vehicles are rebalanced between zones by solving a mathematical programming problem. Wen et al. [172] utilized a reinforcement learning approach to address the vehicle rebalancing problem in a shared MoD system. Their proposed method reduced the fleet size by 14% in a real-world simulation in London. Jiao et al. [88] proposed a practical framework based on deep

reinforcement learning and decision-time planning for rebalancing vehicles in ride-hailing systems. Braverman et al. [30] designed a fluid-based optimization model to model vehicles in ride-hailing systems. Their proposed method resulted in a higher fraction of passengers served compared to benchmark models. Miao et al. [121] introduced a data-driven distributionally robust vehicle rebalancing model to minimize the worst-case vehicle rebalancing cost, which consists of vehicle rebalancing distance and a service quality function indicating the balanced-ness between supply and demand. Their approach was evaluated with real-world taxi data in NYC and achieved a 30% reduction in idle driving distance on average.

With the advent of autonomous vehicles, vehicle rebalancing problems have been studied extensively for AMoD systems as well in recent years [188]. A fluid model was utilized to model passengers and vehicles, and an optimal rebalancing policy was developed by solving a linear program [131]. A queueing-based theoretical model was also proposed to model the vehicle rebalancing problem in the AMoD system. The algorithm was designed to minimize the total number of rebalancing trips while maintaining vehicle availability [194]. Iglesias et al. [84] proposed a Model Predictive Control (MPC) algorithm to compute rebalancing strategies by leveraging short-term demand forecasts utilizing the LSTM neural networks. Their proposed algorithm significantly reduced the average customer wait time compared to the rebalancing strategy proposed in [131]. In a shared AMoD setting, Tsao et al. [155] proposed an MPC algorithm to optimize routes for both vacant and occupied vehicles.

Besides, decentralized vehicle rebalancing systems were proposed as contingency plans when AVs lost connections with central dispatch systems. Chen et al. [44] proposed a decentralized cooperative cruising method for offline operations of AMoD fleets. Their proposed method shows significant performance improvements compared to strategies with random-selected destinations for rebalancing AVs under different fleet sizes.

Most recently, Guo et al. [67] proposed a MIVR model, introducing driver-customer matching component into the vehicle rebalancing problem to produce better rebalancing decisions. Robust optimization was used to better protect rebalancing decisions

against demand uncertainty. Their method could reduce the average customer wait time by 18% compared to approaches proposed in [30] under a real-world simulation with the NYC ride-hailing data.

One common modeling framework to handle demand uncertainty is to predict and optimize in separate steps, in which a prediction model is built first, followed by an optimization model taking the outputs from the prediction model. Few studies have considered combining prediction and optimization into one framework. Al-kanj et al. [2] studied a sequence of decision problems in a ride-hailing system with autonomous electric vehicles, including vehicle dispatching (matching, rebalancing, EV charging), surge pricing, and fleet size problems. They utilized value functions to represent the spatial and temporal patterns of demand in order to incorporate the downstream impact of a decision made now on the future. The vehicle dispatching problem was modeled as a Markov decision process and addressed with the approximate dynamic programming (ADP) approach. Ramezani and Nourinejad [136] proposed a taxi dispatching model using the model predictive control approach. They incorporated the interrelated impact of normal traffic flows and taxi dynamics when generating dispatching decisions.

In summary, demand prediction and decision-making under uncertainty are two flourishing topics being researched in parallel. In this chapter, we will first review literature in demand prediction and decision making under uncertainty separately, and then introduce a data-driven method that optimizes both in one model.

3.2.2 Demand Prediction

In recent years, a lot of studies apply deep learning to forecast ride-hailing demand. The state-of-the-art method is the class of Convolutional LSTM (CNN-LSTM) models because of their capacity in capturing the spatiotemporal travel demand patterns. The appropriate variant of the CNN-LSTM model used in travel demand predictions depends on the structure of the problem. Standard CNN is designed to analyze quantities on urban grids, and convolutions are defined with respect to neighboring cells on an imposed artificial grid [184, 92]. Due to the irregular shapes of taxi zones,

graph neural networks were used and different types of correlations between spatial entities are defined by adjacency matrices [90, 185, 56].

All machine learning methods are concerned with selecting the best estimators via Empirical Risk Minimization (ERM), where weights of the network are obtained via gradient-based algorithms such that the empirical average loss is minimized. The loss functions are often standard, differentiable functions: log-likelihood for predicting distributions, cross-entropy for classification, and mean squared error (MSE) for regression. However, this implicitly assumes that the losses for each sample are equally weighted. For example, with an MSE loss function, over(under)-predicting n people yield the same error regardless of the actual demand. However, in downstream applications such as vehicle rebalancing, the actual decision loss of over(under)-predicting a certain amount of demand is highly likely to be different.

3.2.3 Decision making under uncertainty in OR

The most widely-used method for decision-making under uncertainty in downstream optimization tasks is stochastic optimization [28]. One traditional method in stochastic optimization is Sample Average Approximation (SAA), where empirical distributions are treated as the true distributions [97]. Another notable approach for decision-making under uncertainty is robust optimization [12], and its data-driven variants [18], where the optimization task considers an uncertainty set around the predicted values and optimizes the worst realization. However, none of the optimization approaches mentioned here utilize auxiliary observations besides the uncertain quantities.

To narrow this gap and combine ML with OR approaches, Bertsimas et al. [20] proposed a predictive prescription framework for decision making under uncertainty where auxiliary observations and data are leveraged to prescribe optimal decisions directly from data in the optimization model. In this chapter, the predictive prescription framework is introduced into the vehicle rebalancing problem and it is compared with both point-prediction-driven optimization models with advanced LSTM networks, sample average approximation, and robust optimization models.

3.3 Methodology

3.3.1 Matching-integrated Vehicle Rebalancing Model

In this section, we briefly describe the matching-integrated vehicle rebalancing (MIVR) model proposed by Guo et al. [67], which constitutes the optimization component of proposed data-driven approaches in this chapter.

The operational period is divided into Ω identical time intervals indexed by $k = 1, 2, ..., \Omega$, where each time interval has length Δ. The MIVR model is solved in a rolling-horizon manner illustrated in Figure 3-1, where decision variables are determined repeatedly at the beginning of each time interval. The MIVR model is solved considering four future time intervals ($\kappa = 4$). Red intervals indicate the look-ahead window while green intervals represent the current decision time intervals whose rebalancing decisions will be implemented. It is worth mentioning that the MIVR model is a forward-looking model incorporating κ future time intervals. When solving the MIVR model at the beginning of time interval k, including the demand during time interval k, κ future time intervals are considered. Only the vehicle rebalancing decisions of the current time interval k will be implemented. Then the vehicle locations are observed and submitted to the MIVR model as inputs for the next time interval.

Additionally, the study area is divided into n sub-regions (zones), where each sub-region i has an estimated demand $r_i^k \geq 0$ at time k. We introduce two sets in this model: i) set of sub-regions $N = \{1, ..., n\}$, and ii) set of time intervals $K = \{1, ..., \kappa\}$.

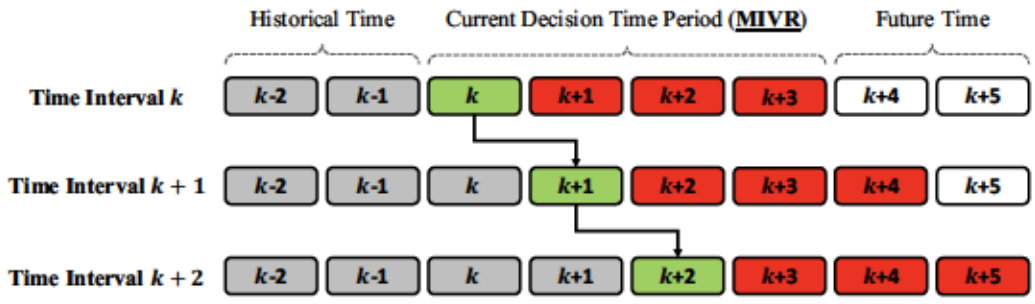


Figure 3-1: Example of rolling horizon manner for solving the MIVR model.

For each time interval, the MIVR model performs two tasks: i) vehicle rebalancing, which happens at the beginning of each time interval, and ii) driver-customer matching, which is conducted at the end of each time interval.

In the vehicle rebalancing phase, decision variables are represented by $x_{ij}^k \in \mathbb{R}_+$ denoting the number of idle vehicles rebalanced from sub-region i to sub-region j at time k. The number of available vehicles in sub-region i at the end of time interval k is indicated by $S_i^k \in \mathbb{R}_+$. Let d_{ij}^k, w_{ij}^k denote the travel distance and time from sub-region i to sub-region j at time k, respectively, which can be approximated by the distance and travel time between the centroids of two sub-regions. Let $a_{ij}^k \in \{0, 1\}$ denote whether an idle vehicle can be rebalanced from sub-region i to sub-region j at time k, where $a_{ij}^k = 0$ if rebalancing between sub-regions i, j is feasible at time k. The vehicle rebalancing from sub-region i to sub-region j at time k is feasible if the vehicle can be rebalanced to the destination within the current time interval, i.e., $w_{ij}^k \leq \Delta$. The feasibility constraint for vehicle rebalancing is given by:

$$a_{ij}^k \cdot x_{ij}^k = 0 \quad \forall i, j \in N, \ \forall k \in K. \tag{3.1}$$

In the MIVR model, since the actual quantity and detailed locations of customers and vehicles are not available, the matching component considers interzonal matchings based on estimated demand. In the matching phase, the decision variables are $y_{ij}^k \in \mathbb{R}_+$ denoting the number of customers in sub-region i matched with vehicles in sub-region j at time k. A maximum pickup time is imposed to guarantee that customers do not experience excessive wait times. Let $\bar{w}$ denote customers' maximum pickup time and parameter $b_{ij}^k \in \{0, 1\}$ denote whether customers in sub-region i can be matched with drivers in sub-region j at time k, where $b_{ij}^k = 0$ indicates a feasible interzonal matching. The matching feasibility constraint is

$$b_{ij}^k \cdot y_{ij}^k = 0 \quad \forall i, j \in N, \ \forall k \in K. \tag{3.2}$$

The number of unsatisfied requests in sub-region i at time k is represented as $T_i^k \in \mathbb{R}_+$. Then constraints related to the matching phase are:

$$\sum_{j=1}^{n} y_{ji}^k \leq S_i^k \quad \forall i \in N,\ \forall k \in K \tag{3.3a}$$

$$\sum_{j=1}^{n} y_{ij}^k \leq r_i^k \quad \forall i \in N,\ \forall k \in K \tag{3.3b}$$

$$T_i^k = r_i^k - \sum_{j=1}^{n} y_{ij}^k \quad \forall i \in N,\ \forall k \in K \tag{3.3c}$$

Constraints (3.3a) and (3.3b) restrict the number of interzonal matchings by the number of available vehicles S_i^k and estimated demand r_i^k. The number of unsatisfied requests is defined as $T_i^k \in \mathbb{R}_+$ by constraints (3.3c), which is the number of customers who have not been assigned drivers within the current matching phase.

To connect matching and rebalancing phases in the MIVR model, we introduce the following decision variables and parameters:

- $V_i^k \in \mathbb{R}_+$: number of vacant vehicles for sub-region i at the beginning of time interval k.
- $O_i^k \in \mathbb{R}_+$: number of occupied vehicles for sub-region i at the beginning of time interval k.
- $V_i^1, O_i^1, \forall i \in N$: initial vehicle locations.
- $P^k(P_{ij}^k)$: the probability that an occupied vehicle located in sub-region i at time k will be in sub-region j and stay occupied at time $k+1$.
- $Q^k(Q_{ij}^k)$: the probability that an occupied vehicle starting in sub-region i at time k will be in sub-region j and become vacant at time $k+1$.

P^k, Q^k are regional transition matrices describing the movement of occupied vehicles. We approximate them with static matrices estimated from historical data. The approach to estimating these matrices and the limitations of such approximations are discussed in [67].

Then, we specify the following relationships between S_i^k, V_i^k, O_i^k and decision variables x_{ij}^k, y_{ij}^k:

$$\sum_{j=1}^{n} x_{ij}^k \leq V_i^k \quad \forall i \in N,\ \forall k \in K \tag{3.4a}$$

$$S_i^k = V_i^k + \sum_{j=1}^{n} x_{ji}^k - \sum_{j=1}^{n} x_{ij}^k \quad \forall i \in N,\ \forall k \in K \tag{3.4b}$$

$$V_i^{k+1} = S_i^k - \sum_{j=1}^{n} y_{ji}^k + \sum_{j=1}^{n} Q_{ji}^k O_j^k \quad \forall i \in N,\ \forall k \in K \setminus \{\kappa\} \tag{3.4c}$$

$$O_i^{k+1} = \sum_{j=1}^{n} y_{ji}^k + \sum_{j=1}^{n} P_{ji}^k O_j^k \quad \forall i \in N,\ \forall k \in K \setminus \{\kappa\} \tag{3.4d}$$

Constraint (3.4a) ensures that the maximum number of vehicles in sub-region i that can be rebalanced to other sub-regions is the number of vacant vehicles at the beginning of time intervals. Constraint (3.4b) states that available vehicles in sub-region i at time k consist of vacant and rebalanced vehicles. Constraint (3.4c) shows that vacant vehicles in sub-region i at time $k+1$ are comprised of currently vacant vehicles at time k and currently occupied vehicles that become vacant in the next time interval. Constraint (3.4d) states that occupied vehicles in sub-region i at time $k+1$ are comprised of currently vacant vehicles that become occupied in the next interval as well as currently occupied vehicles at time k.

The MIVR model minimizes the number of unsatisfied requests and the total vehicle distance traveled, which consists of vehicle rebalancing distance and vehicle pickup distance. Let γ indicate the penalty (in the unit of VMT[1]) induced by each unsatisfied request, and β defines the relative weighting of rebalancing distance and pickup distance. The MIVR model can be formulated as:

[1]VMT stands for vehicle miles traveled.

$$\min_{\boldsymbol{x}\in\mathcal{X}} \quad c(\boldsymbol{x};\boldsymbol{r}) = \sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^{k} d_{ij}^{k} + \min_{(\boldsymbol{y},\boldsymbol{T})\in\mathcal{L}(\boldsymbol{x},\boldsymbol{r})} \left\{ \gamma \cdot \sum_{k=1}^{\kappa}\sum_{i=1}^{n} T_{i}^{k} + \beta \cdot \sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}^{k} d_{ji}^{k} \right\}, \tag{3.5}$$

Where

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{r}) = \left\{ (\boldsymbol{y},\boldsymbol{T}) \in \mathbb{R}_{+}^{n^2\kappa\times n\kappa} : \text{Constraints}(2),(3),(4) \right\},$$
$$\text{and} \quad \mathcal{X} = \left\{ \boldsymbol{x} \in \mathbb{R}_{+}^{n^2\kappa} : \text{Constraints}(1) \right\}.$$

To simplify the notation, we ignore auxiliary variables $\boldsymbol{S}, \boldsymbol{V}, \boldsymbol{O}$ in problem (3.5) and only keep the rebalancing decision vector $\boldsymbol{x}$ and two auxiliary decision vectors $\boldsymbol{y}, \boldsymbol{T}$. The demand vector is denoted as $\boldsymbol{r} \in \mathbb{R}_{+}^{n\kappa}$, which serves as the input parameter of the MIVR model. The MIVR model is a linear programming (LP) problem and can be solved efficiently by off-the-shelf LP solvers, even for large-scale instances (e.g., $n = 500$).

Solving the MIVR model (3.5) requires the prediction of demand $\boldsymbol{r}$ for future κ time periods. Suppose we are given historical data $(\boldsymbol{z}^i, \boldsymbol{r}^i), i = 1, ..., m$, where $\boldsymbol{z}^i \in \mathbb{R}^{n\times\kappa\times p}$ denotes the independent variables with p features, $\boldsymbol{r}^i \in \mathbb{R}^{n\times\kappa}$ is a demand vector which depends upon $\boldsymbol{z}^i$, and m is the number of previous days whose information is provided in the data.

For instance, if we are solving a MIVR model at 9:00 AM today and we would like to predict the future demand $\boldsymbol{r}$ from 9:00 AM to 10:00 AM, we can utilize the historical demand and features between 9:00 AM to 10:00 AM from previous m days, i.e., $\{(\boldsymbol{z}^i, \boldsymbol{r}^i) : \forall i = 1, ..., m\}$, to predict the demand today. Meanwhile, we also have access to a feature vector $\boldsymbol{z}$ with exogenous information such as temperature and precipitation for the time period to be predicted.

There are two ways that demand information can be incorporated into the model: point-predictions or data-driven optimization. The former method follows a two-step

approach where a point prediction is first produced based on historical observations and auxiliary data independent of the optimization model. Then rebalancing decisions are made according to the point predictions. Data-driven optimization methods, on the other hand, directly prescribe rebalancing decisions from historical observations and auxiliary data.

3.3.2 Point-Prediction-Driven Optimization

In point-prediction-driven optimization, a predictive model is first developed. Let $f(\cdot)$ represent such a predictive model to predict the unknown demand vector $\boldsymbol{r}$, i.e., $f(\boldsymbol{z}) = \hat{\boldsymbol{r}}$. $f(\cdot)$ can be established based on the data $\{(\boldsymbol{z}_i, \boldsymbol{r}_i), i = 1, ..., m\}$ with machine learning methods. The predicted demand $\hat{r}$ is then fed into the MIVR model as $\hat{\boldsymbol{r}}$ to get the rebalancing decisions:

$$\hat{\boldsymbol{x}}^{point-pred} = \underset{\boldsymbol{x} \in \mathcal{X}}{\operatorname{argmin}}\, c(\boldsymbol{x}; \hat{\boldsymbol{r}}). \tag{3.6}$$

Recent developments on short-term travel demand prediction focus on capturing the spatial-temporal patterns of travel demand using deep learning. The state-of-the-art architecture is the class of Convolutional Long Short Term Memory (Conv-LSTM) networks, where the standard LSTM is extended by having a convolutional structure in both input-to-state and state-to-state transitions [142]. Since sub-regions do not conform to a grid structure, graph convolution proposed by Kipf and Welling [95] is adopted instead of grid convolutions.

Suppose we have L_g graph convolutional layers and the output of the hidden layers is denoted as $H^{(l)}$, $l = 1, ..., L_g$, we have the following layer-wise propagation rule:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{3.7}$$

where $\sigma(\cdot)$ is an activation function (most commonly ReLU); $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ is the degree matrix; $\tilde{A} = A + I_N$ is the adjacency matrix with added self-connections; and $W^{(l)}$ is the trainable weights of layer l.

Graph convolution layers require upfront access to the global structure of the

graph in the form of adjacency matrices (A). In this case, the Euclidean distance between the centroids of the sub-regions is used to relate to neighboring sub-regions.

$$[A_E]_{ij} = \frac{1}{\text{Euclidean Distance}(i,j)} \tag{3.8}$$

where Euclidean distance is defined as the straight-line distance between the centroids of sub-regions i and j.

In addition to Graph Convolutional LSTM (Graph Conv-LSTM), two LSTM networks without spatial convolution were also constructed as benchmarks. Time series of past demand in each zone are treated as inputs to the model and no spatial correlation between zones is considered. The difference is that in one model, named All Zones LSTM in subsequent discussions, the temporal correlation between different zones is assumed to be the same. Time series from all zones were used to estimate the one-zone model. In the model named Single Zone LSTM, one LSTM is separately trained for each zone. Since All Zones LSTM is a subset of Single Zone LSTM, it is expected that the predictive performance of All Zones LSTM will be the worst among the three models.

3.3.3 Data-driven Optimization

Instead of producing a point estimate, there exist data-driven optimization approaches that can prescribe decisions directly from data. First, we consider a simple data-driven approach, SAA, in this section, which is used as a baseline model. Given a finite sample of data, the SAA approach assumes that the demand vector $\boldsymbol{r}^i$ are drawn uniformly at random from dataset $\{\boldsymbol{r}^i\}_{i=1}^m$. Therefore, the MIVR problem can be written as:

$$\hat{\boldsymbol{x}}^{SAA} = \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} c(\boldsymbol{x};\boldsymbol{r}^i). \tag{3.9}$$

Although SAA accounts for the data uncertainty, it does not utilize any auxiliary information described in $\{\boldsymbol{z}^i\}_{i=1}^m$, which incurs an unacceptable waste of good data. Therefore, we introduce the predictive prescription approach to this problem. Pro-

posed in [20], this framework combines ML and OR techniques and utilizes auxiliary information.

Compared to the traditional SAA approach where only demand vectors $\{\boldsymbol{r}^i\}_{i=1}^m$ are considered for generating rebalancing decisions, the predictive prescription leverages auxiliary observations $\{\boldsymbol{z}^i\}_{i=1}^m$ and solve the following problem:

$$\hat{\boldsymbol{x}}(\boldsymbol{z}) = \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{m} w_i(\boldsymbol{z}) c(\boldsymbol{x}; \boldsymbol{r}^i), \tag{3.10}$$

where $w_i(\boldsymbol{z})$ stands for weight functions derived from historical data $\{(\boldsymbol{z}^i, \boldsymbol{r}^i), i = 1, ..., m\}$ and current observation $\boldsymbol{z}$. The predictive prescription utilizes machine learning algorithms to generate "smarter" weights compared to identical weights used in the SAA approach.

In this chapter, we introduce two machine learning algorithms for generating weights $[w_i(\boldsymbol{z})]_{i=1}^m$. The first algorithm is one of the most commonly used unsupervised learning algorithm, k-nearest-neighbors (KNN). The rebalancing decisions can be generated by solving the following problem:

$$\hat{\boldsymbol{x}}^{KNN}(\boldsymbol{z}) = \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmin}} \sum_{i\in\mathcal{N}_k(\boldsymbol{z})} c(\boldsymbol{x}; \boldsymbol{r}^i), \tag{3.11}$$

where $\mathcal{N}_k(\boldsymbol{z})$ represents the set of k data points that are closest to $\boldsymbol{z}$, i.e., $\mathcal{N}_k(\boldsymbol{z}) = \{i = 1, ..., m : \sum_{j=1}^m \mathbb{1}[\|\boldsymbol{z} - \boldsymbol{z}^i\| \geq \|\boldsymbol{z} - \boldsymbol{z}^j\|] \leq k\}$.

The second algorithm considered in this chapter is the optimal regression tree (ORT) proposed in [17], which generates a regression tree with better prediction accuracy than the standard classification and regression tree (CART) approach. The predictive prescription with ORT is formulated as:

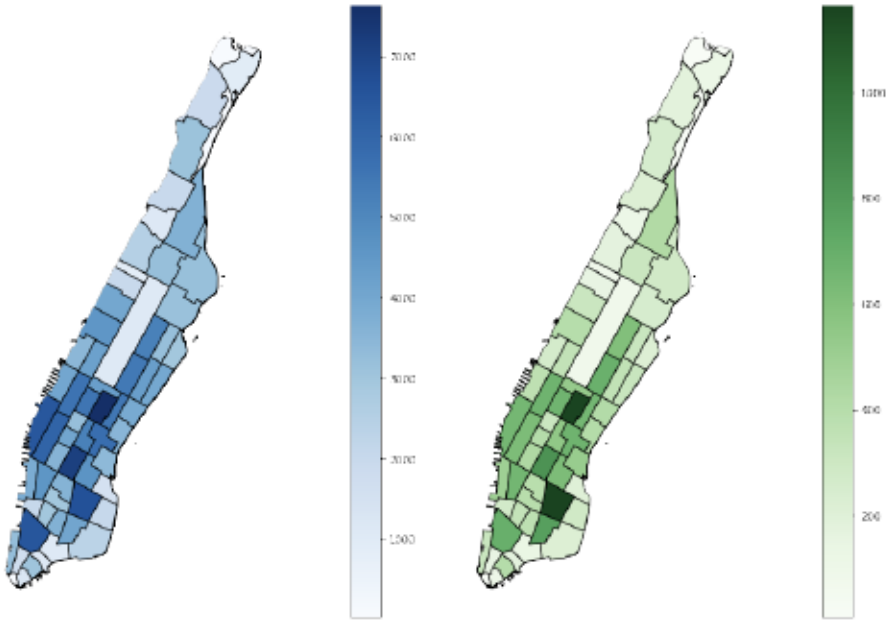
$$\hat{\boldsymbol{x}}^{ORT}(\boldsymbol{z}) = \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmin}} \sum_{i:R(\boldsymbol{z}^i)=R(\boldsymbol{z})} c(\boldsymbol{x}; \boldsymbol{r}^i), \tag{3.12}$$

where $R(\boldsymbol{z})$ corresponds to the leaf of current observation $\boldsymbol{z}$ in the ORT trained on the dataset.

3.4 Data Description

The study area is the island of Manhattan in New York City (NYC) and demand data used in this chapter is the high-volume ride-hailing trip data [130]. The data includes pickup and drop-off times and locations for all trips made using "high-volume" ride-hailing services, defined as any service that dispatches more than 10,000 trips per day within New York City, including Uber, Lyft, Juno, and Via. We use the data from 20 workdays of June 2019 and the demand is aggregated to 5-minute time intervals.

The sub-regions used in the experiments are "taxi zones" defined within the high-volume ride-hailing trip dataset. There are in total 63 taxi zones on the island of Manhattan. Real travel speed data from June 2019 provided by the Uber Movement database [156] is used for generating interzonal travel times w_{ij}^k. The regional transition probability matrices for occupied vehicles, P^k, and Q^k are generated based on the real travel time and demand data, and details can be found in [67]. Figure 3-2 shows the mean and standard deviation of daily regional demand in Manhattan. Regions near lower Manhattan have large standard deviations, which imply that accurately predicting demand is not a trivial task when making vehicle rebalancing decisions.



(a) Mean (b) Standard deviation

Figure 3-2: Daily demand by zone (trips) in Manhattan.

The auxiliary information used in experiments for both predictive and prescriptive models includes:

- **Weather**: hourly weather data, including air temperature, sensible temperature, precipitation, and snowfall.
- **Point of Interests (POIs)**: number of residential, education, recreational, commercial, and health POIs.
- **Public transit accessibility**: number of subway stations and bus stops.
- **Historical demand**: average demand from previous five-time intervals and historical average demand from m previous days.

Since POI and transit stops/stations are time-independent, they are not used in predictive prescription models.

3.5 Experimental Results

In this section, we compare model performances of the following approaches: i) point-prediction-driven optimization, ii) SAA, iii) predictive prescriptions and iv) robust MIVR model proposed in [67], as well as two benchmark models: i) optimization with historical average and ii) optimization with true demand under four different demand scenarios. Linear programs in this chapter are modeled with open-source Julia [25] package JuMP [52] and solved with Gurobi 9.0 [74] on a 3.0 GHz AMD Threadripper 2970WX Processor with 128 GB Memory.

3.5.1 Model Evaluation and Demand Scenarios

To evaluate model performances, we set the last weekday (June 28th, 2019) in our dataset as the test day during which vehicle rebalancing decisions need to be made without knowing the true demand. Data for the previous 19 weekdays are used to construct predictive models and serve as model inputs for data-driven optimization

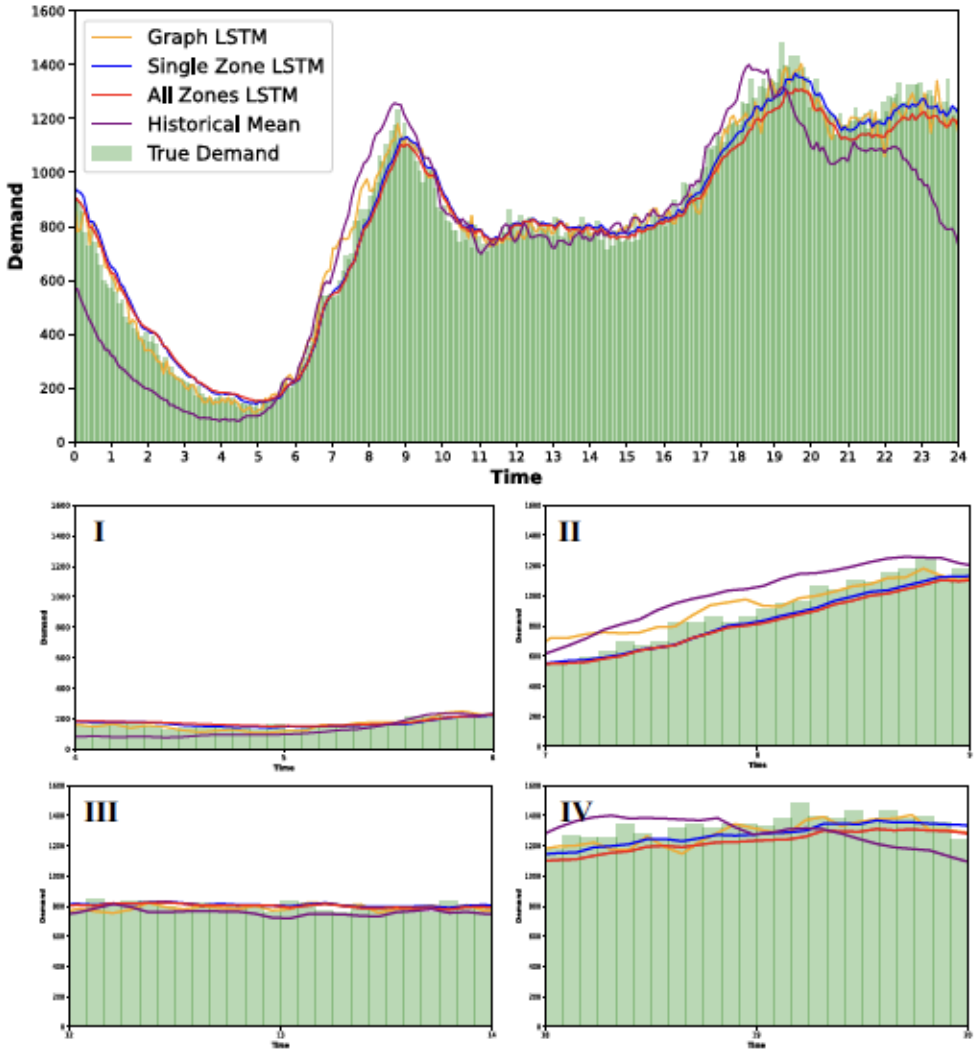


Figure 3-3: Demand levels for four different demand scenarios.

models. Vehicle rebalancing models are evaluated with four different 2-hour demand scenarios which are shown in Figure 3-3[2]:

I Morning off-peak scenario (4 - 6): Total demand level is low while point predictions are accurate.

II Morning peak scenario (7 - 9): Total demand level is high while point predictions are not accurate. Zone-based LSTM underestimates the total demand level.

[2]Green histogram: true demand. Orange line: predicted demand with Graph LSTM. Blue line: predicted demand with Single Zone LSTM. Red line: predicted demand with All Zones LSTM. Purple line: historical average demand.

III Mid-day off-peak scenario (12 - 14): Total demand level is high while point predictions are accurate.

IV Evening rush hour scenario (18 - 20): Total demand level is high while point predictions are not accurate. Both zone-based LSTM and graph-based LSTM underestimate the total demand level.

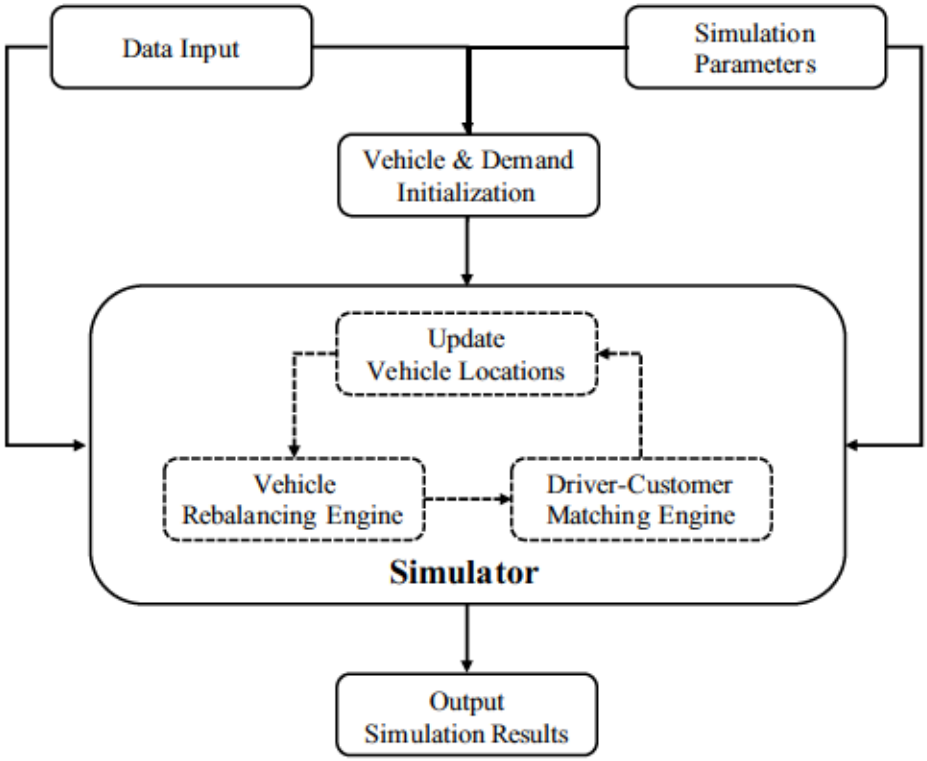


Figure 3-4: Simulation framework for evaluating vehicle rebalancing models.

Model Parameter	Explanation	Value
β	Weight parameter for pickup distance	1
γ	Penalty for unsatisfied requests	10^2
Ω	Total number of time intervals in the simulation for each demand scenario	24
Δ	Decision time interval length for vehicle rebalancing problem	300 (seconds)
δ	Batch size for driver-customer matching problem	30 (seconds)
$\bar{w}$	Maximum pickup time	300 (seconds)
$\tilde{w}$	Maximum wait time	300 (seconds)
n	Number of sub-regions	63
κ	Number of look-ahead time intervals when solving MIVR model	6
m	Number of historical data points	19
N_v	Number of vehicles	2000

Table 3.1: Model parameters and values.

The simulation framework is shown in Figure 3-4. The input data includes road network for the Manhattan area with shortest path distance and predecessor matrices, distance and travel time matrices between taxi zones, regional transition matrices, demand data, and weights for predictive prescriptions. Parameters used in the simulation are shown in Table 3.1. Fleet size is set to be 2000 vehicles in the simulation. With the setup described above, vehicle and demand locations are initialized. Vehicles are all available and equally distributed to the taxi zones at the beginning of the simulation. Given that origins and destinations of demand are at the sub-regional level, road nodes within the sub-regions are randomly assigned as origins and destinations for customers in each sub-region.

After initializing vehicle and demand locations, a simulation consisting of a vehicle rebalancing engine and a driver-customer matching engine is run with different rebalancing models. In the simulator, a vehicle rebalancing problem is solved at the start of each time period of length Δ and the vehicle locations are updated before solving vehicle rebalancing problems. A separate driver-customer matching problem is solved at the end of each time period of length δ with available vehicles and realized demand. Details about the driver-customer matching problem can be found in Appendix B.1. The simulation outputs average customer wait time, unsatisfied customer rate, average non-occupied VMT, and an average number of rebalancing trips for the evaluation of different rebalancing models.

3.5.2 Performance of Point Predictions

To ensure that there are enough training samples for neural networks, we utilized additional workday demand data in April and May 2019 in the model training stage. The hyperparameters used in the LSTMs are shown in Table 3.2.

Prediction accuracy for different LSTM models and the benchmark historical average model for the full day are shown in Table 3.3. All machine learning models significantly outperform the historical average. Among the machine learning models, Graph Conv-LSTM has the most representation power, therefore the training error was the smallest. The test set performances for Graph Conv-LSTM and Single Zone

Hyperparameter	Value
# GCN layers	2
# Units in hidden layers	64
# LSTM layers	1
Weight decay	0.005
Learning rate	0.005

Table 3.2: LSTM model setup.

LSTM were similar. All Zones LSTM has the worst performance since it does not differentiate demand from different zones.

Model	Train MSE	Test MSE	Test MAE
Historical Average	23.18	29.73	3.64
Graph Conv-LSTM	15.63	16.64	2.93
Single Zone LSTM	16.84	16.52	2.93
All Zones LSTM	17.04	18.26	3.04

Table 3.3: Prediction performance.

For the prediction performance under each demand scenario, the MAE is shown in Table 3.4. For both off-peak demand scenarios (I and III), predictive models have higher prediction accuracy compared to peak demand scenarios (II and IV). Meanwhile, higher demand leads to higher prediction errors. For peak demand scenarios, zone-based LSTM underestimates the overall demand. Graph-based LSTM only underestimates the overall demand in scenario IV. In the next subsection, we will show that making inaccurate predictions (demand underestimation) could potentially benefit the system's performance.

3.5.3 Performance of Different Vehicle Rebalancing Models

In this subsection, we compare the model performances of point-prediction-driven optimization and data-driven optimization, along with two benchmark models: optimization with historical average and optimization with true demand. For predictive models, we constructed a Graph Convolutional LSTM model, a Single Zone LSTM, and an All Zones LSTM model for predicting future demand and generated optimal vehicle rebalancing decisions with point estimations by solving the problem (3.6). For

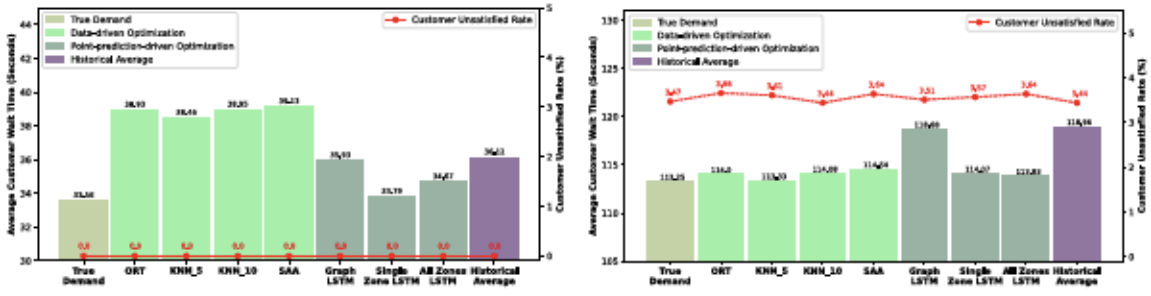
Model	Scenario I (MAE)	Scenario II (MAE)
Historical Average	1.40	3.64
Graph Conv-LSTM	1.36	3.25
Single Zone LSTM	1.36	3.18
All Zones LSTM	1.40	3.28
Model	Scenario III (MAE)	Scenario IV (MAE)
Historical Average	2.89	4.87
Graph Conv-LSTM	2.81	3.76
Single Zone LSTM	2.85	3.82
All Zones LSTM	2.97	4.20

Table 3.4: Prediction performances under four different demand scenarios.

optimization with the historical average, we used the average demand of m previous workdays as point estimations and solved the problem (3.6). Similarly, the optimization with true demand utilized the real demand as point estimations and solved the problem (3.6).

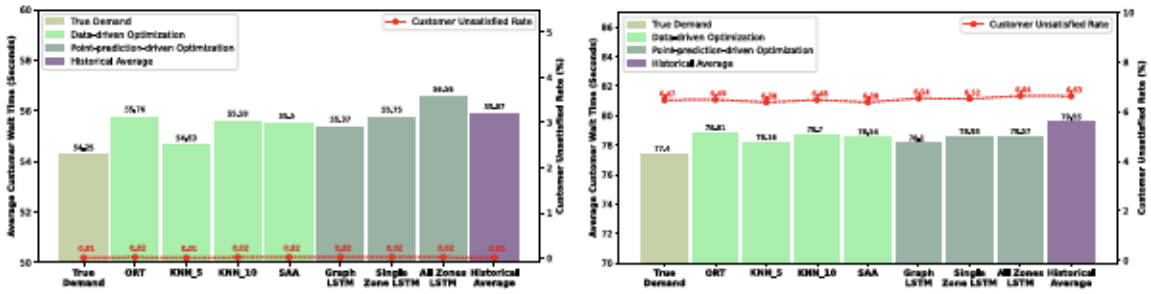
Four data-driven models are considered. The SAA model is included as a benchmark and three predictive prescription models are tested: two KNN models (3.11) with $k = 5$ and $k = 10$, and an ORT model (3.12). Weights for m historical days used in predictive prescriptions were generated in the following way. First, a vector $e \in \mathbb{R}^m$ is initialized with m zero values. Then for each unique pair of zones and time intervals, KNN or ORT algorithms were run with m historical data points $\{z_i, i = 1, ..., m\}$ and the current observation z, After that, i-th value in vector e was increased by 1 if z_i is within k nearest neighbors of z or z_i and z belong to the same branch in the constructed ORT. Finally, the weights were generated by normalizing vector e.

Figure 3-5 shows customer wait times and unsatisfied requests, which are key performance indicators of a ride-hailing system, under four demand scenarios. In each sub-figure, colored bars represent the average customer wait time after matching to vehicles while red dotted lines indicate the customer unsatisfaction rate. To better understand how each vehicle rebalancing model works under each demand scenario, the average non-occupied VMT and the average rebalancing trips for each vehicle are shown in Figure 3-6 and 3-7. In all four demand scenarios, knowing the true demand



(a) Morning off-peak scenario (4 - 6)

(b) Morning peak scenario (7 - 9)



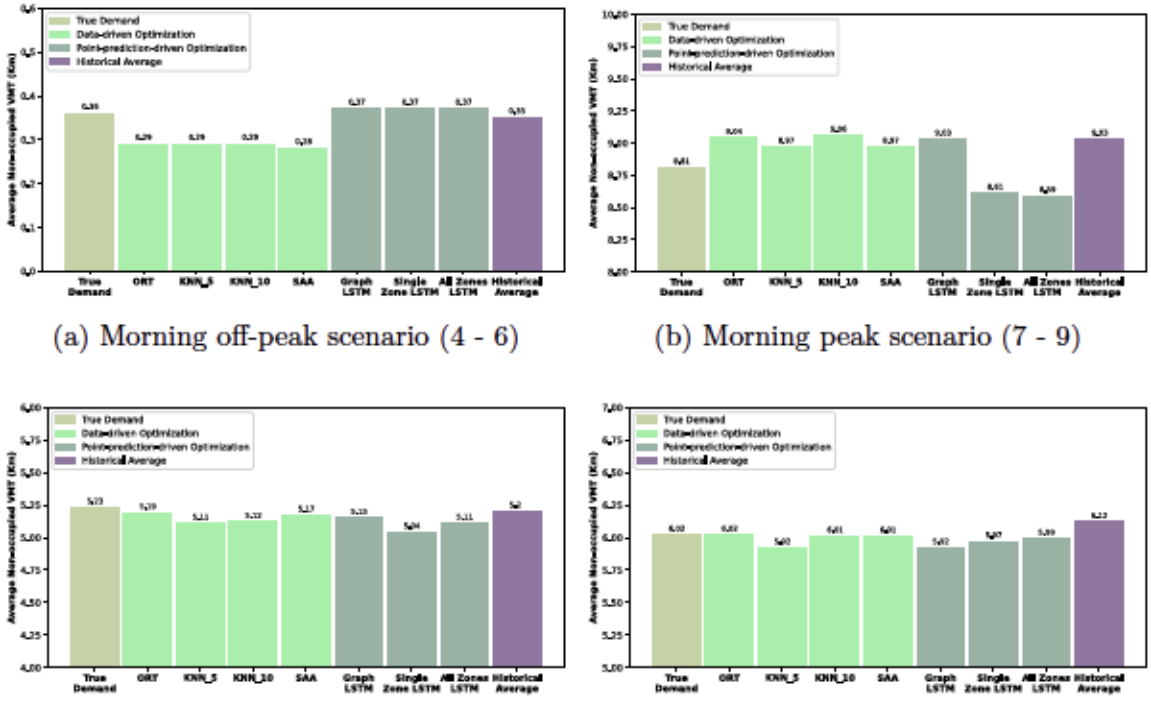
(c) Mid-day off-peak scenario (12 - 14)

(d) Evening rush hour scenario (18 - 20)

Figure 3-5: Customer wait time and unsatisfied rate for different demand scenarios.

leads to the minimum average customer wait time compared to applying any data-driven approaches. The performance comparisons between data-driven optimization and point-prediction-driven optimization vary across different demand scenarios.

In the morning off-peak scenario, the overall demand level is low and all predictive models are more accurate compared to other time periods. Under this demand scenario, point-prediction-driven optimization outperforms data-driven optimization since future demand predictions are very accurate. Figure 3-5a indicates that data-driven optimization approaches perform even worse than only knowing the historical average demand. On the other hand, data-driven optimization approaches conduct much fewer vehicle rebalancing trips according to Figure 3-7a. When combining with the average non-occupied VMT for each vehicle shown in Figure 3-6a, we know that data-driven optimization approaches distribute fewer idle vehicles with longer distances. The poor performances of data-driven optimization models imply that several days with low demand levels are deemed more relevant by the model. To summarize, when demand can be accurately predicted, point-prediction-driven optimization



(a) Morning off-peak scenario (4 - 6)

(b) Morning peak scenario (7 - 9)

(c) Mid-day off-peak scenario (12 - 14)

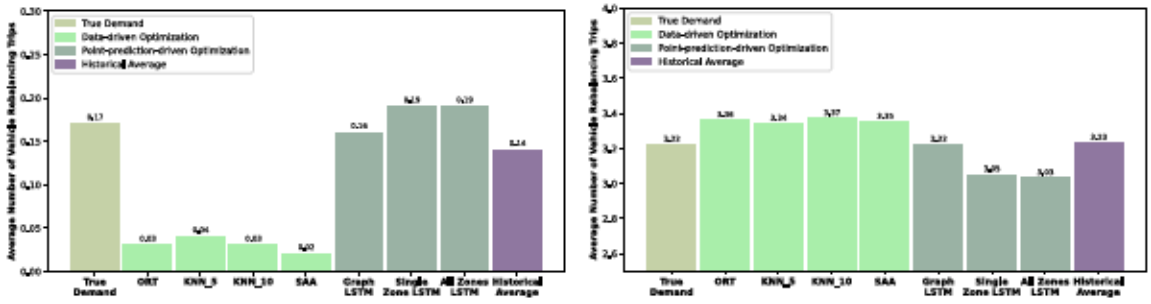
(d) Evening rush hour scenario (18 - 20)

Figure 3-6: Average non-occupied VMT for each vehicle under different demand scenarios.

should be used.

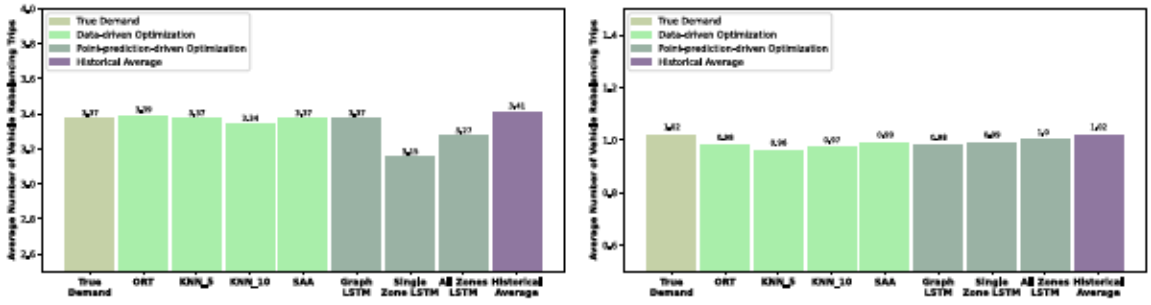
For the morning peak scenario, the overall demand level is high and predictive models have large prediction errors. Figure 3-5b shows the average customer wait time and customer unsatisfied rate for each model. The customer unsatisfaction rate is fairly close across all different models. Under this demand scenario, data-driven optimization models perform better overall compared to point-prediction-driven optimization models. All four data-driven optimization models achieve competitive performances with respect to the optimal case in which true demand is known. For predictive models, the graph-based LSTM has the worst performance while two zone-based LSTMs have competitive performances compared to data-driven optimization models.

It is worth mentioning that Graph LSTM has better prediction accuracy than All Zones LSTM though it has a worse model performance. The main reason for zone-based LSTMs to have satisfying performances is that they underestimate fu-



(a) Morning off-peak scenario (4 - 6)

(b) Morning peak scenario (7 - 9)



(c) Mid-day off-peak scenario (12 - 14)

(d) Evening rush hour scenario (18 - 20)

Figure 3-7: Average number of rebalancing trips made for each vehicle under different demand scenarios.

ture demand, which is shown in Figure 3-3. Also from Figure 3-6b and 3-7b, less rebalancing trips and lower non-occupied VMT imply the demand underestimation by zone-based LSTMs. The "conservativeness" brought by the underestimation leads to better system performances given high volatility in ride-hailing demand. Being conservative is also the key reason for the robust MIVR model proposed in [67] to have satisfying performances. The simulation results justify that a better demand prediction does not necessarily lead to a better rebalancing decision. Meanwhile, underestimation is a more desirable prediction error to make than overestimation when predicting future demand for the purpose of distributing vacant vehicles.

As for the mid-day demand scenario, the overall demand is at a medium level while predictive models are more accurate than the two peak demand scenarios. Under this demand scenario, data-driven optimization models perform better overall compared to point-prediction-driven optimization models based on Figure 3-5c. Figure 3-6c and 3-7c show that they conduct similar number of rebalancing trips with similar

distance. All Zones LSTM performs worse than the model knowing the historical average since it has a worse prediction accuracy. When the demand prediction is not accurate enough, data-driven optimization has a close edge over point-prediction-driven optimization.

Under the evening rush hour scenario, the overall demand level is high and predictive models have the worst performances compared to the other three demand scenarios. In this demand scenario, data-driven optimization and point-prediction-driven optimization have similar performances regarding the average customer wait time and customer unsatisfaction rate according to Figure 3-5d. There are limited idle vehicles that can be rebalanced due to a high demand level in the evening rush hour scenario. Figure 3-7d indicates that the average number of rebalancing trips performed for each vehicle is nearly 1, while the number is over 3 for scenarios II and III with high demand levels. Although data-driven optimization performs better when demand predictions are not accurate, the limited number of idle vehicles leaves no space for data-driven optimization to improve system performances by proactively balancing demand and supply.

Within four data-driven optimization models, predictive prescription with KNN ($k = 5$) performs better than the other three methods by having lower average customer wait times across four demand scenarios. Meanwhile, for scenarios where the demand level is high (morning peak, mid-day off-peak, and evening rush hour), predictive prescription with KNN ($k = 5$) utilizes the minimum VMT over rebalancing idle vehicles. This performance superiority implies that sparsity is an ideal property when applying data-driven optimization. Compared to the predictive prescription with KNN-5, the other three models incorporate more historical demand scenarios, which could diminish the system performance if some demand scenarios are significantly different from the future demand scenario over which rebalancing decisions are made.

To summarize, there are two factors to consider when choosing vehicle rebalancing models: i) supply to demand ratio, and ii) demand prediction accuracy. When the demand can be accurately predicted, point-prediction-driven optimization mod-

els perform the best. When the demand is hard to predict (for example, during rush hour), data-driven optimization models perform the best. System performances can be further improved if the supply to demand ratio is higher, where more idle vehicles are available to be rebalanced. Compared to the standard data-driven optimization approach, SAA, predictive prescriptions perform better by leveraging auxiliary information. On the other hand, when demand cannot be accurately predicted, system performances can benefit from underestimation, so fewer unnecessary rebalancing trips are made. However, predictive models tend to aim for "unbiasedness", where the amount of overestimation and underestimation is the same.

3.5.4 Comparison with the Robust MIVR Model

In this subsection, we compared the best performing data-driven optimization model, prescriptive prescription with KNN-5, with the robust MIVR model proposed in [67] under the morning peak scenario. We evaluated performances of the robust MIVR model under multiple uncertain scenarios defined by uncertain parameters ρ and Γ via the simulation described in section 3.5.1. Parameters ρ and Γ are parameters defining the size of uncertainty set in the robust MIVR model, and details can be found in Appendix B.2.

Figure 3-8 shows the percentage reduction of average customer wait time for the robust MIVR model compared to predictive prescription with KNN-5. Each cell indicates an uncertain scenario (defined by parameters ρ and Γ) in the robust MIVR model. Larger values of uncertain parameters ρ and Γ lead to more conservative rebalancing decisions (since higher demand uncertainty is considered in the model). It is worth mentioning that the uncertain parameter ρ significantly influences the downstream matching performances, while the effect of uncertain parameter Γ is marginal.

In general, the predictive prescription with KNN-5 outperforms the robust MIVR model regarding the average customer wait time. The robust MIVR model could achieve a similar average customer wait time when larger demand uncertainty is considered. On the other hand, the robust MIVR model can satisfy more customers

compared to the predictive prescription with KNN-5, which is shown in Figure 3-9. More customers can be satisfied when considering lower levels of demand uncertainty. The additional customers served by the robust MIVR model are "hard" customers which require longer pickup distances, hence longer wait times. To summarize, predictive prescriptions can reduce the average customer wait time compared to the robust MIVR model. However, a small proportion of customers will not be satisfied, which is likely the reason behind reduced customer wait times.

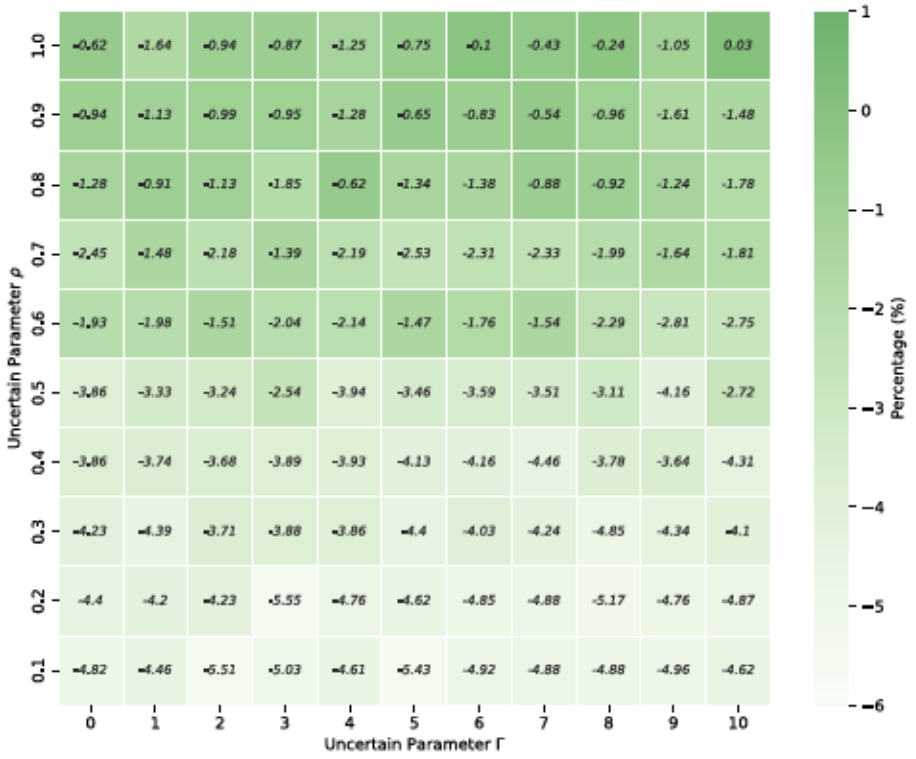


Figure 3-8: Relative percentage reduction of average customer wait time for the robust MIVR model compared to predictive prescription with KNN-5.

Figure 3-10 displays the percentage decrease of average non-occupied VMT for the robust MIVR model compared to predictive prescription with KNN-5. When a certain level of demand uncertainty is considered in the robust MIVR model, it reduces the average non-occupied VMT for each vehicle.

Figure 3-11 exhibits the percentage reduction of average vehicle rebalancing trips for the robust MIVR model compared to predictive prescription with KNN-5. The robust MIVR model significantly reduces the number of rebalancing trips dispatched

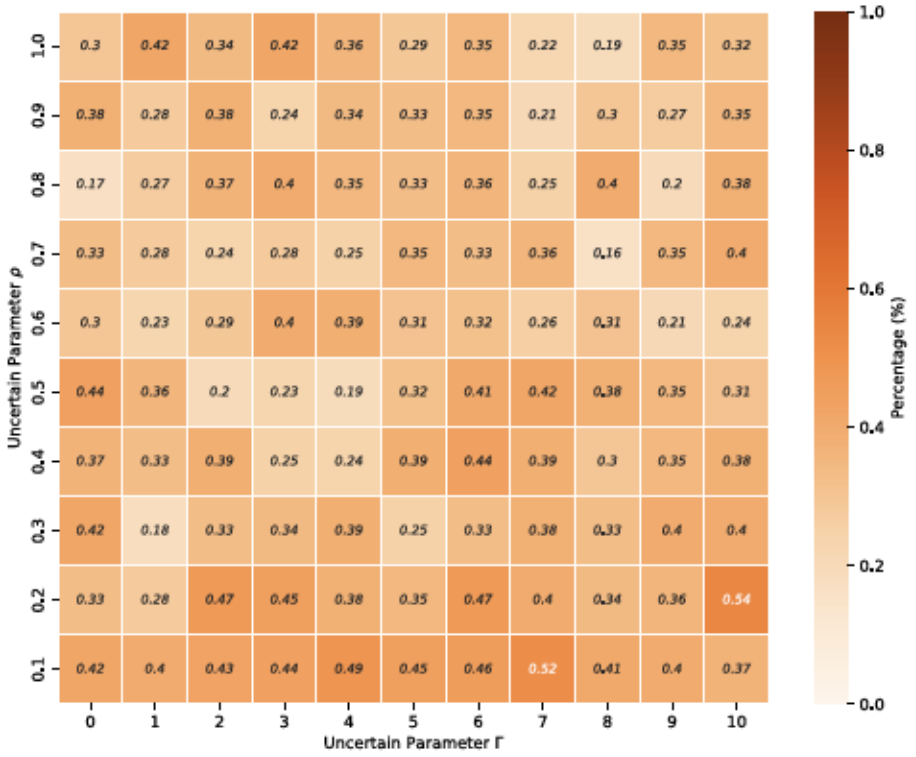


Figure 3-9: Absolute reduction of customer unsatisfied rate for the robust MIVR model compared to predictive prescription with KNN-5.

in the system. Given that robust optimization generates decisions optimal for the worst-case scenario, the robust MIVR model is conservative and few rebalancing trips are made to mitigate the impact of inaccurate demand estimations. On the other hand, predictive prescriptions generate decisions that are optimal for an expected scenario which indicates demand from previous m days. Therefore, they do not maintain the same level of conservativeness as the robust MIVR model.

In conclusion, the robust MIVR model satisfies more customers while conducting fewer rebalancing trips and predictive prescriptions reduce the average customer wait time. From a practical perspective, applying the robust MIVR model requires decision-makers to choose an uncertainty level (ρ and Γ) incorporated in the model for the future demand. While for predictive prescriptions, additional information about the future demand is not required to make rebalancing decisions. Decision-makers should choose the appropriate model based on data availability and confidence about the level of uncertainty in the future demand.

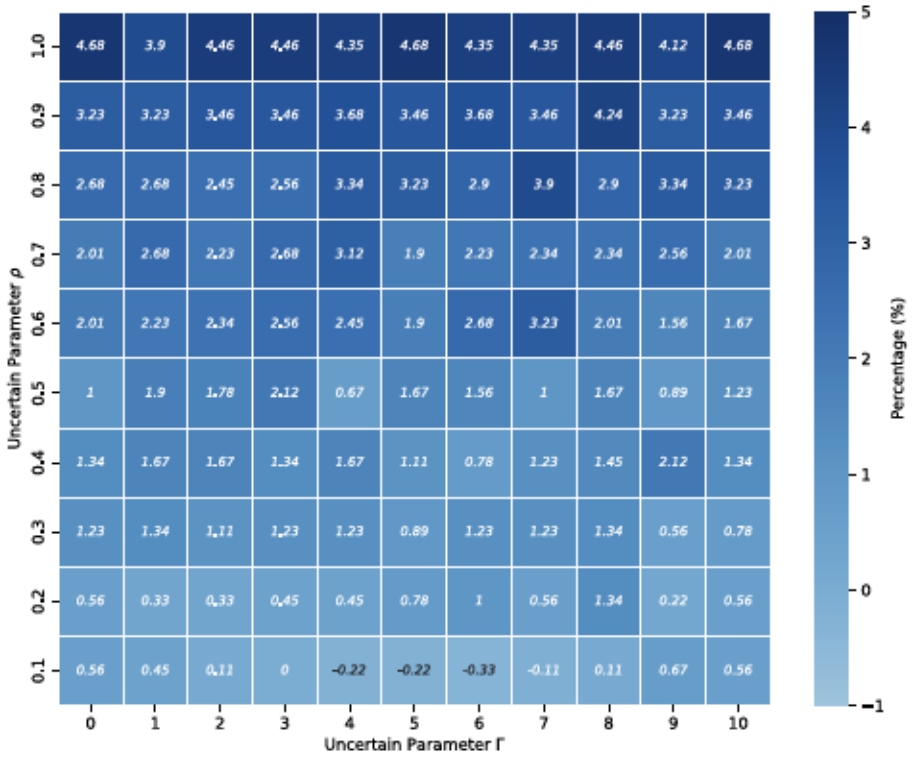


Figure 3-10: Relative percentage decrease of average non-occupied VMT for the robust MIVR model compared to predictive prescription with KNN-5.

3.6 Conclusions

In this chapter, we introduce a novel data-driven optimization approach, predictive prescriptions, into the vehicle rebalancing problem to handle demand uncertainty in the ride-hailing system. Building upon a state-of-the-art vehicle rebalancing model, MIVR proposed by Guo et al. [67], point-prediction-driven optimization models and data-driven optimization models are proposed to improve the model performance against demand uncertainty.

Regarding point-prediction-driven optimization models, a graph convolutional LSTM and two zone-based LSTM models are constructed in this chapter to predict future demand for each sub-region. As for data-driven optimization models, SAA and predictive prescription with KNN and ORT are introduced in this chapter. A real-world simulation with NYC data is used to evaluate performances for point-prediction-driven optimization models, data-driven optimization models, and

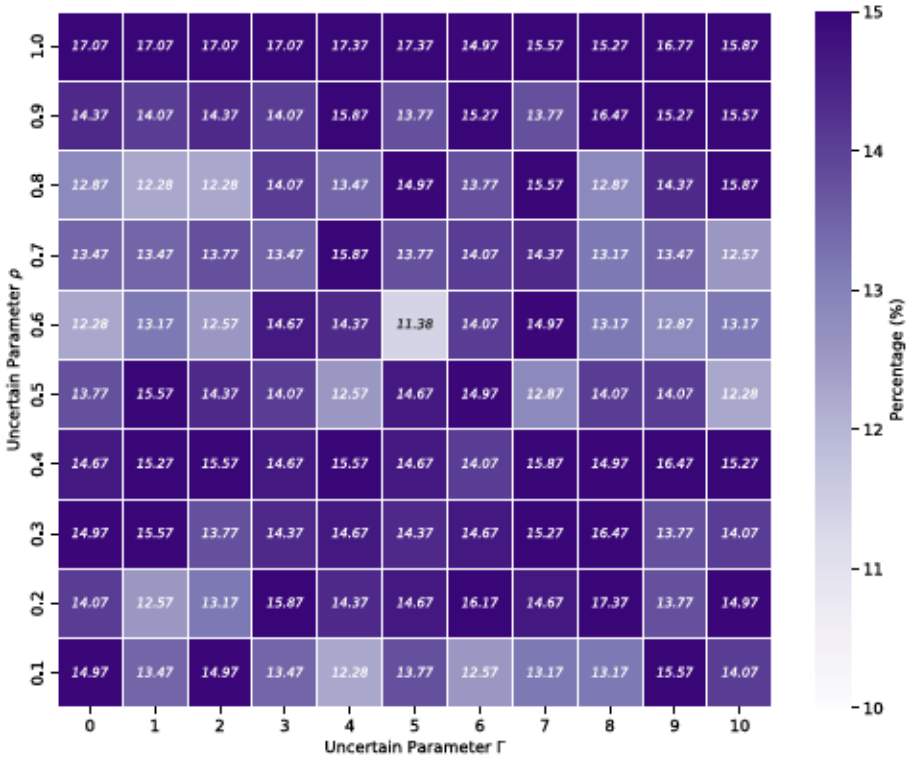


Figure 3-11: Relative percentage reduction of average rebalancing trips for the robust MIVR model compared to predictive prescription with KNN-5.

two benchmark models, optimization with historical average and optimization with true demand, under four different demand scenarios.

Between the data-driven optimization and point-prediction-driven optimization models, one should make a decision based on supply to demand ratio and the prediction accuracy. When the future demand can be predicted accurately, point-prediction-driven optimization models should be adopted. When the demand is volatile and hard to predict, data-driven optimization models perform better. The system performances can be further improved for data-driven optimization models when the supply to demand ratio is higher, indicating more idle vehicles are available to be redistributed. Among all data-driven optimization methods, predictive prescriptions perform better by leveraging auxiliary information.

Meanwhile, prediction errors over the future demand in the vehicle rebalancing problem can be beneficial to system performances when errors come from demand underestimation. The "conservativeness" brought by the demand underestimation

improves the system performance due to highly uncertain demand in the future. The strong performances of the robust MIVR model proposed in [67] are also brought by the "conservativeness" embedded in robust models. However, predictive models usually aim for "unbiasedness", and weights overestimation and underestimation equally. A possible future research direction is to develop predictive models for ride-hailing systems which have an asymmetric loss function that favors underestimation over overestimation. Meanwhile, extra benefits brought by conservativeness due to demand underestimation should have a limit. Future research could identify such underestimation level where the vehicle rebalancing benefits the most.

The best-performing data-driven optimization model, predictive prescription with KNN-5, is also compared with the robust MIVR proposed in [67], which utilizes robust optimization techniques to protect rebalancing decisions against demand uncertainty. The robust MIVR model reduces the customer unsatisfaction rate while conducting fewer vehicle rebalancing trips. On the other hand, predictive prescriptions reduce the average customer wait time but serve fewer customers. In practice, the robust MIVR model should be utilized if knowing the demand uncertainty level in the future. In general, predictive prescriptions can generate competitive rebalancing decisions without knowing any additional future demand information. Another future research direction can be introducing data-driven robust optimization techniques into the MIVR model, which combines the benefits of both data-driven optimization and robust optimization.

From a practical perspective, rebalancing models need to be selected ahead of schedule. When considering a whole day's demand, demand uncertainty and prediction accuracy of predictive models change from time to time. Therefore, a good operation strategy is to separate the whole operation period into high and low uncertainty periods based on historical demand data. For low uncertainty periods, point-prediction-driven optimization models should be adopted. As for high uncertainty periods, data-driven optimization models, including robust and predictive prescription models, can be applied.

Chapter 4

Disparity-Reducing Vehicle Rebalancing in the Ride-hailing System

4.1 Introduction

Since its introduction in 2009, the ride-hailing industry has witnessed significant global growth. Fueled by technological advancements and the widespread adoption of mobile phones, ride-hailing services offered by Transportation Network Companies (TNCs) like Uber, Lyft, and Didi have revolutionized commuting, creating new economic opportunities. With a market size of approximately 30 billion USD and projected to reach 100 billion USD by 2030 [62], the industry continues to meet the increasing demand for convenient and flexible transportation options in today's rapidly urbanizing world.

However, alongside the benefits, the ride-hailing industry has also raised significant societal concerns. Research conducted by Diao et al. [50] indicates that the proliferation of TNCs has exacerbated urban mobility challenges, resulting in increased road congestion and decreased usage of public transit. Underserved communities and low-income neighborhoods have been disproportionately affected by the limited ac-

cessibility and affordability of ride-hailing services. Additionally, the heavy reliance on algorithms in TNC platforms for tasks such as passenger-driver matching, pricing, and operational optimization poses the risk of perpetuating biases and discrimination if not designed and implemented with equity in mind.

One of the major operational problems in the ride-hailing system is the vehicle rebalancing problem, where vacant vehicles are redistributed proactively to under-supplied areas to reduce the discrepancy between supply and demand [144, 164, 121, 172, 67, 71]. Nevertheless, if ride-hailing platforms focus solely on maximizing profits or efficiency without considering equity concerns, their operational approach can trigger a detrimental feedback loop within the system. Figure 4-1 demonstrates how ride-hailing platforms' vehicle rebalancing decisions tend to allocate more vehicles to high-demand areas and fewer to low-demand ones. This disparity results in enhanced service levels in high-demand regions, potentially increasing future demand there. Conversely, areas with less demand experience a scarcity of vehicles, leading to diminished service quality and a loss of passengers over time. Such dynamics exacerbate the demand imbalance, creating a harmful feedback loop that influences future vehicle distribution decisions.

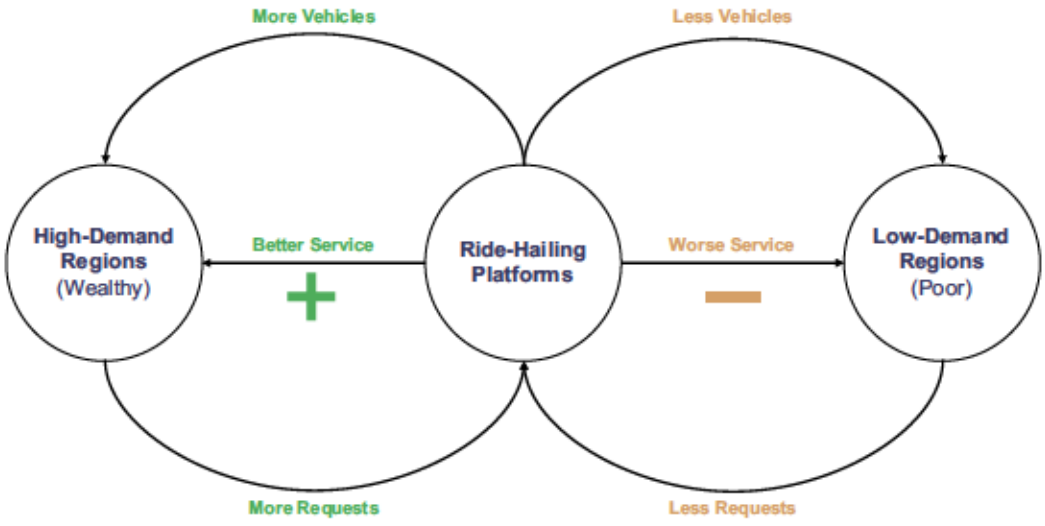
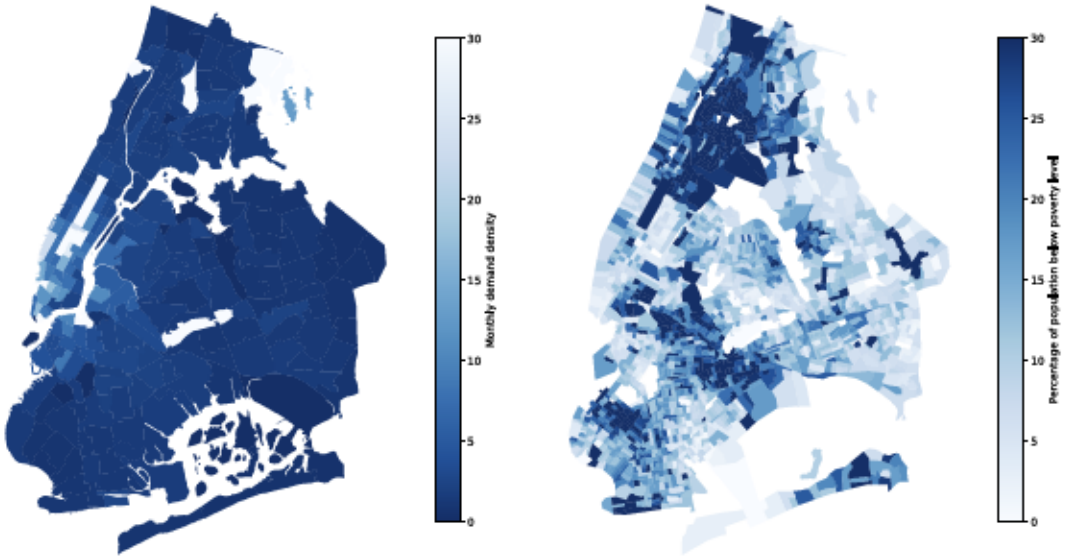


Figure 4-1: Illustration of detrimental feedback loop in vehicle rebalancing operations by ride-hailing platforms.

Meanwhile, the underserved communities typically have low ride-hailing demand

density, therefore, they are particularly disadvantaged by the detrimental feedback loop generated by the vehicle rebalancing operations. Figure 4-2 displays the spatial distribution of ride-hailing demand density and poverty levels in the city. Darker shades represent areas with a higher population living below the poverty line and lower ride-hailing demand density. Predictably, regions with a larger impoverished population exhibit reduced ride-hailing demand, such as upper Manhattan, the Bronx, and lower Brooklyn. If reducing service disparity isn't incorporated into the design of vehicle rebalancing algorithms, these communities will continue to be adversely affected by the prevailing systems.



(a) NYC Ride-Hailing Demand Density Map (by taxi zones)

(b) NYC Poverty Map (by census tracts)

Figure 4-2: Spatial distributions of ride-hailing demand and poverty in New York City (NYC).

Rebalancing vehicles require the knowledge of future demand distributions. There are two critical component in the ride-hailing vehicle rebalancing operation: upstream demand forecasting and downstream vehicle repositioning. In this chapter, we reduce disparity concerns in both components. While not directly addressing fairness issues within the ride-hailing system, understanding disparity is a foundation for understanding fairness. We outline two levels of disparity aimed at mitigation:

1. **Error disparity in Demand Prediction Algorithms:** we aim to minimize disparity in prediction errors across regions, irrespective of spatial locations and historical demand levels, involved in vehicle rebalancing operations.

2. **Service disparity in Ride-Hailing Operations:** our goal is to reduce disparities in quality of services accessed by customers, indicated by waiting times, regardless of the regions from which trips originate.

Disparity considerations are not prevailing in either component. For the upstream demand forecasting, data-driven approaches, including traditional time series analysis [193, 104] and modern machine learning models [175, 107, 92, 66, 103, 166], have been utilized in generating reliable predictions. However, many studies focus solely on prediction accuracy, disregarding the social consequences of travel demand forecasting. Very few works have been addressed error disparity issues in the demand prediction [177, 178, 199]. For the downstream ride-hailing operations, researchers have attempted to enhancing particular facets of disparity. For instance, disparity concerning driver earnings [149, 29, 146, 34, 135] and disparity regarding rider pricing [128, 113, 93, 34, 135]. To the authors' knowledge, no studies have yet tackled disparity issues in vehicle rebalancing operations, particularly diminishing disparity in both demand forecasting and vehicle distribution simultaneously.

This chapter introduces a disparity-reducing vehicle rebalancing framework, taking into account disparity issues in both demand prediction and vehicle repositioning. The framework aims to tackle the two levels of disparity outlined previously. The key contributions of this chapter are summarized as follows:

- This work represents a novel contribution to the realm of ride-hailing vehicle rebalancing by aiming to mitigate disparity in both upstream demand prediction and downstream vehicle rebalancing operations.

- To decrease error disparity in upstream demand prediction, a Socio-Aware Spatial-Temporal Graph Convolutional Network (SA-STGCN) that builds upon the STGCN model [187] is proposed. This new framework incorporates a socio-

enriched adjacency matrix and a bias-mitigation regularization method to minimize prediction discrepancies across regions.

- Building on the Matching-Integrated Vehicle Rebalancing (MIVR) algorithm [67], a disparity-included weighted objective function is proposed to reduce service disparity in the process of vehicle rebalancing. The proposed function is informed by the socio-enriched adjacency matrix from the SA-STGCN model and is designed to grant more service accessibility to underserved communities.

- Several metrics are implemented to evaluate the prediction accuracy and error disparity of the upstream demand prediction module. Real-world ride-hailing data is utilized to evaluate the downstream vehicle rebalancing outcomes. The proposed framework diminishes service disparity—indicated by a more uniform distribution of wait times across regions—by 6.5% while not diminishing the system efficiency—measured by customer wait times.

The remainder of this chapter is organized as follows. Section 4.2 provides a comprehensive review of the existing literature on ride-hailing vehicle rebalancing problems, demand prediction approaches, and Equity issues within the ride-hailing system. In Section 4.3, the disparity-reducing vehicle rebalancing framework is proposed. Results of numerical experiments are presented in Section 4.4. Section 4.5 discussed the policy and practical implications from the results. Finally, Section 4.6 summarizes the chapter and outlines future research directions.

4.2 Literature Review

4.2.1 Ride-Hailing System and Vehicle Rebalancing Problem

The field of ride-hailing systems is extensively studied, as outlined in Wang and Yang [165]. This body of research covers a range of topics including the structure of the market [70, 190, 167], analyses of labor supply [68], operations of matching drivers with passengers [3], strategies for vehicle rebalancing [144, 164, 121, 172, 67, 71],

designs of surge pricing [39], among other areas. A key operational strategy in these systems is the rebalancing of vacant vehicles. This process is vital in complementing the primary function of matching customers with available drivers.

A major challenge in ride-hailing systems is the spatial mismatch between where demand arises and where vehicles are available. To address this, there's a need for relocating idle vehicles to areas where future demand is anticipated to exceed the current supply of vehicles. By adopting this proactive rebalancing approach, ride-hailing platforms can significantly reduce the distance traveled by empty vehicles, also known as 'empty miles', and concurrently decrease the waiting times for customers. This strategy is essential for optimizing operational efficiency and enhancing user satisfaction in ride-hailing services.

The vehicle rebalancing problem is initially studied by Godfrey and Powell [59, 60] under the context of dynamic fleet management. Over the past decade, with the rapid growth of Mobility-on-Demand (MoD) and ride-hailing systems, more attention has been devoted to solving this challenge [172, 88, 30, 121]. Wen et al. [172] used reinforcement learning to tackle vehicle rebalancing in a shared MoD system, achieving a 14% fleet size reduction in a London simulation. Braverman et al. [30] designed a fluid-based optimization model for ride-hailing vehicle management, resulting in improved passenger service compared to benchmark models. Miao et al. [121] introduced a data-driven vehicle rebalancing model, minimizing the worst-case rebalancing cost using real-world NYC taxi data, achieving an average 30% reduction in idle driving distance. Guo et al. [67] proposed the Matching-Integrated Vehicle Rebalancing (MIVR) model for solving the vehicle rebalancing problem considering future iterations and incorporated the demand uncertainty with the Robust Optimization (RO) techniques. Guo et al. [71] expanded on this concept by exploring multiple data-driven strategies to handle demand uncertainty within the framework of the MIVR model.

Meanwhile, various studies have been focused on the control of Autonomous MoD system [188, 131, 194, 84, 155], where vehicles are dispatched in the system by the proposed control strategy. Pavone et al. [131] used a fluid-based model and linear

program for generating optimal rebalancing policy. Zhang and Pavone [194] proposed a queueing-based algorithm for AMoD rebalancing. Iglesias et al. [84] utilized LSTM neural networks in a Model Predictive Contro (MPC) algorithm for rebalancing with short-term demand forecasts. Tsao et al. [155] further introduced an MPC algorithm in the shared AMoD setting.

Despite the extensive research on vehicle rebalancing, none of the existing studies have explicitly tackled equity concerns in their proposed algorithms. This chapter aims to fill this gap by presenting a disparity-reducing vehicle rebalancing framework building upon the MIVR model [67] that systematically addresses issues related to the perceived services experienced by customers.

4.2.2 Demand Prediction in the Transportation System

Predicting the accurate travel demand of a given transportation system is crucial for efficient system operations and regulations. A great volume of research has studied diverse methods for travel demand forecasts. Traditional methods include the Historical Averages, Moving Averages, autoregressive integrated moving average (ARIMA) and its variants, and some basic machine learning models such as support vector machines (SVM) [193, 85, 108, 163]. However, these models focused more on temporal links but failed to account for spatial and relational information in the transportation network [108].

In recent years, with the rise of large machine learning models and the booming computing power, there has been a shift from using traditional statistical time series analysis to deep learning sequential networks [163]. Many studies have leveraged the convolutional neural networks (CNN) such as ResNet to capture spatial features [192, 191] and used the recurrent neural networks (RNN) such as the long-short term memory (LSTM) and Attention mechanism to learn the time series function [183, 92, 184].

On the other hand, graph neural networks (GNN) gained popularity because of their ability to capture spatial dependency and the non-Euclidean structure of the street network [87]. There are many types of graph convolutions in GNNs, including

the spectral-based graph convolutional networks (GCN) [96, 49] and spatial-based convolutional GNNs [5, 57, 76, 160]. Many architectures are proposed to solve traffic forecasting problems. For instance, Li et al. [105] proposed the Diffusion Convolutional Recurrent Neural Network (DCRNN), leveraging bidirectional graph random walk and the recurrent neural network to capture both the spatial and temporal dependencies. Yu et al. [187] proposed the Spatial-Temporal Graph Convolutional Network (STGCN), using graph and temporal convolution layers to build up the basic block of the architecture. Wu et al. [173] developed the Graph WaveNet, which utilized a self-adaptive adjacency matrix to capture hidden relations in the graph and leveraged dilated causal convolution to work with long-range sequences. Our research adopted the STGCN model as the main structure for its state-of-art performance and ease of implementation.

4.2.3 Equity in the Transportation System

Equity Definition

There is a wide range of debates regarding the definition of equity or fairness in political philosophy, computer science, and transportation. Major theories of fairness in political philosophy can be separated into four categories, which are 1) ensure equal share or proportional share for each individual, 2) ensure market equilibrium, 3) maximize total welfare, and 4) ensure subgroup welfare [102, 27]. In computer science, there are also various evaluation metrics to measure algorithmic fairness, the first set is Disparate Impact Analysis versus Disparate Treatment Analysis, where the former aims to achieve fair impact or results for the unit of comparison while the latter aims to achieve fair treatment [120, 132]. Another set of evaluation notions based on the unit of comparison includes group-based and individual-based fairness, where the former focuses on the same outcome or treatment of different groups, while the latter concentrates on the same outcome or treatment of individuals [41, 120]. In transportation, *horizontal equity* and *vertical equity* are often alluded to. Horizontal equity refers to the goal of similar people receiving similar treatment, while vertical

equity refers to the goal of disadvantaged people being taken care of [109, 177]. This fairness notion aligns with the many existing concepts of fairness and equity from the fields of political philosophy and computer science.

Equity Research in the Transportation System

Research in equity in the transportation system has been emerging in recent years, especially in public transit planning, where the fair sharing of public resources is important. Previously, a lot of the focus was put on the equity analysis of the existing or forthcoming systems. Bills and Walker [26] measured and compared consumer surplus distributions for different population segments across planning scenarios in the Bay area. Cascetta et al. [38] estimated the horizontal equity of the travel time accessibility by calculating the Gini index for Italian high-speed railways. On the other hand, Zheng et al. [200] demonstrated the inequity in the Deep Neural Network and the Discrete Choice Model, unlike preceding research, they also provided a disparity mitigation solution through an absolute correlation regularization method.

In the realm of ride-hailing, researchers have been focused on improving algorithmic fairness, driver equity, and rider equity. Yan and Howe [177] addressed the challenge of socio-economic inequity perpetuated by new mobility services like car-sharing, bike-sharing, and ride-hailing. Yan and Howe [178] discussed a novel unsupervised learning architecture, named EquiTensors, for integrating heterogeneous spatio-temporal urban data to counteract bias and produce fair and reusable representations. Both Zheng et al. [199] and Zhang et al. [197] have looked at the demand prediction fairness in ride-hailing, however, they only leveraged group fairness as a fairness definition and focused on the prediction result without considering the downstream impact.

In exploring driver-side equity in ride-hailing systems, various studies have made significant contributions. Bokányi and Hannák [29] employed a city simulation to highlight critical issues of wage inequality. Sühr et al. [149] analyzed drivers' income equity over time, suggesting a framework that enhances the utility for both customers and drivers in a dual-sided market. Sun et al. [146] tackled the dual challenge of

efficiency and equity, introducing a multi-agent reinforcement learning framework that assists drivers in making equitable income decisions through order selection and repositioning. Raman et al. [135] offered two strategies within a Markov decision process to mitigate inequality in ride-pooling services, aiming to balance profitability and equity for both passengers and drivers.

As for rider-side equity in transportation systems, Qian et al. [134] developed a novel route recommendation system that considers shared roads while maintaining cost-effectiveness for each customer. Cao et al. [34] aimed at enhancing both efficiency and equity in ride-sharing by optimizing routes for ride-hailing services. Ke and Qian [93] introduced the concept of Fleet-Optimal Behavior with Service Constraint (FOSC), focusing on striking a balance between reducing total fleet costs and ensuring fair travel times for riders. Nanda et al. [128] proposed a driver-customer matching algorithm that addresses the challenge of maximizing profit while maintaining equity in rider matching rates. Lu et al. [113] suggested a decentralized smart price auditing system using block-chain technology and smart contracts to ensure price equity and prevent discrimination among riders in ride-hailing services.

The term equity has been misused in many existing transportation literature without a proper definition. In this chapter, we are addressing the disparity issues, which is one important component of equity issues. Meanwhile, understanding disparity is a foundation for understanding equity in the system. We present a new vehicle rebalancing framework aimed at simultaneously reducing error disparity in the demand prediction and mitigating service disparity in downstream vehicle rebalancing, aspects not jointly considered in existing research. This innovative approach ensures equitable and efficient operations, serving as foundations for studying equity in vehicle rebalancing.

4.3 Methodology

In this section, we will introduce the disparity-reducing vehicle rebalancing framework, consisting of a Socio-Aware STGCN (SA-STGCN) demand prediction compo-

nent and an equity-enhanced MIVR component for downstream vehicle rebalancing operations. The SA-STGCN method aims to diminish prediction error disparity in upstream demand forecasting, while the equity-enhanced MIVR model seeks to augment the number of rebalanced vehicles in underserved areas, thus lessening service accessibility disparities across regions. Although these components appear to pursue distinct disparity objectives, subsequent numerical experiments reveal that minimizing the disparity in demand prediction errors directly contributes to reduced service provision disparities in the vehicle rebalancing phase. Integrating both components effectively addresses customer-side equity issues, leading to a synergy where the combined effect is greater than the sum of their individual impacts, embodying a "$1 + 1 > 2$" scenario.

4.3.1 Disparity-Reducing Demand Prediction Framework

The original STGCN (Spatio-Temporal Graph Convolutional Network) framework, while effective in numerical prediction accuracy, primarily concentrated on leveraging spatial locality information and temporal autocorrelations. This approach, though beneficial in certain contexts, inadvertently leads to a regional imbalance in the model outputs. Such imbalances can propagate inequity in downstream applications, particularly in tasks that are sensitive to regional disparities. Recognizing this limitation, our research proposes the innovative SA-STGCN (Social Aware STGCN) framework, complemented by a disparity-reducing methodology specifically tailored for vehicle rebalancing optimization in ride-hailing operations. The conceptual architecture of this approach is detailed in Figure 4-3:

As depicted in Figure 4-3, our modifications to the STGCN framework are multifaceted and designed to address the disparity issue at its core. These modifications include:

1. Integrating a demographic matrix (e.g., representing the black and low-income population) with the original adjacency matrix to infuse social context into the data.

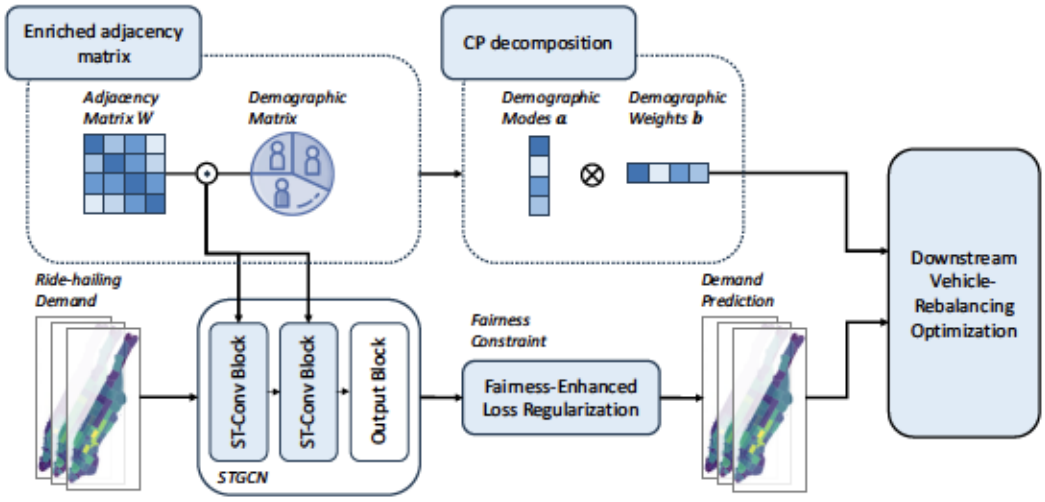


Figure 4-3: Detailed illustrations of the SA-STGCN framework and its integration with the vehicle rebalancing optimization task.

2. Implementing a disparity-reduced loss regularization approach to penalize demand overestimations and reduce output disparities.

3. Employing matrix decomposition on the augmented adjacency matrix to derive demographic weights, which serve as adjustment factors in the downstream rebalancing task.

In the following sections, we delve deeper into each of these components, discussing their implementation and the specific methodologies employed. We aim to provide a comprehensive understanding of how each aspect contributes to the overall effectiveness and equity of the SA-STGCN framework.

Socio-Aware Spatial-Temporal Graph Convolutional Network

This module starts by formulating the demand prediction problem. Given the ride-hailing system, we can regard it as a directed weighted graph $G = (V, E, W)$, where $|V| = n$ is the set graph vertices representing sub-regions (zones), the E represents the set of road networks between each pair of vertices as edges, and W represents the weights of each edge in the form of weighted adjacency matrix, calculated by:

$$W_{ij} = \exp\left(-\frac{\bar{d}_{ij}^2}{\sigma_d^2}\right), \tag{4.1}$$

where W_{ij} is the edge weight between graph vertices v_i and v_j, $\bar{d}_{ij}$ is the Euclidean distance between the centroids of vertices v_i and v_j, σ_d is the standard deviation of the set of distances begins at each vertex v_i [49].

To capture the spatial demographic features that exist in the network, this study establishes a socio-demographically enriched adjacency matrix by incorporating an additional demographic matrix. In this study, we leverage census data and focus on the ratio of the minority race population and the population in poverty as the variables of interest. The first step is to construct a demographic matrix. Consider the demographic variables in preceding N years in each graph vertice v_i as a vector z_i where $|z_i| = 2N$ since we only focused on two variables here. Then we can construct the relationship between the demographic features of each pair of vertices z_i and z_j with a correlation matrix.

$$Corr_{ij} = \frac{\sum_{m=1}^{2N}(z_i^m - \bar{z}_i)(z_j^m - \bar{z}_j)}{\sqrt{\sum_{m=1}^{2N}(z_i^m - \bar{z}_i)^2}\sqrt{\sum_{m=1}^{2N}(z_j - \bar{z}_j)^2}}, \tag{4.2}$$

where z_i^m and z_j^m are the individual elements in z_i and z_j. When $z_i = z_j = 0$, the corresponding value in the matrix will be filled as zero.

We derived a new socio-demographically adjacency matrix that incorporates this demographic matrix into the original adjacency matrix, specifically, we use the Hadamard product of the original adjacency matrix and the demographic correlation matrix as the new adjacency matrix to feed into the prediction model described below. To introduce sparsity, the element value of the new matrix is only kept when it is greater than or equal to 0.1, otherwise, it is converted to 0.

$$W_{ij}^* = \begin{cases} W_{ij} \circ Corr_{ij}, & \text{if } W_{ij} \circ Corr_{ij} \geq 0.1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

In terms of the representation of the ride-hailing demand, let $r_i^k \in \mathbb{N}$ be the demand at each time period k that originates at vertex v_i (or sub-region i). Given the demand observations from previous M time periods at each vertex $[r^{k-M}, ..., r^{k-1}]$, we want to forecast the upcoming demand $\hat{r}^k$.

The study refers to the Spatial-Temporal Graph Convolutional Network (STGCN) proposed by [187] as the main model structure. STGCN has been widely used in traffic forecasting for its excellent ability to capture temporal and spatial features compared to traditional CNN and time-series models because it utilizes both graph and gated temporal convolutions. The model we adopted consists of two Spatial Temporal Convolutional (ST-Conv) blocks followed by a fully connected layer. Each ST-Conv block comprises two temporal gated convolution layers and a spatial graph convolution layer in between. To reduce the computing cost of graph convolution, we adopted the Chebyshev Polynomials Approximation and the first-Order Approximation strategies [49, 96, 187].

Disparity Reduction

In this study, we addressed the prediction error disparity issue with regularization inspired by the work of [199] and [24]. The set of demand predictions $\hat{\boldsymbol{r}}$ and all the trainable parameters in the model $\boldsymbol{W_\theta}$ gives the cost function. In general, the cost function consists of two parts, each optimizes the accuracy and the disparity of the model performance:

$$J_{total} = J_{accuracy} + param * J_{disparity}, \tag{4.4}$$

where $param$ is a hyperparameter that adjusts the relative weight of the disparity optimization term. The detailed formulation of training objective can be formulated as

$$J(\hat{\boldsymbol{r}}, \boldsymbol{W_\theta}) = \sum_{t=1}^{\kappa}\sum_{i=1}^{n}(r_i^k - \hat{r}_i^k)^2 + \lambda \sum_{k=1}^{\kappa} \frac{\sum_{i=1}^{n}(SAPE_i^k - \bar{SAPE}^k)^2}{n-1} + \gamma \sum_{k=1}^{\kappa}\sum_{i=1}^{n} max(0, \hat{r}_i^k - r_i^k) \tag{4.5}$$

where λ and γ are weight parameters for the case of limiting error distribution and penalizing overestimation, respectively. κ is the number of time periods in each training iteration.

Limit error distribution. To reduce disparity for the whole population and regulate the error distribution variance, we add the variance of the symmetric absolute percentage error (SAPE) to the cost function. The SAPE is a bounded error that is more robust to regions with zero or near-zero demand compared to metrics like Mean Absolute Percentage Error (MAPE). First, $SAPE$ at each vertex in a given time period is defined as:

$$SAPE_i^k = \begin{cases} \frac{|r_i^k - \hat{r}_i^k|}{|r_i^k| + |\hat{r}_i^k|}, & |r_i^k| + |\hat{r}_i^k| \neq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{4.6}$$

where r_i^k and $\hat{r}_i^k$ are the original demand and the predicted demand in vertex v_i at time k. Measuring the percentage error in this way is useful in the case of ride-hailing as the prediction and original demand are sometimes equal or close to zero, making the traditional percentage error term fraction undefined. In the case where both the prediction and original demand are zero, we set the error term as zero. Then, we can calculate the variance of the $SAPE$ for all vertices in a given time period as regularization term to limit error distribution shown in Equation (4.5).

Penalize overestimation. As Guo et al. [71] pointed out in their study of ride-hailing vehicle rebalancing operations, when the model prediction is not accurate enough, underestimation is preferred compared to overestimation because this avoids unnecessary vehicle relocation, resulting in a better rebalancing result. Moreover, the resultant smaller overestimation will help to reduce the total variance of the prediction error, thus providing a fairer outcome. To reduce overestimation, we add a penalizing term to the cost function Equation (4.5) with the parameter γ. In this way, whenever the prediction overestimates the demand, that is, when $\hat{r}_i^k > r_i^k$, the regularization term will penalize it to be closer to zero.

Weighted Adjacency Matrix Decomposition

The first two components of the SA-STGCN framework focus on reducing error disparity in demand prediction. This section introduces a matrix decomposition method

to create equity weights for sub-regions, indicating the importance of repositioning idle vehicles to each sub-region. This vector is then applied in the subsequent vehicle rebalancing task, aiming to enhance equity in the delivery of ride-hailing services.

In order to obtain the equity weights of different regions to guide the downstream rebalancing task, we decomposed our self-designed adjacency matrix to obtain a set of equity weights that consider both the spatial dependencies as well as the disparity of socio-demographic features in the city. Therefore, two consecutive steps are conducted through the process: (1) enrich the weighted adjacency matrix W with sociodemographic information, which was explained by Equation 4.2 and 4.3; (2) decompose the adjacency matrix to obtain the spectrum and use it as the output weights, which will be explained in this section.

The decomposition of the adjacency matrix is usually applied in the graph spectrum analysis because it provides a powerful means to reveal and quantify the latent structural properties of graphs, such as node centrality, community structure, and connectivity patterns [47, 63]. In our context, we use Canonical polyadic (CP) decomposition in our experiment. The CP decomposition, also known as PARAFAC or tensor factorization, is a multilinear algebraic framework that generalizes the matrix singular value decomposition to higher-order tensors [79, 77, 168]. For our adjacency matrix W^*, viewed as a two-dimensional tensor, CP decomposition factorizes W^* into a sum of component rank-one tensors, providing insights into multi-way interactions. Mathematically, if W^* is a tensor of order two, the CP decomposition is represented as:

$$W^* \approx \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \otimes \mathbf{b}_r, \tag{4.7}$$

where $\otimes$ denotes the outer product, R is the rank of the decomposition, λ_r are the weights indicating the importance of each component, and $\mathbf{a}_r$ and $\mathbf{b}_r$ are the corresponding factor vectors representing the demographic modes and weights of the decomposed W^*. This decomposition facilitates the distillation of complex network data into a form that accentuates inherent spatial and sociodemographic relation-

ships, allowing for compressing W^* to capture nuanced regional disparities in a comprehensible manner. In this study, we define $R = 1$ to produce a pair of vectors that encapsulate the maximum amount of information from the adjacency matrix W^*.

The equity weights for the downstream optimization problem will be derived from the vector $\boldsymbol{b}$ by following these procedures: i) Normalize $\boldsymbol{b}$ using a min-max scaling method, ii) decrease the magnitude of each element by multiplying 0.1, and iii) obtain the final equity weights for each sub-region by subtracting each scaled value from 1. Specifically, for each sub-region i, the normalized value $\bar{b_i}$ is computed as

$$\bar{b_i} = \frac{b_i - \min(\boldsymbol{b})}{\max(\boldsymbol{b}) - \min(\boldsymbol{b})}.$$

Subsequently, the equity weight for each sub-region i is determined as

$$\omega_i = 1 - 0.1 \times \bar{b_i}.$$

Our enriched adjacency equation (Equation 4.3) indicates that sub-regions nearer to other sub-regions, with lesser poverty rates and smaller minority race populations, are likely to have higher b_i values. In contrast, city outskirts, typically marked by higher poverty and more significant minority race populations, are distinct from urban centers. The aim of the equity weights is to prioritize these underserved peripheral areas, which often face greater socio-economic challenges. These weights are calculated following the earlier mentioned steps, ensuring that sub-regions with higher b_i values correspondingly have lower ω_i values, thus aligning with our objective of equitable weights distribution.

4.3.2 Equity-Enhanced Vehicle Rebalancing Component

Tackling error disparity in the demand prediction module doesn't directly mitigate disparity issues related to the services received by customers, who are the primary indicators when evaluating ride-hailing operations. Therefore, this study provides a equity-enhanced Matching-Integrated Vehicle Rebalancing (MIVR) model, building

upon the MIVR model proposed by Guo et al. [67].

The operational period is divided into Ω identical time intervals, each denoted by an index $k = 1, 2, ..., \Omega$, and lasting Δ time units. Furthermore, the study area is partitioned into n sub-regions (zones), with each sub-region i exhibiting an estimated demand $\hat{r}_i^k \geq 0$ at time k. To formulate the model, we introduce two sets: i) the set of sub-regions denoted as $N = 1, 2, ..., n$, and ii) the set of time intervals represented by $K = 1, 2, ..., \kappa$.

The MIVR model is solved in a rolling-horizon manner: when solving the MIVR model at the start of time interval k, it considers the demand during time interval k as well as the demand for κ future time intervals; however, only the vehicle rebalancing decisions for the current time interval k are put into action; following this, vehicle locations are observed and updated as inputs to the MIVR model for the subsequent time interval.

The MIVR model consists of two components: matching and rebalancing. The matching component uses decision variables $y_{ij}^k \in \mathbb{R}^+$ to represent the number of customers matched between sub-regions i and j at time k. The rebalancing component uses decision variables $x_{ij}^k \in \mathbb{R}^+$ to denote the number of idle vehicles rebalanced from sub-region i to sub-region j at time k. Travel distance d_{ij}^k between sub-regions i and j at time k is approximated by the distance between their centroids.

The equity considerations are incorporated by introducing the equity weights $\boldsymbol{\omega} \in (\mathbb{R}^+)^n$, consisting of a weight parameter ω_i for each region i. The equity weights $\boldsymbol{\omega}$ are decomposed from the enriched adjacency matrix from the SA-STGCN framework. The weighted objective function of the MIVR model can then be formulated as

$$\min_{\boldsymbol{x},\boldsymbol{y}} \quad c(\boldsymbol{x}, \boldsymbol{y}; \hat{\boldsymbol{r}}) = \sum_{k=1}^{\kappa} \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij}^k d_{ij}^k + \alpha \cdot \sum_{k=1}^{\kappa} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i y_{ij}^k d_{ji}^k + \beta \cdot \sum_{k=1}^{\kappa} \sum_{i=1}^{n} \omega_i \left(\hat{r}_i^k - \sum_{j=1}^{n} y_{ij}^k \right), \tag{4.8}$$

where $\hat{\boldsymbol{r}}$ stands for the vector of estimated demand, α and β are weight parameters indicating weights for matching distance and penalty for unsatisfied demand, respec-

tively. The objective function $c(\boldsymbol{x}, \boldsymbol{y}; \hat{\boldsymbol{r}})$ defines a weighted generalized cost for the vehicle rebalancing problem with the consideration of the matching component. The equity weights $\boldsymbol{\omega}$ play a crucial role in rebalancing decisions by imposing extra costs and penalties on specific regions, thereby guiding vehicles towards these areas. The constraints used in this optimization problem can be found in Guo et al. [67].

With a given demand prediction $\hat{\boldsymbol{r}}$, problem 4.8 provides a vehicle rebalancing strategy for the upcoming decision time interval. The demand prediction $\hat{\boldsymbol{r}}$ is provided by the upstream SA-STGCN framework, which balances the prediction accuracy and disparity metrics.

4.4 Experimental Results

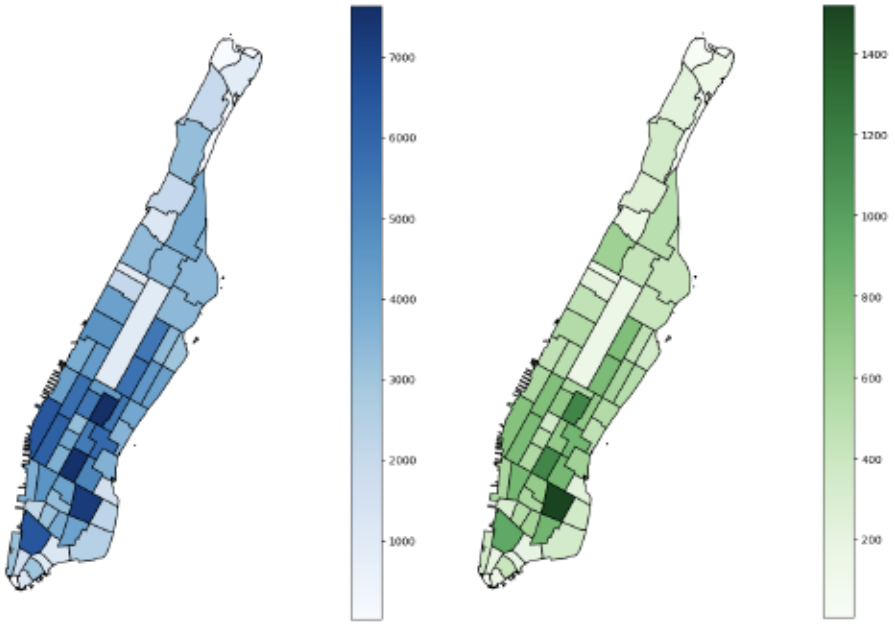
4.4.1 Data

This research utilizes For-Hire Vehicle trip records obtained from the NYC Taxi and Limousine Commission [130]. The experiments are conducted using "taxi zones", which are well-defined regions within the high-volume ride-hailing trip dataset, comprising a total of 63 zones on Manhattan Island.

To best emulate typical travel behaviors, five months of demand data from February to June 2019 are utilized, encompassing a total of 47,009,841 trips within the study period. The demand data is aggregated into 5-minute time intervals ($\Delta = 300$ seconds) for analysis and evaluation. Figure 4-4 illustrates the average and standard deviation of daily regional demand in Manhattan, revealing significant demand volatility in the lower Manhattan area.

In the demand prediction module, these taxi zones are employed as vertices in the graph G. The dataset is divided into three subsets: the training set, validation set, and test set, which account for 70%, 15%, and 15% of the overall dataset, respectively. To get the demand prediction $\hat{r}^k$ for a given taxi zone at time k, the module utilizes the previous 12 observations of demand $[r^{k-12}, r^{k-11}, ..., r^{k-1}]$.

To assess the performance of both the demand prediction and vehicle rebalanc-



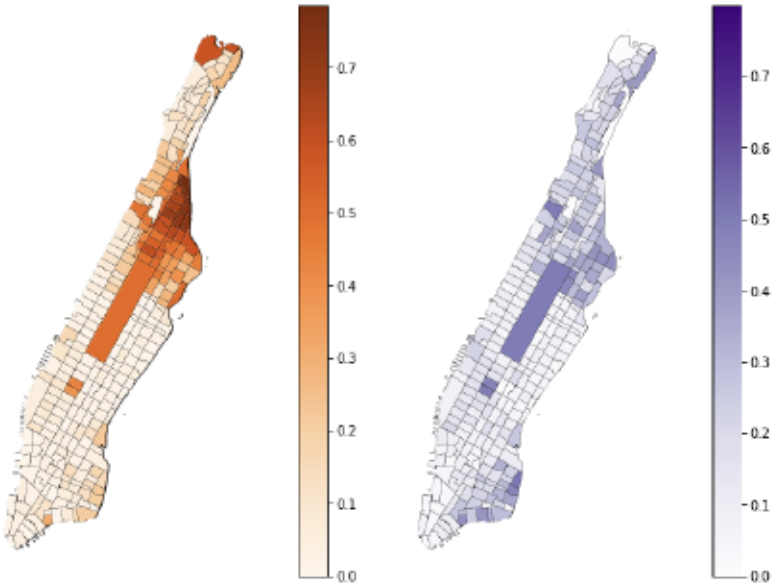
(a) Mean (b) Standard deviation

Figure 4-4: Daily demand by zone (trips) in Manhattan.

ing modules, real-world demand data from June 26, 2019, is used as the evaluation set. This one-day demand data is excluded from the training stage to ensure a fair assessment of the modules' capabilities.

To incorporate the socio-demographic information into both the upstream prediction and downstream operation, we used the American Community Survey (ACS) census data. Specifically, we focused on two variables, the racial and poverty compositions of a given region to represent the population vulnerable to insufficient ride-hailing service provision. The poverty level is pre-determined by the ACS dataset. We also selected five years of data from 2015-2019 to capture a fuller picture of the demographic patterns in the city. Figure 4-5 illustrates the spatial distributions of the two variables in 2019. There is apparent spatial clustering concerning how the marginalized population resides in the city. The original census data is at the census tract level, and to aggregate it to the taxi zone level for future estimation, we calculated the centroid of each census tract, assigned them to the corresponding taxi zone, and summed up the population of each demographic variable in each taxi zone. To calculate the ratio of each demographic, we divided the population of the target

demographic population by the total population for each taxi zone.



(a) Black Population Ratio (b) Poverty Population Ratio

Figure 4-5: Demographic Variables Distribution in Manhattan in 2019.

4.4.2 Performance Evaluation

Upstream Prediction Evaluation

The study adopted two sets of metrics to evaluate the accuracy and disparity of the models respectively. To measure accuracy, we measured the error magnitude, relative error percentage, and error direction. Two commonly used metrics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), are adopted to evaluate error magnitude. They are defined as follows:

$$MAE = \frac{1}{\kappa \times n} \sum_{k=1}^{\kappa} \sum_{i=1}^{n} |r_i^k - \hat{r}_i^k|, \tag{4.9}$$

$$RMSE = \sqrt{\frac{1}{\kappa \times n} \sum_{k=1}^{\kappa} \sum_{i=1}^{n} (r_i^k - \hat{r}_i^k)^2}. \tag{4.10}$$

In addition, to measure the relative error percentage, we adopted the Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{\kappa \times n} \sum_{k=1}^{\kappa} \sum_{i=1}^{n} \frac{|r_i^k - \hat{r}_i^k|}{r_i^{k'}}, \text{ where } r_i^{k'} = min(r_i^k, 0.1). \tag{4.11}$$

Note that the denominator $r_i^{k'}$ is adjusted to ensure the fraction is defined. Lastly, to measure the overall prediction error direction compared to the original demand, we adopted the Mean Error (ME):

$$ME = \frac{1}{\kappa \times n} \sum_{k=1}^{\kappa} \sum_{i=1}^{n} (r_i^k - \hat{r}_i^k), \tag{4.12}$$

where a positive ME value denotes the model underestimating the demand in general, and vice versa.

On the other hand, to evaluate the disparity of the model predictions, we utilized two metrics. The first metric is the Mean-Variance of the Percentage Error (MVPE). This metric originates from the goal to ensure equal error distribution for all predictions. Given the percentage error in vertex i at time t, $PE_i^k = \frac{r_i^k - \hat{r}_i^k}{r_i^k}$, the MVPE is defined as:

$$MVPE = \frac{1}{\kappa} \sum_{k=1}^{\kappa} \frac{\sum_{i=1}^{n} (PE_i^k - \bar{PE^k})^2}{n-1} \tag{4.13}$$

where $\bar{PE^k}$ is the mean of the percentage errors in a given period k.

The second disparity metric refers to the Generalized Entropy Index proposed by [143]. It is a metric originated from economics that measures inequality, or in the point of information theory, it can be interpreted as the *redundancy* in data. In our case, it measures the spread between individual prediction and the average prediction error. A smaller value represents a more even distribution between errors and their mean. We adjusted the metric to evaluate the distribution of the percentage errors in our research setting:

$$f_i^k = PE_i^k + m, \text{ where } m = \max(0, \max(-PE_i^k)), \tag{4.14}$$

$$GEI = \frac{1}{\kappa}\sum_{k=1}^{\kappa}[\frac{1}{n\eta(\eta-1)}\sum_{i=1}^{n}[(\frac{f_i^k}{\bar{f}^k})^{\eta}-1]]. \tag{4.15}$$

Specifically, a temporal GEI for each period is first calculated and then we take the average of them across all periods. $\bar{f}^k$ is the mean of f_i^k in a given period k. η is a constant that regulates how much attention is put on the larger percentage of errors, and here we set $\eta = 2$.

Downstream Vehicle Rebalancing Evaluation

For the downstream performance evaluation of the vehicle rebalancing module, we adopt the ride-hailing simulator utilized in Guo et al. [71]. For the simulation and evaluation of various approaches, demand data from June 26, 2019, is employed. The simulator is configured with a fleet size of 2000 vehicles. With this setup, vehicle and demand locations are initialized. At the start of the simulation, all vehicles are made available and evenly distributed across the taxi zones.

Since demand origins and destinations are at the sub-regional level, random road nodes within each sub-region are assigned as origins and destinations for customers in that particular sub-region. The simulator comprises two components: a matching engine, which is solved every 30 seconds, and a vehicle rebalancing engine, which is solved every 300 seconds. In the vehicle rebalancing model (4.8), the parameters are set as $\alpha = 1$ and $\beta = 10^2$. The vehicle rebalancing problem takes into account $\kappa = 6$ time intervals ahead for optimization.

The simulation yields several key metrics for evaluating different rebalancing models, including:

1. Average customer waiting time
2. Standard deviation of customer waiting time across zones
3. Customer unsatisfaction rate
4. Average non-occupied VMT
5. Average number of rebalancing trips

These metrics are essential in assessing the performance and effectiveness of various disparity-reducing rebalancing approaches. The variation in customer waiting times, as measured by standard deviation, reflects the disparity of services provided to customers. Meanwhile, other metrics are indicative of the system's overall efficiency.

4.4.3 Upstream Demand Prediction Results

In this section, we discuss the effect of the SA-STGCN framework on the performance of the upstream prediction of the ride-hailing demand.

Performance Summary

In this section, we demonstrate the demand prediction performance of the SA-STGCN model and different combinations of regularization terms compared to the Historical Average model and the pure STGCN model as the baseline models. Specifically, the Historical Average model is defined by using the historical average of the demand on the same date of the week and time interval of the day in the training set to predict the testing data. The baseline SA-STGCN model without regularization ($\boldsymbol{\lambda} = \mathbf{0}, \boldsymbol{\gamma} = \mathbf{0}$) is shortened as SA-STGCN in this section.

Table 4.1 shows the prediction performance comparison among the two benchmark models, the SA-STGCN model, and models with regularization terms implemented in terms of accuracy and disparity. The first four metrics in the table (*i.e.* MAE, RMSE, MAPE, and ME) demonstrate the accuracy performance of the models, while the last two metrics (*i.e.* MVPE and GEI) illustrate the error disparity performance. The first section of the table 4.1 illustrates how SA-STGCN compare with the baseline models, and the second, third, and fourth sections of table 4.1 refer to the performance of the models with regularization terms penalizing overestimation, limiting error distribution, and the models with both regularization terms with different weight parameters.

As shown in the first section of the table 4.1, it is evident that using the socio-demographically enriched matrix helps to improve the accuracy metrics and reduce

	MAE	RMSE	MAPE	ME	MVPE	GEI (10^{-4})
Historical Average	3.699	5.293	0.721	-0.885	8.347	5.463
STGCN	3.130	4.334	0.738	-0.046	9.054	5.948
SA-STGCN	3.128	4.352	0.693	0.204	8.002	5.230
SA-STGCN ($\gamma = 0.01$)	3.109	4.318	0.713	0.118	8.624	5.657
SA-STGCN ($\gamma = 0.03$)	3.113	4.350	0.659	0.463	7.426	4.846
SA-STGCN ($\gamma = 0.05$)	3.138	4.396	0.606	0.733	6.170	4.003
SA-STGCN ($\gamma = 0.07$)	3.173	4.458	0.577	0.988	5.637	3.648
SA-STGCN ($\gamma = 0.09$)	3.216	4.522	0.549	1.206	5.045	3.257
SA-STGCN ($\lambda = 0.5$)	3.152	4.360	0.744	-0.057	9.004	5.917
SA-STGCN ($\lambda = 1$)	3.159	4.376	0.760	0.064	9.183	6.038
SA-STGCN ($\lambda = 3$)	3.197	4.415	0.728	0.108	8.143	5.320
SA-STGCN ($\lambda = 5$)	3.255	4.481	0.739	0.240	7.973	5.198
SA-STGCN ($\lambda = 0.5, \gamma = 0.09$)	3.136	4.359	0.723	0.143	8.934	5.863
SA-STGCN ($\lambda = 1, \gamma = 0.07$)	3.228	4.489	0.548	0.974	4.751	3.060
SA-STGCN ($\lambda = 3, \gamma = 0.03$)	3.346	4.575	0.568	0.786	4.540	2.929
SA-STGCN ($\lambda = 5, \gamma = 0.07$)	3.548	4.850	0.510	1.486	3.324	2.132

Table 4.1: Model cross-comparison.

the disparity metrics compared the benchmarks. Compared to the STGCN, the SA-STGCN is able to reduce both the MVPE and GEI by 11.6% and 12.1% while not harming the accuracy metrics. It even slightly improved the MAE and the MAPE results.

By cross-comparing the evaluation metrics of the performance of different regularization terms, the proposed regularization terms mitigate overall disparity significantly while not sacrificing model accuracy too much. In fact, the accuracy metrics in all three regularization models are improved in most cases compared to benchmarks. With extremely large weight parameters such as when $\lambda = 5$ and $\gamma = 0.07$, the disparity metrics MVPE and GEI can be reduced by 63.3% and 64.2%, respectively, while the accuracy metrics such as MAE only increased by 13.4% compared to the pure STGCN. Moreover, the positive outcomes of the relative error metric ME show that adding the regularization helps the prediction to shift from overestimation to underestimation. This transformation, as we will later illustrate, contributes significantly to the enhancement of downstream vehicle rebalancing. The model penalizing overes-

timation performs the best as it improves accuracy metrics and diminishes disparity metrics at the same time.

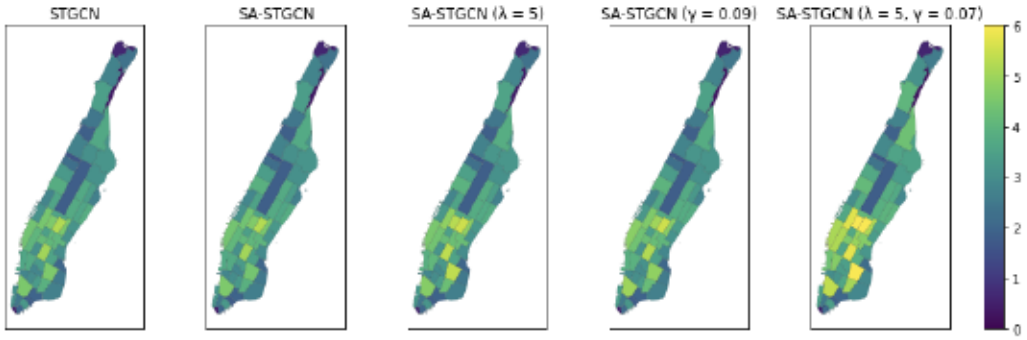
The second and third sections of the table 4.1 also serve as the sensitivity analysis of the regularization weight's effect on the model performances. In the case of overestimation penalization, there is a consistent accuracy-disparity trade-off pattern, as the weight of regularization (γ) increases, the error magnitudes monotonically increase, the percentage error decreases, and the disparity diminishes. On the other hand, when restricting the error distribution, as we increase the magnitude of the weight ($\boldsymbol{\lambda}$) of the regularization term, only the MAE, RMSE, and ME monotonically increase, but the MAPE and the two disparity metrics first increase then decrease.

Fairer Prediction in Space

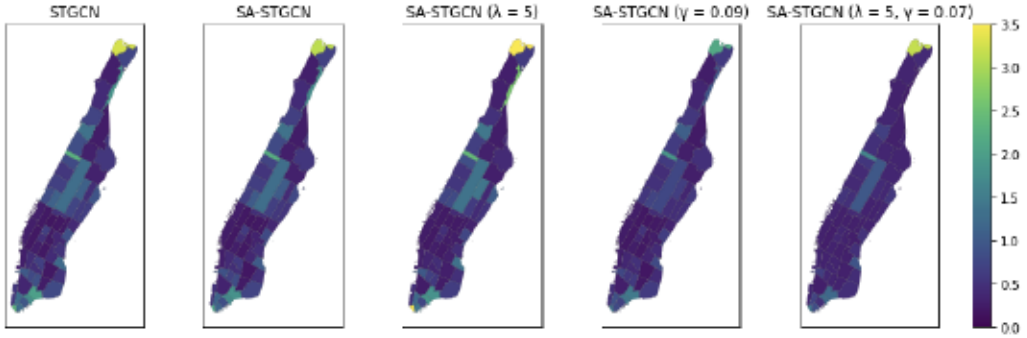
Figure 4-6 demonstrates the spatial distribution of the error metrics across the study area of the pure STGCN, SA-STGCN, and the selected models with regularization terms. In general, the prediction errors are larger in magnitude in the downtown regions, while the percentage errors have more variation in the regions with low demand, such as the northern tip of the island. When comparing various models, the MAE plots reveal similar distributions, indicating that the prediction performances are not highly sensitive to the proposed changes in the model. In contrast, the MAPE plots demonstrate that models incorporating regularization terms exhibit a more uniform error distribution across space, reflecting a fairer prediction. The ME distributions suggested the general error direction shifted from negative to positive as regularization terms are introduced, which means a transition from overestimation to underestimation. We later demonstrate that this shift could be beneficial for reducing service disparity in downstream vehicle rebalancing operations.

Fairer Prediction in Time

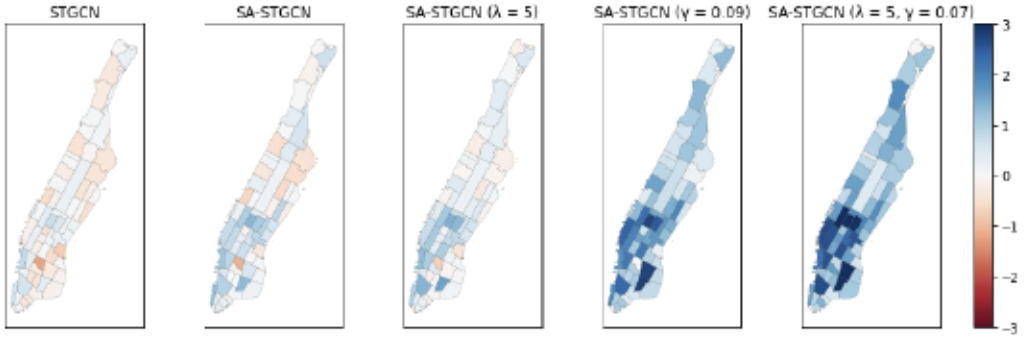
Figure 4-7 illustrates the temporal distribution of the error metrics across the time of the day. In general, the error magnitude and percentage distribution pattern follow the demand level, with smaller MAE values and larger MAPE and MVPE when



(a) MAE spatial distribution



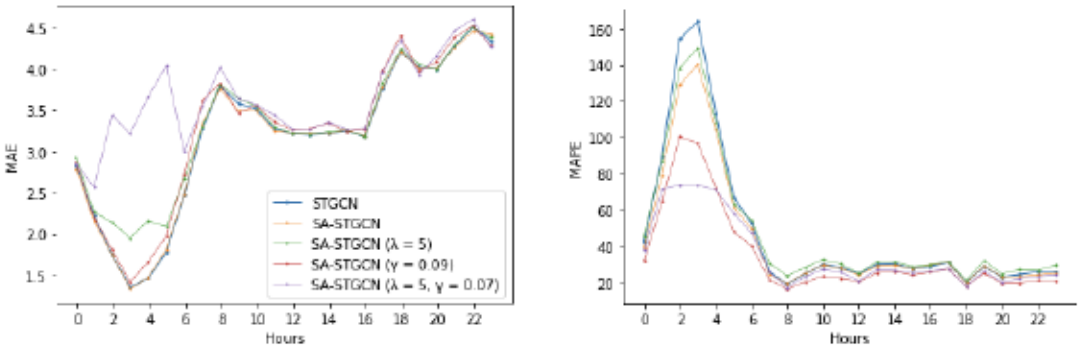
(b) MAPE spatial distribution



(c) ME spatial distribution

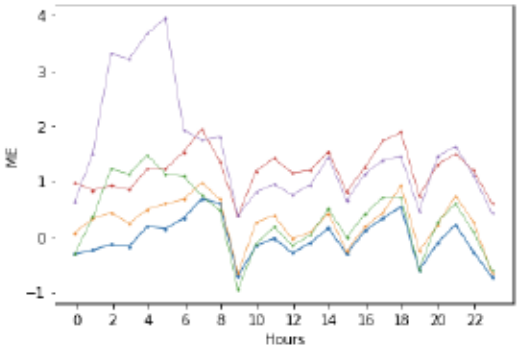
Figure 4-6: Error spatial distribution in Manhattan.

demand is low and vice versa. In contrast, the error direction displays a less regular pattern. Compared to the benchmark performance of the pure STGCN model, the proposed SA-STGCN model and the method of adding regularization terms can reduce the peak errors in all scenarios. Specifically, SA-STGCN adding both regularization terms has the most direct impact on smoothing the MAE, MAPE, and MVPE temporal distribution. The smoothing effect is most significant in the early morning time when error values are more extreme. In addition, regularization terms

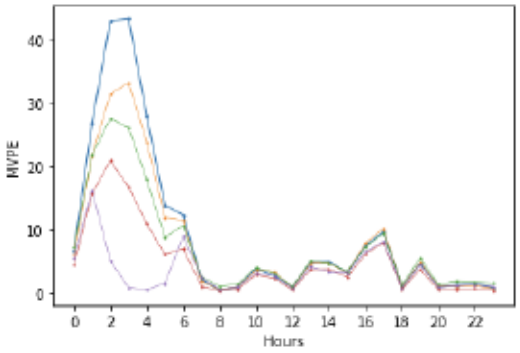


(a) MAE temporal distribution

(b) MAPE temporal distribution



(c) ME temporal distribution



(d) MVPE temporal distribution

Figure 4-7: Error temporal distribution across times of the day.

have various effects on the error directions across time. When we add both regularization terms, the ME distribution shifts to positive the most as shown by the purple line. On the other hand, adding the error restriction regularization term has the least effect on day-time ME distribution. The temporal trend of RMSE is similar to MAE and the GEI trend is similar to that of MVPE, so the figures didn't include these two metrics.

4.4.4 Downstream Vehicle Rebalancing Performances

In this section, we discuss the effectiveness of our models when applied to vehicle rebalancing tasks, utilizing a simulator based on real-world ride-hailing data.

Performance Summary

Table 4.2 presents a summary of how various demand models perform in addressing the vehicle rebalancing problem, specifically when employing the equity-enhanced MIVR model. This evaluation includes scenarios where true demand is known and those relying solely on average historical demand for future predictions, serving as benchmark models. Additionally, a standard STGCN and a baseline SA-STGCN model without additional regularization terms are also employed for benchmarking purposes.

Moreover, the evaluation also encompasses three variations of the SA-STGCN model: (i) SA-STGCN with a focus on penalizing overestimation, (ii) SA-STGCN with limited error distribution, and (iii) SA-STGCN incorporating both regularizations. Crucial customer service metrics assessed include the rate of customer dissatisfaction, average waiting times, and the variability of these waiting times across different sub-regions. To demonstrate the effectiveness of the vehicle rebalancing algorithms in enhancing operational efficiency, we examine metrics like the average Vehicle Miles Traveled (VMT) without passengers per vehicle and the average number

	Unsatisfaction Rate (%)	Wait Time Avg (seconds)	Wait Time Std (seconds)	Non-occupied VMT (miles)	Rebalancing Trip Number
True Demand	1.76	89.65	22.07	60.82	20.03
Historical Demand	1.75	90.55	22.76	60.81	19.47
Baseline STGCN	1.79	89.83	22.52	60.34	19.40
Baseline SA-STGCN	1.74	89.75	22.13	60.20	19.25
SA-STGCN ($\gamma = 0.01$)	1.77	89.88	22.08	60.49	19.50
SA-STGCN ($\gamma = 0.03$)	1.76	89.95	22.53	60.29	19.25
SA-STGCN ($\gamma = 0.05$)	1.76	89.51	21.72	60.21	19.36
SA-STGCN ($\gamma = 0.07$)	1.72	89.67	21.77	60.00	19.08
SA-STGCN ($\gamma = 0.09$)	1.73	89.58	21.78	60.00	19.15
SA-STGCN ($\lambda = 0.5$)	1.73	90.16	22.34	60.37	19.28
SA-STGCN ($\lambda = 1$)	1.77	90.03	22.17	60.34	19.28
SA-STGCN ($\lambda = 3$)	1.78	89.98	22.15	60.36	19.33
SA-STGCN ($\lambda = 5$)	1.82	89.99	21.75	60.15	19.14
SA-STGCN ($\lambda = 0.5, \gamma = 0.09$)	1.75	89.39	21.06	59.87	19.09
SA-STGCN ($\lambda = 1, \gamma = 0.07$)	1.77	89.70	21.72	60.04	19.17
SA-STGCN ($\lambda = 3, \gamma = 0.03$)	1.73	90.02	22.10	59.99	18.90
SA-STGCN ($\lambda = 5, \gamma = 0.07$)	1.74	90.44	20.62	60.11	18.70

Table 4.2: Model performance summary.

of rebalancing trips each vehicle undertakes.

Fair Prediction Leads to Fair Service

The first insight from the simulation results is the superior performance of models with fair predictions, compared to benchmark models. These enhanced models excel in two main aspects: reducing average customer waiting times and achieving a more consistent waiting time across different regions. This improvement suggests that factoring in disparity reduction in demand prediction can lead to more efficient and equitable service outcomes.

Further analysis reveals that the SA-STGCN model, when compared to the baseline STGCN, is more effective in satisfying a larger number of customers while simultaneously reducing the average waiting time. Notably, this model also succeeds in minimizing the variance in service provision across different regions, addressing the issue of regional disparities in service quality.

The benefits of the SA-STGCN model stem from two primary factors. Firstly, it outperforms the traditional STGCN model in terms of both prediction accuracy and disparity indicators. These enhanced performances translate into more effective rebalancing operations, ensuring a fairer distribution of services with less disparity in error rates across regions. Secondly, the integration of socio-demographic information within the STGCN framework leads to a more conservative approach in demand estimation. This conservatism, while acknowledging the inherent inaccuracies in demand prediction, results in fewer non-occupied VMT and a reduced number of rebalancing trips. Consequently, this conservative approach benefits the overall system operation.

Moreover, the incorporation of additional regularization terms in the SA-STGCN models could potentially enhance model performance further. This improvement would be due to the reasons discussed above, emphasizing that fairer predictions not only improve service efficiency but also contribute to equitable service distribution. Ultimately, the proposed disparity-reducing vehicle rebalancing approach fosters a win-win scenario, enhancing the efficiency and mitigating the disparity of ride-hailing services.

The Power of Demand Underestimation

The second insight from the simulation results highlights the significant role of demand underestimation in ride-hailing rebalancing operations. This underestimation, prompted by the addition of a penalization term (associated with the parameter γ) to errors in demand overestimation, results in more equitable predictions. Despite a decrease in prediction accuracy, this approach leads to improved service provision for customers and reduces disparities in services across regions.

The advantage of demand underestimation lies in its inherent conservativeness. As Guo et al. [71] suggest, in scenarios where future demand is uncertain, adopting a less aggressive approach often yields better outcomes. This conservatism allows for a more efficient distribution of limited vehicle resources, particularly important given the high costs associated with incorrect rebalancing decisions.

However, excessively penalizing demand overestimation in ride-hailing operations is not ideal. The simulation indicates that the optimal balance is achieved when γ is set to 0.05. This parameter setting results in the best average customer wait times and the fairest service distribution across different regions. It's important to note that while increasing γ does lead to fairer predictions, the impact on downstream services is not monotonic. Excessive conservatism in vehicle rebalancing can be counterproductive. Therefore, identifying an appropriate level of conservativeness is crucial for making optimal rebalancing decisions.

The Most Fair Model

In the comparative analysis of models presented in Table 4.2, the SA-STGCN model with $\lambda = 5$ and $\gamma = 0.07$ emerges as the most equitable in terms of service provision to customers. This model significantly reduces the variability in customer wait times by 8.43% compared to the baseline STGCN model, albeit at a slight increase in average wait times of 0.68%. Notably, this model also reduces disparity regarding demand prediction, achieving the lowest MVPE of 3.32 and GEI of 2.13. These results underscore the correlation between fairer prediction models and more equitable

service provision in the downstream.

Conversely, when considering a balance between efficiency and disparity in vehicle rebalancing, the SA-STGCN model with $\lambda = 0.5$ and $\gamma = 0.09$ stands out. In comparison to the baseline STGCN model, it successfully lowers both the standard deviation and average of customer wait times by 6.48% and 0.49%, respectively. Impressively, this model even outperforms scenarios with perfect demand knowledge in terms of average customer wait times. This can be attributed to the equity-enhancing MIVR model, which approximates the matching component to account for future time intervals. Such approximations, while not perfectly optimal for real-world ride-hailing scenarios, demonstrate that sometimes, less accurate demand predictions can paradoxically lead to more efficient system operation.

Trade-offs of Introducing Equity Weights

	Unsatisfaction Rate (%)	Wait Time Avg (seconds)	Wait Time Std (seconds)	Non-occupied VMT (miles)	Rebalancing Trip Number
True Demand	1.81	89.49	22.55	60.70	20.13
Historical Demand	1.78	90.51	22.62	60.73	19.53
Baseline STGCN	1.79	89.79	22.80	60.36	19.50
Baseline SA-STGCN	1.75	89.74	22.69	60.15	19.23
SA-STGCN ($\gamma = 0.01$)	1.79	89.87	22.78	60.29	19.38
SA-STGCN ($\gamma = 0.03$)	1.77	90.11	22.58	60.29	19.21
SA-STGCN ($\gamma = 0.05$)	1.75	89.57	22.28	60.04	19.25
SA-STGCN ($\gamma = 0.07$)	1.73	89.50	22.21	59.82	19.02
SA-STGCN ($\gamma = 0.09$)	1.76	89.32	22.56	59.88	19.16
SA-STGCN ($\lambda = 0.5$)	1.79	89.77	22.37	60.17	19.30
SA-STGCN ($\lambda = 1$)	1.76	90.14	22.59	60.21	19.13
SA-STGCN ($\lambda = 3$)	1.78	90.02	23.01	60.07	19.10
SA-STGCN ($\lambda = 5$)	1.77	89.80	22.50	60.03	19.14
SA-STGCN ($\lambda = 0.5, \gamma = 0.09$)	1.74	89.21	21.84	59.89	19.24
SA-STGCN ($\lambda = 1, \gamma = 0.07$)	1.80	89.70	21.96	59.83	19.05
SA-STGCN ($\lambda = 3, \gamma = 0.03$)	1.78	90.03	22.35	60.03	18.99
SA-STGCN ($\lambda = 5, \gamma = 0.07$)	1.75	90.35	21.79	59.91	18.63

Table 4.3: Model performance summary without considering equity weights in the MIVR model.

Finally, we explore the implications of incorporating equity weights into the MIVR model. Table 4.3 presents the performance metrics of the MIVR model sans the application of equity weights in determining vehicle rebalancing strategies.

When analyzing the results from both tables, two significant insights are drawn. First, incorporating equity weights into the MIVR model notably reduces service disparity, especially by diminishing discrepancies in customer wait times. Additionally, this approach results in serving a higher proportion of customers. This improvement is attributed to better servicing of customers in previously underserved areas, as more vehicles are strategically repositioned there. Evidence of this is seen in the increased average distance traveled by non-occupied vehicles (VMT) and the reduced number of rebalancing trips, which supports our hypothesis. However, the second observation reveals a compromise: while service disparity is reduced, there is an observable rise in the average customer wait time across the system.

4.5 Discussion

4.5.1 Policy Discussion

On October 30, 2023, President Joe Biden issued an executive order focusing on "safe, secure, and trustworthy artificial intelligence" [152], emphasizing the critical need to address the risks associated with artificial intelligence and establish new standards for its safety and security. Disparity is an important component for understanding equity. This study reveals disparity issues within ride-hailing service algorithms, prompting consideration for regulatory interventions that would compel ride-hailing companies to integrate equity considerations into their algorithmic design. For instance, governments can mandate that the variation in waiting times across neighborhoods should not exceed a predefined threshold. Governments may also prescribe specific weights ω_i for individual regions within the vehicle rebalancing objective function (Equation 4.8), thereby exerting control over the importance assigned to successful matching in each distinct region i.

Concerning ride-hailing companies, the disparity-reducing algorithm introduced in this research offers an effective solution to reduce service disparity in ride-hailing. By adopting the proposed disparity-reducing strategies in both demand prediction and

service rebalancing, ride-hailing companies can greatly reduce variations in passenger waiting times across the region. It's worth highlighting that with the proper selection of hyperparameters λ and γ in the objective function of demand prediction (Equation 4.5), both the variation and average customer waiting time can be reduced. This achievement positions ride-hailing service companies to attain a Pareto improvement, concurrently enhancing efficiency and reducing service disparity in their operations and achieving a "win-win" scenario.

To enhance equity in vehicle distribution, ride-hailing services can adopt strategies like Lyft's 'Power Zones'. These zones, located in underserved areas, offer drivers bonuses for accepting rides originating from these regions. This approach addresses the income disparity caused by low-demand zones, where surge pricing is less frequent and earning opportunities are diminished. By incentivizing drivers to operate in these areas, companies can not only ensure better service coverage but also support drivers in maintaining consistent earnings.

From the perspective of passengers, this approach, which aims for a more equitable distribution of wait times among communities, especially benefits individuals facing challenges with transportation options. By fine-tuning both demand and rebalancing models, it has the potential to simultaneously attain a lower and more equitable waiting time distribution among people from different communities. This improvement enhances mobility and connectivity across all communities, fostering trust among passengers in the ride-hailing service, especially within disadvantaged communities.

The ride-hailing service plays a pivotal role as a transportation solution for individuals seeking employment and opportunities. The decrease in waiting times translates to more efficient access to opportunities for residents in diverse communities, offering significant relevance for areas with elevated unemployment rates or limited local job options. Through the adoption of this disparity-reducing algorithm, ride-hailing companies not only showcase their commitment to social responsibility but also actively contribute to the overall well-being and inclusivity of the communities they serve.

4.5.2 Practical Discussion

One might question why the current ride-hailing system does not reach extreme despite the negative feedback loop illustrated in Figure 4-1. We identify four factors that help maintain the system's equilibrium.

Firstly, the rebalancing algorithms for vehicles assume that drivers will strictly adhere to the guidance provided by the platforms. This assumption holds true for autonomous fleets like Waymo and Cruise but is unrealistic for traditional ride-hailing services such as Uber and Lyft. These platforms incentivize drivers to relocate to specific areas to fulfill requests, using incentives as a rough approximation for the rebalancing decisions that contribute to the negative feedback loop.

Secondly, there is a heterogeneity in driver behavior within the system. Many drivers prefer to work within specific areas, sometimes even those with fewer incentives, due to proximity to their homes or greater familiarity, which enhances their efficiency in handling local customer requests.

Thirdly, the platforms employ various strategies to serve customers even when no vehicles are immediately available in the vicinity. It's not uncommon for a request in a remote area to be fulfilled by a driver 10 to 20 minutes away. The goal of these ride-hailing platforms is to accommodate every customer request, striving to find a match for each one.

Lastly, the demand in areas with low request rates never completely disappears. Under certain conditions, customers are willing to endure longer wait times for a ride. This patience is particularly noted among customers who have a safe place to wait or when ride-hailing services present the only viable option, such as in inclement weather conditions.

4.6 Conclusion

This chapter presents a pioneering framework aimed at reducing disparity in both predicting ride-hailing demand and delivering equitable service to riders. The framework introduces a Socio-Aware Spatio-Temporal Graph Convolutional Network (SA-

STGCN), which integrates a socio-enriched adjacency matrix and bias-reduction regularization methods. Additionally, it features a vehicle rebalancing engine that incorporates equity considerations into its objective function. This framework was evaluated using a simulator with real-world ride-hailing data, demonstrating that the SA-STGCN model not only outperforms standard demand prediction models in increasing accuracy but also in reducing error disparity. Significantly, mitigation in disparity at the demand prediction stage lead to more equitable service delivery in the vehicle rebalancing process. The vehicle rebalancing module, enhanced with equity weights, showed a notable reduction in the standard deviation of customer wait times by 6.5%, while not diminishing the system efficiency for ride-hailing platforms.

The proposed framework offers a viable approach for ride-hailing companies to reduce service disparity into their operations, and it provides a basis for government regulations aimed at preventing service imbalances across different areas. However, realizing the win-win scenario highlighted in the study involves addressing practical challenges. A key solution lies in developing driver incentive mechanisms. These mechanisms should ensure that drivers are motivated to serve in underserved communities and that their earnings remain stable despite such commitments. As the role of ride-hailing services becomes more central in our everyday activities, it's crucial to make certain that these platforms maintain a strong commitment to social responsibility and proactively enhance the well-being and inclusiveness of the communities they operate in.

This chapter does not make a formal judgement on what is a *fair* vehicle rebalancing operation. Instead, it tries to understand and reduce disparity within the system, which serves as the foundation for understanding the fairness in the system. More analyses can be done to better understand and achieve the fairness in the vehicle rebalancing.

For future research, it would be beneficial to include a focus on driver behaviors and earnings, which this study has not addressed. A comprehensive framework could be developed to reduce disparity across all aspects of the ride-hailing vehicle rebalancing operations: error disparity in demand prediction, pricing disparity for riders, and

earning disparity for drivers. Such a framework should ensure that drivers who are redirected to serve underserved communities are compensated equitably, comparable to those serving in city centers. This approach would create a more balanced and fair environment for all parties involved in the ride-hailing ecosystem.

Chapter 5

Robust Transit Frequency Setting Problem with Demand Uncertainty

5.1 Introduction

The past century has witnessed one of the most dramatic evolutions in human history, urbanization. More than half of the world now lives in urban areas. By 2050, over two-thirds of the world's population is expected to live in urban areas [137]. Urban mobility, defined as moving people from one place to another within or between urban areas, is critical to the functionality of people's daily lives in urban areas. It allows people to access housing, jobs, and recreational services. However, urban mobility is also the largest contributor to greenhouse gas emissions in the United States, accounting for over 27% of the total greenhouse gas emissions [157]. Therefore, an efficient and sustainable urban mobility system is necessary to support future urban development.

Although emerging urban mobility services, e.g., ride-hailing and bike-sharing, have provided people with various options for traveling, public transit systems keep serving as the backbone of a sustainable urban mobility system, which allows more efficient travel across cities for a mass number of people. Meanwhile, public transit systems provide an affordable travel option for everyone regardless of travel distances within cities. Hence, it is important to design a public transit system with a good

level of service and operate it efficiently.

The COVID-19 pandemic has imposed an enormous impact on public transit systems. The national public transportation ridership stays around 60% of the pre-pandemic ridership level at the beginning of 2022 [4]. One of the main driving forces for the ridership drop is the flexible or remote working adopted by many employers worldwide during the pandemic. However, remote working won't be a temporary strategy for companies because the US is projected to have an average of 30% paid full days working from home for people in the future compared to a 5% pre-pandemic level [10]. Remote working implies that a proportion of commute trips in transit may be lost permanently, which motivates transit agencies to redesign their transit networks and schedules with the new demand patterns. Also, transit demand has become more volatile. Predicting future demand becomes more challenging.

While transit networks have been developed for years and are hard to change by transit agencies within a short period of time, changing transit schedules is straightforward. In this chapter, we focus on the transit frequency setting problem (TFSP), where transit schedules are optimized given a set of transit stops to serve. Though TFSP has been explored in previous literature, there is a limited number of papers incorporating uncertainty (such as volatile demand) into consideration for the TFSP [83, 180, 81]. Ignoring demand uncertainty when setting up transit schedules may diminish the level of service for transit systems.

To handle demand uncertainty for transit systems, especially during the post-COVID and remote work era [10, 36, 37], we first propose a baseline TFSP model for a single transit line. Next, we introduce the Robust Optimization (RO) technique into the TFSP to incorporate demand uncertainty. Furthermore, the Transit Downsizing (TD) approach is proposed to reduce problem dimensionality and generate optimal transit schedules efficiently. Also, the proposed TD approach can be utilized in other transit-related problems. A benchmark TFSP model using the Stochastic Programming (SP) technique is proposed and compared with the robust TFSP model. Overall, the contribution of this chapter can be summarized as follows:

- Propose a nominal TFSP model under a single transit line setting and an ex-

tended TFSP model considering crowding levels.

- Address demand uncertainty issues by introducing a robust TFSP model, which generates transit schedules that are optimized for the worst-case demand scenario. To the best of the authors' knowledge, this is the first study to utilize RO techniques to address demand uncertainty in TFSP.

- Design the TD approach to reduce problem sizes and make the model tractable given large-scale demand matrices from real-world transit instances. Theoretically prove that the optimal objective function of the problem after TD is close to that of the original problem (i.e., the difference is bounded from above).

- Compare the current transit schedule with the schedules solved by nominal, stochastic, and robust optimization, respectively, under multiple demand scenarios over the same corridor with two real-world transit lines (Routes 49 and X49) operated in Chicago. The robust TFSP model outperforms the benchmark stochastic TFSP model and existing transit schedules by simultaneously reducing passenger wait times and in-vehicle travel times under scenarios with significant demand uncertainty.

The remainder of the chapter is organized as follows. Section 5.2 reviews the relevant literature. Section 5.3 describes the nominal, robust, and benchmark stochastic TFSP models and proposed dimensionality reduction algorithms. Section 5.4 outlines experimental setups, including utilized data and transit lines, and displays experiment results and sensitivity analyses. Finally, Section 5.5 recaps the main contributions of this work, outlines the limitations, and provides future research directions.

5.2 Literature Review

5.2.1 Transit Frequency Setting Problem

The design and planning of urban public transit systems consist of a series of decisions before operating the system, which is known as Transit Network Planning

(TNP) problem. In literature, TNP is commonly divided into sub-problems that range across tactical, strategical, and operational decisions, including Transit Network Design (TND), Frequency Setting (FS), Transit Network Timetabling (TNT), Vehicle Scheduling Problem (VSP), Driver Scheduling Problem (DSP), Driver Rostering Problem (DRP). A thorough review of TNP and its sub-problems can be found in [83, 54, 42].

The TFSP is defined as a problem to determine the number of trips for a given set of lines that provide a high level of service in a planning period. The TFSP is first studied by Newell et al.[129] using analytic models. Given a fixed number of vehicles and constant passenger arrival rate, Newell et al.[129] produced vehicle dispatching time in order to minimize the total waiting time of all passengers. The study concluded that the optimal headway should be approximated as the square root of the arrival rate of passengers. The proposed model assumes fixed passenger demand and overlooks vehicle capacity constraints.

Furth and Wilson [55] formulated the TFSP as a non-linear program that computed the optimal headway for bus routes in order to maximize the net social benefits, consisting of ridership benefits and wait-time savings. Sets of constraints incorporated in their model were total subsidy, maximum fleet size, and acceptable level of loading. A key assumption they made was considering responsive demand which was a function of headway in the model. Furthermore, a heuristic-based algorithm was designed to solve non-linear programs.

More recently, Verbas and Mahmassani [161] extended the model proposed by Furth and Wilson [55] by incorporating service patterns into transit routes. A service pattern corresponds to a unique set of stops that need to be served by transit vehicles along a transit route. They formulated two non-linear optimization problems with different objectives: i) maximize the number of riders and wait time savings, and ii) minimize the net cost. Non-linear optimization solvers were directly used to solve non-linear programs. Additionally, Verbas et al. [162] discussed the impact of demand elasticity over solutions from the TFSP which is similar to models proposed by Furth and Wilson [55] and Verbas and Mahmassani [161]. They introduced three method-

ologies for estimating demand elasticity within transit networks and solved TFSPs under multiple demand elasticity scenarios on a large-scale network. Although the impact of demand uncertainty is discussed in this paper, their proposed methods are not equipped with abilities to generate optimal schedules considering demand uncertainty explicitly.

One could argue that one of the modeling contributions in formulations based on Furth and Wilson's [55] model is the introduction of responsive demand. However, the authors claim that it is more reasonable to consider a fixed demand matrix when solving the TFSP. There are short-term and long-term objectives in the TFSP: i) minimizing wait times for existing passengers, and ii) attracting more passengers to use transit networks. Minimizing wait times for the existing passengers leads to an increase in the level of service, which in turn attracts more passengers to take transit. On the contrary, maximizing ridership when considering responsive demand could lead to a waste of resources since it takes weeks for demand to respond to service changes. Meanwhile, transit schedules are modified frequently in practice, e.g., Chicago Transit Authority (CTA) publishes new transit schedules quarterly. An updated demand matrix can be utilized when generating new transit schedules every time. Therefore, minimizing wait times for existing passengers is a better objective in the authors' opinion.

Although limited papers take demand uncertainty into consideration when setting transit frequencies, Li et al. [106] utilized stochastic programming techniques to solve the headway optimization problem for a single bus route considering random passenger arrivals, boarding, alighting, and vehicle travel time. A metaheuristic algorithm consisting of a stochastic simulation and a genetic algorithm was designed to solve the proposed model. Their proposed approach was compared with three traditional headway determination models and bringing both demand and travel time uncertainty improved model performances. The main critique for Li et al. [106]'s work is the lack of discussions on the optimality gap given a heuristic-based solution algorithm. It is worth noting that incorporating travel time uncertainty when setting transit frequency can lead to better transit schedules. One approach is to utilize ad-

vanced techniques to better predict the traffic conditions [66, 196, 65] and use more realistic travel time information in the model.

Gkiotsalitis et al. [58] considered the frequency setting problem for autonomous minibuses with demand uncertainty. They utilized a traditional stochastic optimization approach, Sample Average Approximation (SAA), to handle the demand uncertainty. Moreover, no existing studies on the TFSP have incorporated the RO technique to address data uncertainties (e.g., demand uncertainty and travel time uncertainty). One major barrier to building the RO-based TFSP model is the dimensionality issue, where the robust counterpart (a solvable formulation of the RO model) significantly expands the problem size. In this paper, the TD approach is proposed to make the robust TFSP model computationally tractable given large-scale transit instances. The proposed TD approach can be generalized to any transit-related problems with large-scale demand matrices.

5.2.2 Robust Optimization and Applications in Urban Mobility

Robust Optimization (RO) is one of the widely-used approaches for decision-making under uncertainty in the Operations Research (OR) domain [12]. RO and its data-driven variants [18] are effective options to handle uncertain parameters. The underlying idea for RO is to specify a range for an uncertain parameter, namely an uncertainty set, and optimize over the worst-case realizations given the bounded uncertainty set. The solution method for RO problems involves generating a deterministic equivalent, called the robust counterpart. A practical guide on RO can be found in [61].

Urban mobility systems have various sources of uncertainty brought by human behaviors and environmental impacts (e.g., weather). Considering uncertainty when designing and operating urban mobility systems is crucial and necessary. There are several applications for applying RO techniques to solve urban mobility problems. For transit systems, Yan et al. [180] proposed a robust framework for solving the bus transit network design problem considering stochastic travel times. Mo et al. [125]

utilized the RO technique to solve the individual path recommendation problem under rail disruptions considering demand uncertainty. For shared mobility systems, Guo et al. [67] formulated a robust matching-integrated vehicle rebalancing (MIVR) model to balance vacant vehicles in the ride-hailing operations given demand uncertainty. Guo et al. [71] extended the MIVR model proposed by Guo et al. [67] by introducing predictive prescriptions approach [20] to handle demand uncertainty, which is an advanced approach for handling data uncertainty based on the stochastic optimization framework.

5.3 Methodology

5.3.1 Basic Optimization Model

We consider the TFSP for a single urban transit line (either rail or bus services) with a sequence of N stops. Let the set of stops be $\mathcal{S}$. A single line is the basic element of a transit network. Future studies can be extended to the network-level design by considering potential interactions between different lines. Without loss of generality, we assume each bi-directional transit line is considered as two separate transit lines with distinct sets of stops in this chapter. For an urban transit line, there exists a set of potential service patterns $\mathcal{P}$, where each pattern $p \in \mathcal{P}$ consists of a subset of stops $\mathcal{S}_p \subseteq \mathcal{S}$, indicating where the vehicles should stop if traveling with this pattern. Common examples of patterns are short-turnings and limited-stop lines in bus operations.

Let $\mathcal{V}$ represent the set of vehicle types that can be operated on the transit line. For instance, $\mathcal{V} = \{\text{standard bus, articulated bus, minibus}\}$ includes three types of buses, and $\mathcal{V} = \{\text{four-car train, six-car train, eight-car train}\}$ consists of three types of rail cars with a different number of carriages. For each type of vehicle $v \in \mathcal{V}$, the number of seats is C_v and the maximum vehicle capacity is $\bar{C}_v$. Furthermore, we discretize the full planning period $[T_{start}, T_{end}]$ into time periods $t = 1, ..., T$, where each time interval t has the same length Δ.

Let *passenger flow* (o, d, t) stand for passengers with origin station (stop) $o \in \mathcal{S}$ and destination station (stop) $d \in \mathcal{S}$ who arrives at the boarding station (stop) o at the beginning of time interval t. The set of passenger flows is indicated by $\mathcal{F}$. For each transit line, we have a demand matrix $\boldsymbol{u} = (u_t^{o,d})$, where $u_t^{o,d}$ indicates demand for the passengers flow (o, d, t). The decision variables for the TFSP are $\boldsymbol{x} = (x_t^{p,v})$, where $x_t^{p,v} = 1$ denotes a vehicle with type $v \in \mathcal{V}$ operating on a pattern $p \in \mathcal{P}$ departures from the terminal station of pattern p at the beginning of time interval t. Hence, unlike typical headway-based design, this chapter allows non-even dispatching of vehicles according to the service needs, where transit schedules can be better tailored to demand patterns.

In real-world transit line operations, transit agencies usually have a limited number of operating patterns for each line due to practical constraints. Having too many service patterns will confuse both transit operators and passengers. Therefore, we impose a *sparsity* constraint on operating patterns. Define an auxiliary decision variable $y_p, \forall p \in \mathcal{P}$, where $y_p = 1$ indicates that the pattern p can be operated on the transit line. Let P represent the maximum number of patterns operated on a single transit line. The sparsity constraint can be formulated as

$$x_t^{p,v} \leq y_p, \quad \forall t = 1, ..., T, \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \tag{5.1a}$$

$$\sum_{p \in \mathcal{P}} y_p \leq P. \tag{5.1b}$$

Let $c^{p,v}$ stand for the cost parameter associated with operating a vehicle of type v on a pattern p. The budget for scheduling transit services over the transit line is represented by B. The set of feasible schedules is denoted by

$$\mathcal{X}_B = \Bigg\{ \boldsymbol{x} \in \{0,1\}^{|\mathcal{P}|\times|\mathcal{V}|\times T} : \sum_{p \in \mathcal{P}} \sum_{v \in \mathcal{V}} \sum_{t=1}^{T} c^{p,v} x_t^{p,v} \leq B; \\ \sum_{v \in \mathcal{V}} x_t^{p,v} \leq 1, \ \forall t = 1, ..., T, \forall p \in \mathcal{P}; \text{Constraints (5.1)} \Bigg\}. \tag{5.2}$$

The feasibility constraints in Equation (5.2) ensure that the total scheduled transit

services do not exceed the budget B and only one type of vehicle can be operated on each pattern during each time interval t[1]. Equation (5.2) imposes a general budget constraint, which can be modified to incorporate more complicated cases. For instance, the budget constraint can be adapted to ensure a limited number of vehicles for each vehicle type:

$$\sum_{p\in\mathcal{P}}\sum_{t=1}^{T} x_t^{p,v} \leq B_v, \quad \forall v \in \mathcal{V}, \tag{5.3}$$

where B_v is the number of available vehicles for each vehicle type v and the cost parameter $c^{p,v} = 1, \forall p \in \mathcal{P}, \forall v \in \mathcal{V}$. Meanwhile, additional constraints can be added to incorporate agency-specific constraints. For example, $\sum_{v\in\mathcal{V}} x_t^{p,v} \geq 5$ implies that at least 5 buses need to be scheduled to operate with pattern p during time t.

It's important to note that the set $\mathcal{X}_B$ defines viable schedules for bus lines. However, for rail lines, we need to consider a minimum departure interval constraint due to the physical limitations of the rail system. We assume that this minimum departure interval is equal to the length Δ. When dealing with rail lines, we introduce the following additional constraint:

$$\sum_{p\in\mathcal{P}}\sum_{v\in\mathcal{V}} x_t^{p,v} \leq 1, \quad \forall t = 1, ..., T. \tag{5.4}$$

To capture boarding for passenger flows, we define decision variables $\boldsymbol{\lambda} = (\lambda_{t,\tau}^{o,d,p,v})$, where $\lambda_{t,\tau}^{o,d,p,v} \in \mathbb{R}_+$[2] indicates the number of passengers in the passenger flow (o, d, t) who board on a vehicle v that departs at the first station of pattern p at time τ.

The waiting time, in-vehicle travel time, and dwell time used in this chapter are defined as

- **Waiting time**: let $w_{t,\tau}^{o,d,p,v}$ represent the waiting time for the passenger flow (o, d, t) to board the vehicle v which departs at the first station of the pattern p at time τ.

[1]It is worth mentioning that multiple patterns are allowed to be operated within the same time period.

[2]We relax the integer variable $\boldsymbol{\lambda}$ to continuous variable to increase tractability for solving the problem while maintaining a satisfying model performance.

- **In-vehicle travel time**[3]: let $\phi^{o,d,p}$ represent he in-vehicle travel time for passengers with an origin-destination pair (o,d) to take a transit vehicle operating on pattern p.

- **Dwell time**: dwell times are ignored in the model since they are generally small compared to in-vehicle times.

To compute waiting and in-vehicle travel times within the model, it's possible to derive pattern-specific travel times using vehicle location data from transit agencies. The waiting time, denoted as $w_{t,\tau}^{o,d,p,v}$, is calculated by knowing the passenger's start time t, alongside the departure time τ and pattern p of the transit vehicle they are boarding. This calculation leverages the fixed travel times specific to each pattern.

Let $L_{\tau}^{p,v,s}$ stands for the vehicle load after visiting the station $s \in \mathcal{S}_p$ of vehicle v which departures at the first station of the pattern p at time τ, i.e.,

$$L_{\tau}^{p,v,s} = \sum_{o \in \mathcal{S}_p^{\text{before}}(s)} \sum_{d \in \mathcal{S}_p^{\text{after}}(s)} \sum_{t=1}^{T_{\tau,p}^{o,d}} \lambda_{t,\tau}^{o,d,p,v}, \quad \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall s \in \mathcal{S}_p, \forall \tau = 1, ..., T, \quad (5.5)$$

where $\mathcal{S}_p^{\text{before}}(s), \mathcal{S}_p^{\text{after}}(s)$ indicate sets of stations in $\mathcal{S}_p$ which are before (include station s) and after the station s, respectively. $T_{\tau,p}^{o,d}$ indicates the latest time interval such that a passenger with the origin-destination pair (o,d) can board a transit vehicle that departs from the first station at the time τ with pattern p.

To guarantee the feasibility of the model, we introduce an auxiliary decision variable $\boldsymbol{\eta} = (\eta_t^{o,d} \geq 0)$ indicating the number of unsatisfied passenger flow (o,d,t) (i.e., passengers who can not be served by the transit system). $\boldsymbol{\eta}$ serves as a slack variable to guarantee the problem always has feasible solutions. Hence, the flow conservation constraints can be represented as:

$$\sum_{v \in \mathcal{V}} \sum_{p \in \mathcal{P}^{o,d}} \sum_{\tau = \tau_t^{o,d,p}}^{T} \lambda_{t,\tau}^{o,d,p,v} = u_t^{o,d} - \eta_t^{o,d}, \quad \forall (o,d,t) \in \mathcal{F}, \quad (5.6)$$

[3]We assume a pattern-specific fixed travel time to maintain the linearity of the optimization model, which can be extended to time-dependent travel time.

where $\tau_t^{o,d,p}$ represents the earliest departure time for vehicles that are operated on a pattern $p \in \mathcal{P}^{o,d}$ and can be boarded by the passenger flow (o,d,t). The set $\mathcal{P}^{o,d} \subseteq \mathcal{P}$ denote the set of patterns that includes both stations o and d. Equation (5.6) means that all passengers from a passenger flow will board vehicles or stay unsatisfied.

Then, we have the following Integer Linear Programming (ILP) formulation for setting optimal frequencies for urban transit lines:

$$\min_{\boldsymbol{x}\in\mathcal{X}_B,\boldsymbol{\lambda},\boldsymbol{\eta}} \quad \sum_{(o,d,t)\in\mathcal{F}} \sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \left(w_{t,\tau}^{o,d,p,v} + \gamma\phi^{o,d,p}\right) \lambda_{t,\tau}^{o,d,p,v} + M \sum_{(o,d,t)\in\mathcal{F}} \eta_t^{o,d} \tag{5.7a}$$

s.t. Constraints (5.5) and (5.6)

$$L_\tau^{p,v,s} \le \bar{C}_v x_\tau^{p,v},$$

$$\forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall s \in \mathcal{S}_p, \forall \tau = 1, ..., T; \tag{5.7b}$$

$$\lambda_{t,\tau}^{o,d,p,v} \ge 0,$$

$$\forall (o,d,t) \in \mathcal{F}, \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall \tau = 1, ..., T; \tag{5.7c}$$

$$\eta_t^{o,d} \ge 0, \quad \forall (o,d,t) \in \mathcal{F}. \tag{5.7d}$$

The objective function (5.7a) minimizes the total generalized journey time for passengers who take transit services and the penalty of unsatisfied passenger flows given the set of feasible transit schedules $\mathcal{X}_B$. γ is a weight parameter controlling the importance between wait times and in-vehicle travel times. $\gamma = 0$ leads to a problem that only minimizes passengers' wait times and $\gamma = 1$ generates a problem that minimizes passengers' journey times (i.e., wait plus in-vehicle times). M stands for a large number that dominates the objective function (5.7a), indicating that all passenger flows should be served in the transit system. Constraints (5.7b) guarantee that passenger loads on vehicles do not exceed the vehicle capacity. Constraints (5.7c) and (5.7d) ensure that decision variables $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ are non-negative.

5.3.2 Optimization Model with Crowding Extension

Passengers may have different comfort levels depending on the degree of crowding in a vehicle and whether they can have a seat or not. Also, the potential infection risks of COVID-19 require transit agencies to control the vehicle load. To enhance the model's capability in managing crowding levels on transit vehicles, we've integrated a binary auxiliary variable, $\boldsymbol{z} = (z_t^{p,v,s})$. Here, $z_t^{p,v,s} = 1$ signifies that a vehicle of type v, following route pattern p and setting off from the terminal at time t, is *crowded* on the segment (s, s'), where s' is the next station following the station s. A transit vehicle v is *crowded* if the passenger load on the vehicle is greater than the seated capacity C_v[4].

Let ω represent the penalty cost per unit of travel time of a crowded transit vehicle. For a vehicle operating on a pattern p, let $\phi^{p,s}$ denote the vehicle running time of the segment after passing through station s. The ILP with crowding extension can be formulated as follows:

$$\min_{\boldsymbol{x}\in\mathcal{X}_B,\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{z}} \sum_{(o,d,t)\in\mathcal{F}} \sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \left(w_{t,\tau}^{o,d,p,v} + \gamma\phi^{o,d,p}\right) \lambda_{t,\tau}^{o,d,p,v} + M \sum_{(o,d,t)\in\mathcal{F}} \eta_t^{o,d}$$
$$+ \omega \sum_{p\in\mathcal{P}} \sum_{v\in\mathcal{V}} \sum_{s\in\mathcal{S}_p} \sum_{\tau=1}^{T} z_\tau^{p,v,s} \phi^{p,s} \tag{5.8a}$$

s.t. Constraints (5.5), (5.6), (5.7c), (5.7d)

$$L_\tau^{p,v,s} \le C_v x_\tau^{p,v} + (\bar{C}_v - C_v) z_\tau^{p,v,s}, \quad \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall s \in \mathcal{S}_p, \forall \tau = 1, ..., T; \tag{5.8b}$$

$$z_\tau^{p,v,s} \le x_\tau^{p,v}, \quad \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall s \in \mathcal{S}_p, \forall \tau = 1, ..., T; \tag{5.8c}$$

$$z_\tau^{p,v,s} \in \{0, 1\}, \quad \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall s \in \mathcal{S}_p, \forall \tau = 1, ..., T. \tag{5.8d}$$

Besides the objective for the baseline TFSP problem (5.7), the crowding penalty for transit vehicles is also added to the objective function as (5.8a), which leads to a transit schedule and passenger boarding choices minimizing the crowding levels.

[4]When passenger loading exceeds seated capacity, the proportion of passengers must stand and standees perceive up to 2.25 times actual travel time [154].

When $\omega = 0$, the problem (5.8) is equivalent to the problem (5.7), leading to transit schedules that minimize the total generalized journey time for passengers given passengers will board the first available transit vehicles. When $\omega > 0$, we assume passengers can wait for the next transit vehicle in order to reduce the crowding levels. It is worth noting that, in reality, passengers may or may not board a crowded vehicle depending on their comfort level requirement [126]. Our model simplifies the modeling of passengers' willingness to board and assumes that their boarding behavior minimizes the objective function. Hence, the objective function is a lower bound of the actual system cost. In this way, our model is useful for providing a perspective of system optimum and showing the trade-off between passengers' total waiting time and crowding levels in transit vehicles. Constraints (5.8b) are the modified capacity constraints with crowding level. Constraints (5.8c) restrict that a vehicle can only be crowded if it is operated in the system. Constraints (5.8d) specify decision variable z is binary.

In the following section, we introduce the robust TFSP model, developed through a robust optimization methodology. Both the standard TFSP model (5.7) and its crowding extension (5.8) can be extended to a robust version. However, for the simplicity of the chapter, we only discuss the robust version of the baseline TFSP model (5.7). The robust TFSP model with crowding extension can be derived following the same steps.

5.3.3 Robust Optimization Model Formulation

RO [12] is a widely-used approach in literature for decision-making under uncertainty. Compared to SP where the generated transit schedules are optimal for an "average" demand scenario, RO produces transit schedules that are optimized against the worst-case demand scenario. The motivation for introducing RO into transit frequency setting is that transit operators would prefer no passengers suffer from excessive wait times given any demand scenarios.

To construct a robust TFSP model, we define an uncertainty set around the uncertain demand parameter $u_t^{o,d}$. The uncertainty set specifies a range for the uncertain

demand $u_t^{o,d}$ where $u_t^{o,d}$ can change to any level within the range. Transit schedules are then generated using RO techniques with respect to the worst-case demand scenario in the uncertainty set.

We adopted the budget uncertainty set introduced by Bertsimas et al. [22], which is widely used in literature, to quantify the demand uncertainty in the TFSP. Let $\mu_t^{o,d}, \sigma_t^{o,d}$ denote the mean and standard deviation of the demand of passenger flow (o,d,t) derived from the historical data, respectively. The budget uncertainty set is defined as

$$\mathcal{U}(\Gamma) = \left\{ \boldsymbol{u} : \left| \frac{u_t^{o,d} - \mu_t^{o,d}}{\sigma_t^{o,d}} \right| \leq 1,\ \forall (o,d,t) \in \mathcal{F}; \sum_{(o,d,t)\in\mathcal{F}} \left| \frac{u_t^{o,d} - \mu_t^{o,d}}{\sigma_t^{o,d}} \right| \leq \Gamma \right\}, \tag{5.9}$$

where Γ is a parameter controlling the level of uncertainty for the budget uncertainty set. The budget uncertainty set implies that the demand can deviate from its historical average by at most one standard deviation, and the total absolute deviations for all passenger flows is upper-bounded by Γ. Define an uncertain parameter $\zeta \in \mathbb{R}^{|\mathcal{F}|}$ and let $u_t^{o,d} = \mu_t^{o,d} + \sigma_t^{o,d}\zeta_t^{o,d}$. We have the following reformulated uncertainty set:

$$\mathcal{U}(\Gamma) = \{\zeta : \|\zeta\|_\infty \leq 1, \|\zeta\|_1 \leq \Gamma\}. \tag{5.10}$$

With the defined uncertainty set over demand vector $\boldsymbol{u}$, we propose the robust TFSP model:

$$\min_{\boldsymbol{x}\in\mathcal{X}_B, \boldsymbol{\lambda}, \boldsymbol{\eta}} \quad \sum_{(o,d,t)\in\mathcal{F}} \sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \left(w_{t,\tau}^{o,d,p,v} + \gamma\phi^{o,d,p} \right) \lambda_{t,\tau}^{o,d,p,v} + M \sum_{(o,d,t)\in\mathcal{F}} \eta_t^{o,d} \tag{5.11a}$$

$$\text{s.t.} \quad \sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \lambda_{t,\tau}^{o,d,p,v} = \mu_t^{o,d} + \sigma_t^{o,d}\zeta_t^{o,d} - \eta_t^{o,d}, \quad \forall (o,d,t) \in \mathcal{F}, \forall \zeta \in \mathcal{U}(\Gamma); \tag{5.11b}$$

$$\sum_{o\in\mathcal{S}_p^{\text{before}}(s)} \sum_{d\in\mathcal{S}_p^{\text{after}}(s)} \sum_{t=1}^{T_{\tau,p}^{o,d}} \lambda_{t,\tau}^{o,d,p,v} \leq \bar{C}_v x_\tau^{p,v}, \quad \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall s \in \mathcal{S}_p, \forall \tau = 1, ..., T; \tag{5.11c}$$

$$\lambda_{t,\tau}^{o,d,p,v} \geq 0, \quad \forall (o,d,t) \in \mathcal{F}, \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall \tau = 1, ..., T; \tag{5.11d}$$

$$\eta_t^{o,d} \geq 0, \quad \forall (o,d,t) \in \mathcal{F}. \tag{5.11e}$$

Constraints (5.11b) in the robust formulation are equality constraints with uncertain parameters which often restrict the feasibility region drastically or even lead to infeasibility [61]. Therefore, we eliminate variables $\eta_t^{o,d}$ via substitution. Equality constraints (5.11b) can be reformulated as

$$\eta_t^{o,d} = \mu_t^{o,d} + \sigma_t^{o,d}\zeta_t^{o,d} - \sum_{v \in \mathcal{V}} \sum_{p \in \mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \lambda_{t,\tau}^{o,d,p,v}, \quad \forall (o,d,t) \in \mathcal{F}, \forall \boldsymbol{\zeta} \in \mathcal{U}(\Gamma). \tag{5.12}$$

Substituting Constraints (5.12) into the objective function (5.11a) and introducing a dummy variable α transform the original robust formulation into a problem formulation without equality constraints:

$$\min_{\boldsymbol{x} \in \mathcal{X}_B, \boldsymbol{\lambda}} \quad \alpha \tag{5.13a}$$

$$\text{s.t.} \quad \sum_{(o,d,t) \in \mathcal{F}} \sum_{v \in \mathcal{V}} \sum_{p \in \mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \left(w_{t,\tau}^{o,d,p,v} + \gamma \phi^{o,d,p} \right) \lambda_{t,\tau}^{o,d,p,v} - M \sum_{(o,d,t) \in \mathcal{F}} \sum_{v \in \mathcal{V}} \sum_{p \in \mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \lambda_{t,\tau}^{o,d,p,v}$$
$$+ M \sum_{(o,d,t) \in \mathcal{F}} \left(\mu_t^{o,d} + \sigma_t^{o,d}\zeta_t^{o,d} \right) \leq \alpha, \quad \forall \boldsymbol{\zeta} \in \mathcal{U}(\Gamma); \tag{5.13b}$$

$$\sum_{o \in \mathcal{S}_p^{\text{before}}(s)} \sum_{d \in \mathcal{S}_p^{\text{after}}(s)} \sum_{t=1}^{T_{\tau,p}^{o,d}} \lambda_{t,\tau}^{o,d,p,v} \leq \bar{C}_v x_\tau^{p,v}, \quad \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall s \in \mathcal{S}_p, \forall \tau = 1, ..., T; \tag{5.13c}$$

$$\mu_t^{o,d} + \sigma_t^{o,d}\zeta_t^{o,d} - \sum_{v \in \mathcal{V}} \sum_{p \in \mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \lambda_{t,\tau}^{o,d,p,v} \geq 0, \quad \forall (o,d,t) \in \mathcal{F}, \forall \boldsymbol{\zeta} \in \mathcal{U}(\Gamma); \tag{5.13d}$$

$$\lambda_{t,\tau}^{o,d,p,v} \geq 0, \quad \forall (o,d,t) \in \mathcal{F}, \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall \tau = 1, ..., T. \tag{5.13e}$$

However, equivalent formulations do not necessarily lead to equivalent robust counterparts, which are solvable reformulations of robust optimization problems. To

guarantee an identical robust counterpart, the substituted variable $\boldsymbol{\eta}$ needs to be *adaptive*, meaning that $\boldsymbol{\eta}(\boldsymbol{\zeta})$ becomes a function of uncertain parameter $\boldsymbol{\zeta}$. Linear Decision Rules (LDRs) are a commonly-used approximation method in literature to handle adaptive robust optimization problems [12, 16], which achieve satisfying performances in practice. Gorissen et al. [61] suggests that making uncertain variables adaptive and applying LDRs is equivalent to eliminating these variables, given coefficients of such variables do not include uncertain parameters and equality constraints are linear in uncertain parameters. Therefore, our reformulated robust optimization problem (5.13) is an approximated formulation of the original robust formulation (5.11), which is more tractable to solve without equality constraints.

To solve the problem (5.13), we need to derive the robust counterpart, which is a solvable formulation of the robust model. Details on the derivation of the robust counterpart can be found in Appendix C.1. The robust counterpart of the problem (5.13) is:

$$\min_{\boldsymbol{x}\in\mathcal{X}_B,\boldsymbol{\lambda},\boldsymbol{\nu}} \quad \alpha \tag{5.14a}$$

$$\text{s.t.} \quad \sum_{(o,d,t)\in\mathcal{F}} \sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \left(w_{t,\tau}^{o,d,p,v} + \gamma\phi^{o,d,p}\right) \lambda_{t,\tau}^{o,d,p,v} - M \sum_{(o,d,t)\in\mathcal{F}} \sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \lambda_{t,\tau}^{o,d,p,v}$$

$$+ M \sum_{(o,d,t)\in\mathcal{F}} \mu_t^{o,d} + \sum_{(o,d,t)\in\mathcal{F}} \nu_1^{o,d,t} + \Gamma\nu_2 \leq \alpha; \tag{5.14b}$$

$$\nu_1^{o,d,t} + \Gamma\nu_2 \geq M\sigma_t^{o,d}, \quad \forall (o,d,t) \in \mathcal{F}; \tag{5.14c}$$

$$\nu_1^{o,d,t} + \Gamma\nu_2 \geq -M\sigma_t^{o,d}, \quad \forall (o,d,t) \in \mathcal{F}; \tag{5.14d}$$

$$\nu_1^{o,d,t} \geq 0, \quad \forall (o,d,t) \in \mathcal{F}; \tag{5.14e}$$

$$\nu_2 \geq 0; \tag{5.14f}$$

$$\sum_{(o',d',t')\in\mathcal{F}} \nu_{o',d',t',3}^{o,d,t} + \nu_4^{o,d,t} \leq \mu_t^{o,d} - \sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \lambda_{t,\tau}^{o,d,p,v}, \quad \forall (o,d,t) \in \mathcal{F}; \tag{5.14g}$$

$$\nu_{o,d,t,3}^{o,d,t} + \nu_4^{o,d,t} \geq \sigma_t^{o,d}, \quad \forall (o,d,t) \in \mathcal{F}; \tag{5.14h}$$

$$\nu_{o,d,t,3}^{o,d,t} + \nu_4^{o,d,t} \geq -\sigma_t^{o,d}, \quad \forall (o,d,t) \in \mathcal{F}; \tag{5.14i}$$

$$\nu^{o,d,t}_{o',d',t',3} + \nu^{o,d,t}_{4} \geq 0, \quad \forall (o',d',t') \neq (o,d,t) \in \mathcal{F}; \tag{5.14j}$$

$$\nu^{o,d,t}_{o',d',t',3} \geq 0, \quad \forall (o',d',t'),(o,d,t) \in \mathcal{F}; \tag{5.14k}$$

$$\nu^{o,d,t}_{4} \geq 0, \quad \forall (o,d,t) \in \mathcal{F}; \tag{5.14l}$$

$$\sum_{o \in \mathcal{S}_p^{\text{before}}(s)} \sum_{d \in \mathcal{S}_p^{\text{after}}(s)} \sum_{t=1}^{T^{o,d}_{\tau,p}} \lambda^{o,d,p,v}_{t,\tau} \leq \bar{C}_v x^{p,v}_{\tau}, \quad \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall s \in \mathcal{S}_p, \forall \tau = 1,...,T; \tag{5.14m}$$

$$\lambda^{o,d,p,v}_{t,\tau} \geq 0, \quad \forall (o,d,t) \in \mathcal{F}, \forall p \in \mathcal{P}, \forall v \in \mathcal{V}, \forall \tau = 1,...,T. \tag{5.14n}$$

Constraints (5.14b) - (5.14f) are the robust counterpart corresponds to constraints (5.13b) while constraints (5.14g) - (5.14l) are the robust counterpart corresponds to constraints (5.13d). Compared to problem (5.13), the robust counterpart (5.14) introduces $(|\mathcal{F}|^2 + 2|\mathcal{F}| + 1)$ additional auxiliary non-negative continuous variables and $(|\mathcal{F}|^2 + 4|\mathcal{F}|)$ additional inequality constraints. When the number of distinct passenger flows $|\mathcal{F}|$ is not large (e.g., below 1,000), the robust counterpart (5.14) can be directly solved by off-the-shelf ILP solvers. However, the problem (5.14) can be intractable when $|\mathcal{F}|$ is large (e.g., above 10,000). In the later section, we will discuss the scalability issues for the TFSP under a single-line context and propose methods to handle large-scale TFSPs.

5.3.4 Benchmark Stochastic Programming Model Formulation

In this section, we propose an SP-based TFSP model as a benchmark model used in the experimental section. For the SP approach [28], the most traditional method is SAA, where the true distributions over uncertain parameters are approximated by empirical distributions obtained from the data [97]. The SAA is also utilized in the recent transit frequency setting work with demand uncertainty by Gkiotsalitis et al. [58].

Given a set of demand scenarios $\mathcal{E}$, the corresponding demand matrix $\boldsymbol{u}_e$ for a demand scenario $e \in \mathcal{E}$ has probability p_e. By introducing demand scenarios into the

frequency setting problem, we adjust the boarding decision variables for passengers to $\boldsymbol{\lambda}_e = (\lambda^{o,d,p,v}_{t,\tau,e})$ for each demand scenario $e \in \mathcal{E}$, where $\lambda^{o,d,p,v}_{t,\tau,e} \in \mathbb{R}_+$ represents the number of passengers in the passenger flow (o,d,t) who board on a vehicle v which departures at the beginning of pattern p at time τ under demand scenario e. Similarly, auxiliary variables $\boldsymbol{\eta}$ are extended to $\boldsymbol{\eta}_e = (\eta^{o,d}_{t,e})$ for each demand scenario $e \in \mathcal{E}$. Then the stochastic TFSP model can be formulated as:

$$\min_{\boldsymbol{x}\in\mathcal{X}_B,\boldsymbol{\lambda},\boldsymbol{\eta}} \quad \sum_{e\in\mathcal{E}} \left[\sum_{(o,d,t)\in\mathcal{F}} \sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau^{o,d,p}_t}^{T} \left(w^{o,d,p,v}_{t,\tau} + \gamma\phi^{o,d,p} \right) \lambda^{o,d,p,v}_{t,\tau,e} + M \sum_{(o,d,t)\in\mathcal{F}} \eta^{o,d}_{t,e} \right] \cdot p_e \tag{5.15a}$$

$$\text{s.t.} \quad L^{p,v,s}_{\tau,e} = \sum_{o\in\mathcal{S}^{\text{before}}_p(s)} \sum_{d\in\mathcal{S}^{\text{after}}_p(s)} \sum_{t=1}^{T^{o,d}_{\tau,p}} \lambda^{o,d,p,v}_{t,\tau,e}, \quad \forall p\in\mathcal{P}, \forall v\in\mathcal{V}, \forall s\in\mathcal{S}_p, \forall \tau = 1,...,T, \forall e\in\mathcal{E}; \tag{5.15b}$$

$$\sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau^{o,d,p}_t}^{T} \lambda^{o,d,p,v}_{t,\tau,e} = u^{o,d}_{t,e} - \eta^{o,d}_{t,e}, \quad \forall (o,d,t)\in\mathcal{F}, \forall e\in\mathcal{E}; \tag{5.15c}$$

$$L^{p,v,s}_{\tau,e} \le \bar{C}_v x^{p,v}_{\tau}, \quad \forall p\in\mathcal{P}, \forall v\in\mathcal{V}, \forall s\in\mathcal{S}_p, \forall \tau = 1,...,T, \forall e\in\mathcal{E}; \tag{5.15d}$$

$$\lambda^{o,d,p,v}_{t,\tau,e} \ge 0, \quad \forall (o,d,t)\in\mathcal{F}, \forall p\in\mathcal{P}, \forall v\in\mathcal{V}, \forall \tau = 1,...,T, \forall e\in\mathcal{E}; \tag{5.15e}$$

$$\eta^{o,d}_{t,e} \ge 0, \quad \forall (o,d,t)\in\mathcal{F}, \forall e\in\mathcal{E}. \tag{5.15f}$$

The problem (5.15) is a stochastic extension of the nominal optimization problem (5.7), and we minimize the expected total generalized journey time and penalties induced by unsatisfied demand across all demand scenarios. The number of variables and constraints grows linearly regarding the number of demand scenarios $|\mathcal{E}|$.

The model (5.15) provides a stochastic version of the TFSP model, which utilizes a different approach to handle the demand uncertainty compared to the robust TFSP model (5.11). It uses the same approach as Gkiotsalitis et al. [58] and it will be used as the benchmark model to evaluate the performance of our proposed robust TFSP model.

5.3.5 Optimization with Large-Scale Demand Matrix

In this section, we propose the Transit Downsizing (TD) approach to reduce the problem dimensionality and increase the tractability for proposed TFSP models given a large-scale demand matrix. As complexity issues are inherent in real-world transit problems, the proposed TD approach can be generalized to other design and operation problems in transit systems.

Consider a bus line, such as Route 49, and a rail line, like the Blue Line, operated by an organization like CTA, for instance. The inbound direction of the CTA Blue line includes 33 stations in total, which leads to 528 distinct OD pairs for passengers. When solving the transit frequency setting problem under a one-hour time interval with 12 decision time periods of length $\Delta = 5$ min, the number of passenger flows is $|\mathcal{F}| = 6,336$. Formulating the robust counterpart (5.14) introduces $40,157,569$ new continuous variables, which is a large-scale problem but might still be able to solve.

On the other hand, the northbound direction of the CTA route 49 bus contains 82 stops overall, which gives $1,176$ distinct OD pairs for passengers. Under the same setting as the Blue line, there will be $14,112$ unique passenger flows and the robust counterpart (5.14) introduces $199,176,768$ new continuous variables. The problem becomes intractable due to the excessive problem size. These two instances imply that large-scale demand matrices commonly exist in practice. Methods need to be designed to reduce the size of demand matrices in robust transit frequency setting problems.

The TD approach consists of two components: i) an optimality-preserved dimensionality reduction component, and ii) a heuristic-based dimensionality reduction component. The optimality-preserved component is proposed to reduce demand matrices based on the following observation: transit demand matrices are *sparse* and only a subset of passenger flows are chosen by passengers. Passengers using transit services have clear spatial and temporal patterns, which lead to sparsity in demand matrices.

Proposition 1. *For the nominal TFSP model (5.7) with a demand matrix* $\mathbf{u}$*, it is*

equivalent to solving the problem with a reduced set of passenger flow $\bar{\mathcal{F}}$, where $\bar{\mathcal{F}}$ only contains passenger flows with positive demand, i.e., $\bar{\mathcal{F}} = \{(o,d,t) : u_t^{o,d} > 0\}$.

Proof. For a passenger flow (o,d,t), when the demand is zero, i.e., $u_t^{o,d} = 0$, constraints (5.6) ensure that $\lambda_{t,\tau}^{o,d,p,v} = 0, \forall \tau = 1,...,T, p \in \mathcal{P}, v \in \mathcal{V}$, in the optimal solution given a minimization problem. Therefore, we can reach the same optimal solution by only considering passenger flows $\bar{\mathcal{F}}$ with positive demand only, i.e., $\bar{\mathcal{F}} = \{(o,d,t) : u_t^{o,d} > 0\}$. □

Proposition 1 reduces the problem size of the nominal TFSP model (5.7) and (5.8). It can also be applied to stochastic formulation (5.15) and robust formulation (5.11). For the stochastic TFSP model (SP), each demand scenario $e \in \mathcal{E}$ with demand matrix $\boldsymbol{u}_e$ leads to a reduced passenger flow set $\bar{\mathcal{F}}_e$, i.e., $\bar{\mathcal{F}}_e = \{(o,d,t) : u_{t,e}^{o,d} > 0\}$. For the robust TFSP model (RO), the reduced passenger flow set $\bar{\mathcal{F}}$ is constructed based on mean demand $\boldsymbol{\mu}$, i.e., $\bar{\mathcal{F}} = \{(o,d,t) : \mu_t^{o,d} > 0\}$.

The optimality-preserved component of the TD approach is extremely effective when solving nominal and stochastic models, where reduced passenger flow sets are established based on daily demand. When applying it to the robust problem with the average demand $\boldsymbol{\mu}$, the approach becomes less effective because the number of non-zero mean demand is still large. Considering the demand data from one month, a passenger flow (o,d,t) has to be incorporated in $\bar{\mathcal{F}}$ if it has demand for at least one day. We utilize a probabilistic scenario to better explain this issue. If a passenger flow (o,d,t) has a 90% probability to have zero demand in one day, the probability of not having a positive mean demand for 30 days is $0.9^{30} = 4.24\%$. When considering a month of demand data, the probability of excluding the passenger flow (o,d,t) from the problem shrinks from 90% to 4.24%, indicating that the first component of the TD approach is not effective for robust problems when considering demand data across multiple days.

Therefore, a heuristic-based dimensionality reduction component of the TD approach is further proposed to reduce the problem size of the robust TFSP model (5.11). It is constructed based on the following observation: if a passenger flow

(o, d, t) only appears once in a long period of time (e.g., one month), it is reasonable to exclude it from setting transit schedules given the same passenger flow (o, d, t) will most likely not be seen again in the future. The heuristic-based component introduces an adjusted passenger flow set $\tilde{\mathcal{F}} = \{(o, d, t) : \mu_t^{o,d} > \epsilon\}$, where passenger flows with mean demand below or equal to ϵ will be excluded from the optimization model.

The new problem after TD has a smaller scale and can be solved efficiently in practice. Compared to the original problem, the new problem has less number of constraints (i.e., a larger feasible space). Hence, its optimal objective function will be better (i.e., smaller in the minimization context). In the following analysis, we show that the difference between the objective functions of new and original problems is bounded. The bound is a function of ϵ. A smaller value of ϵ implies a tighter bound.

Define $Z^*(\mathcal{F})$ as the optimal objective function of the robust TFSP model (5.11) with passenger flow set $\mathcal{F}$. Then the optimal objective function of the problem after TD can be represented as $Z^*(\tilde{\mathcal{F}})$. We have the following lemma:

Lemma 1. For any given passenger flow set $\mathcal{F}_1$, define $\mathcal{F}_2$ as the passenger flow set by eliminating one passenger flow tuple (o, d, t) (i.e., $|\mathcal{F}_1|$ - $|\mathcal{F}_2| = 1$). Then, we have:

$$Z^*(\mathcal{F}_1) - Z^*(\mathcal{F}_2) \leq 2M \cdot \ell \tag{5.16}$$

where $\ell = \max_{(o,d,t)\in\mathcal{F}}(\mu_t^{o,d} + \sigma_t^{o,d})$.

Proof. When changing the passenger flow set $\mathcal{F}_1$ to $\mathcal{F}_2$ by excluding one passenger flow tuple (o, d, t), the objective value of the problem (5.11) decreases. The reduction of the objective value is induced by two reasons: i) less demand considered in the objective function, hence less total journey time and unsatisfied penalty, and ii) reallocation of passengers given more available vehicle capacity.

The robust TFSP model (5.11) minimizes the worst-case demand scenario. Therefore, we consider the worst-case objective loss when excluding one passenger flow (o, d, t). For the objective loss induced by demand reduction, it is upper-bounded by $M \cdot (\mu_t^{o,d} + \sigma_t^{o,d})$, since M dominants passengers' journey time and $(\mu_t^{o,d} + \sigma_t^{o,d})$ represents the largest demand for passenger flow (o, d, t) defined in the uncertainty

set $\mathcal{U}(\Gamma)$. Let $\ell = \max_{(o,d,t)\in\mathcal{F}}(\mu_t^{o,d} + \sigma_t^{o,d})$ and ℓ is a finite value since demand values in TFSP are finite integers. Then the objective loss from demand reduction is upper-bounded by $M \cdot \ell$.

For the objective loss induced by demand reallocation, excluding one passenger flow (o, d, t) equals having $(\mu_t^{o,d} + \sigma_t^{o,d})$ more vehicle capacity. The worst-case scenario is other unsatisfied passenger flows become satisfied when having more available capacity, which is upper-bounded by $M \cdot (\mu_t^{o,d} + \sigma_t^{o,d})$. Similar to the previous argument, it is upper-bounded by a finite value $M \cdot \ell$.

Combining two sources of the objective decrease, the maximum reduction of the objective value in (5.11) is upper-bounded by $2M \cdot \ell$ when excluding one passenger flow (o, d, t) from $\mathcal{F}_1$. □

Definition 1 (Dimensionality Reduction Function). Given the value of ϵ in the heuristic-based component of the TD approach, the dimensionality reduction function is defined as

$$f(\epsilon) = \left|\{(o, d, t) : \mu_t^{o,d} \leq \epsilon\}\right|, \tag{5.17}$$

which is the size of passenger flows excluded from $\mathcal{F}$. The dimensionality reduction function $f(\epsilon)$ has the following properties:

1. $f(\epsilon = 0) = 0$ (assuming all $\mu_t^{o,d} > 0$) and $\lim_{\epsilon\to\infty} f(\epsilon) = |\mathcal{F}|$.
2. $f(\epsilon)$ monotonically increases when ϵ increases.
3. $0 \leq f(\epsilon) \leq |\mathcal{F}| < +\infty$.

The first property holds because we do not exclude any passenger flows with when $\epsilon = 0$, and all passenger flows are excluded when ϵ is a large enough value. The second property holds since more passenger flows will be excluded when increasing ϵ. The last property is directly derived from the first two. Note that $f(\epsilon)$ is finite because the total number of passenger flows is finite considering a finite network and time interval in practice. By defining the dimensionality reduction function, we have the following proposition:

Proposition 2. *For the robust TFSP model (5.11) applying the TD approach, the objective reduction is upper-bounded by a finite value* $\Lambda(\epsilon) = 2M \cdot \ell \cdot f(\epsilon)$. *Mathematically:*

$$Z^*(\mathcal{F}) - Z^*(\tilde{\mathcal{F}}) \leq 2M \cdot \ell \cdot f(\epsilon) \tag{5.18}$$

$\Lambda(\epsilon)$ *has the following properties:*

1. $\Lambda(\epsilon = 0) = 0$.

2. $\Lambda(\epsilon)$ *monotonically increases when* ϵ *increases.*

Proof. Lemma 1 implies that the objective reduction due to excluding one passenger flow tuple (o, d, t) from $\mathcal{F}$ is upper-bounded by $2M\ell$. The size of passenger flow tuples excluding from $\mathcal{F}$ given ϵ is $f(\epsilon)$. Therefore, the objective reduction is upper-bounded by $2M\ell f(\epsilon)$, which is a finite value since $f(\epsilon)$ is upper-bounded by $|\mathcal{F}|$. Define $\Lambda(\epsilon) = 2M\ell f(\epsilon)$ and we have shown the objective loss $\Lambda(\epsilon)$ is upper-bounded.

According to the definition of dimensionality reduction function, when $\epsilon = 0$, we have $f(\epsilon = 0) = 0$, thus $\Lambda(\epsilon = 0) = 0$. Moreover, since $f(\epsilon)$ monotonically increases when ϵ increases, $\Lambda(\epsilon)$ also monotonically increases when ϵ increases. □

Proposition 2 indicates that the objective change due to the heuristic-based component of the TD approach is upper-bounded by a finite value. Meanwhile, decreasing the value of ϵ leads to a tighter bound. This shows that our proposed TD method is a valid approximation of the original problem with bounded errors. This proposition is validated with the experiments on the sensitivity analysis of ϵ in Section 5.4.3.

Setting the value of ϵ is critical in the proposed method. The value of ϵ should be chosen to balance the trade-off between transit schedule performance and problem complexity. Let m represent the number of days considered in the problem. The proposed heuristic approach works well in practice when setting $\epsilon = \frac{1}{m}$, indicating that passenger flows that appear only once over m days will be excluded from the problem.

Overall, the proposed TD approach helps to solve TFSPs with large-scale demand matrices. The first component maintains optimality and the second heuristic-based component could lead to sub-optimal solutions.

5.4 Results

In this section, the numerical results of the proposed models will be covered. All experimental results in this chapter were generated on a machine with a 3.0 GHz AMD Threadripper 2970WX Processor and 128 GB Memory. The linear programs in the experiments for generating optimal transit schedules and evaluating solution performances were solved with Gurobi 9.0.3 [74].

The results section is organized as follows. Section 5.4.1 describes data, parameter values, and experimental setups. Section 5.4.2 displays performance comparisons between the optimized schedule without considering demand uncertainty and the current schedule. Sensitivity analyses, crowding extensions, and optimization with multiple service patterns are also discussed in this section. Section 5.4.3 shows performance comparisons between robust, stochastic, and current transit schedules. Lastly, the computational performance of all proposed models is summarized in Section 5.4.4.

5.4.1 Data Description

Parameter values used in the experiments are shown in Table 5.1. The study transit lines used in the experiments are Route 49 northbound and Route X49 northbound operated by the CTA. Route 49 and Route X49 both serve Western Avenue in western Chicago. Route X49 is an expressed version of Route 49 with limited stops. Route 49 has 82 bus stops and Route X49 has 35 bus stops. Both routes share the same terminals and connect multiple rail line services: Orange, Pink, Green, Blue, and Brown lines.

In practice, transit schedules for Route 49 and Route X49 are determined separately. In our proposed optimization model, we will consider two routes as two patterns for a single transit line and generate both schedules simultaneously, i.e.,

Table 5.1: Model parameters and base case value.

Model Parameter	Explanation	Base Case Value
T_{start}	Start time of planning period	07:00
T_{end}	End time of planning period	09:00
Δ	Decision time interval length	5 (minutes)
T	Number of decision time periods	24
$\mathcal{P}$	Set of patterns for the transit line	{49, X49}
$\mathcal{V}$	Set of bus types	{standard, articulated}
C_v	Number of seats on buses	{37, 58}
$\bar{C}_v$	Maximum vehicle capacity	{70, 107}
$c^{p,v}, \forall p \in \mathcal{P}, \forall v \in \mathcal{V}$	Cost parameter for bus with pattern p and vehicle type v	1
B	Maximum bus supply during the planning period	20
M	Penalty for an unsatisfied passenger	10^5
γ	Weight parameter for in-vehicle travel time	1
m	Number of demand scenarios	22
ϵ	Heuristic parameter for demand matrix size reduction	0.05

$\mathcal{P} = \{49, X49\}$. The position of both patterns within the CTA transit network and stop overviews are shown in Figure 5-1. In the later section, we will explore the model's performance when introducing additional generated patterns to the studied transit line.

Figure 5-1: Positions and stop overviews of Pattern 49 and Pattern X49 in the CTA network.

The data utilized in the experiments are from 22 weekdays in October 2020. The current transit schedule information is from an open-source Generalized Transit Feed

Specification (GTFS) dataset, which is published by CTA every month. Regarding the running times between any two stops for different patterns, they are calculated based on the Automatic Vehicle Location (AVL) dataset of October 2020 provided by CTA. The OD matrix is generated based on CTA's ODX dataset from October 2020.

The "ODX" stands for "origin, destination, and transfer inference algorithm", an algorithm developed by Gabriel et al. [148] and currently implemented within the CTA. The CTA transit network is equipped with a "tap-on" only fare collection system, indicating that alighting information is not reported in the system. The ODX algorithm is utilized to infer the alighting information and details can be found in [148, 35, 198].

Besides the real-world data, the synthetic demand data is also generated to simulate the scenarios with heavy demand. The synthetic demand data is generated as follows: for each passenger flow (o, d, t) with a non-zero average demand value $\mu_t^{o,d}$ over 22 real-world demand scenarios, generate the new demand level according to a Poisson distribution $u_t^{o,d} \sim Pois(\beta \cdot \mu_t^{o,d})$, where β indicates an expansion factor.

The study period is a two-hour time interval from 7:00 AM to 9:00 AM. The length of each decision time interval is $\Delta = 5$ minutes, therefore, there are 24 time intervals considered in the transit frequency setting problem. For the existing transit schedule, there are 20 buses operating in total. The current northbound schedules for the study transit line are shown in Figure 5-2. Each colored dot represents a departure with a specific operation pattern from the terminal stop.

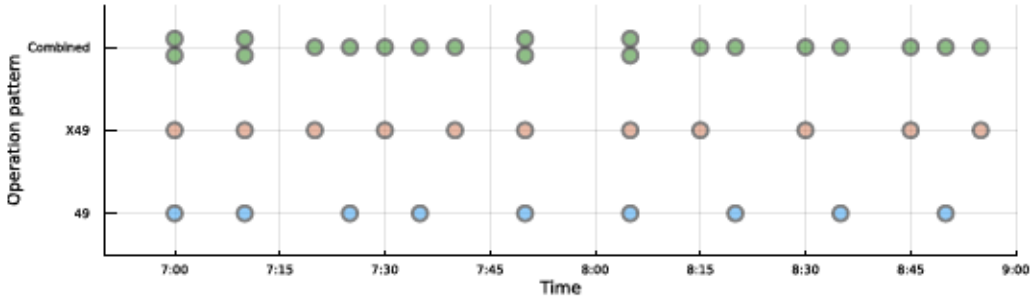


Figure 5-2: The current northbound transit schedule for Pattern 49, Pattern X49, and the combined transit line.

In the experiments, the budget constraint in Equation (5.2) ensures that the total number of buses operating within the overall time interval does not exceed the maximum bus supply, i.e., $c^{p,v} = 1, \forall p \in \mathcal{P}, \forall v \in \mathcal{V}$, and $B = 20$.

For buses used in the experiments, we consider two types of buses: regular buses and articulated buses, i.e., $\mathcal{V} = \{regular, articulated\}$. The regular bus has 37 seats and a maximum capacity is 70, while the articulated bus has 58 seats with a maximum capacity of 107. The current schedule only utilizes regular buses for Route 49 and Route X49. Therefore, only regular buses are considered in the base case scenario.

5.4.2 Baseline Model Performances

Optimal Transit Schedules

To evaluate the performances of the nominal TFSP model (5.7), we randomly choose a demand scenario from 22 weekdays to generate the optimal transit schedule, which is then compared with the current schedule over the remaining 21 demand scenarios. For the base case scenario, wait and travel times are equally important, i.e., $\gamma = 1$. The TD approach without the heuristic-based component is applied when solving the optimization model.

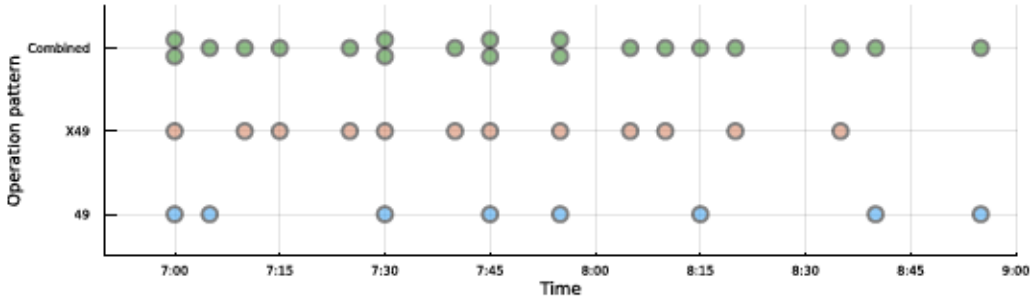


Figure 5-3: The optimized transit schedule without considering demand uncertainty based on a one-day demand scenario.

Figure 5-3 shows the optimized transit schedule without considering demand uncertainty based on a randomly selected one-day demand scenario. Each colored dot represents a departure with a specific operation pattern from the terminal stop. Compared to the current schedule shown in Figure 5-2, more buses are dispatched during

the first hour. The optimized transit schedule without considering demand uncertainty becomes irregular due to serving a specific demand scenario. Meanwhile, it shifts one bus from Route 49 to Route X49.

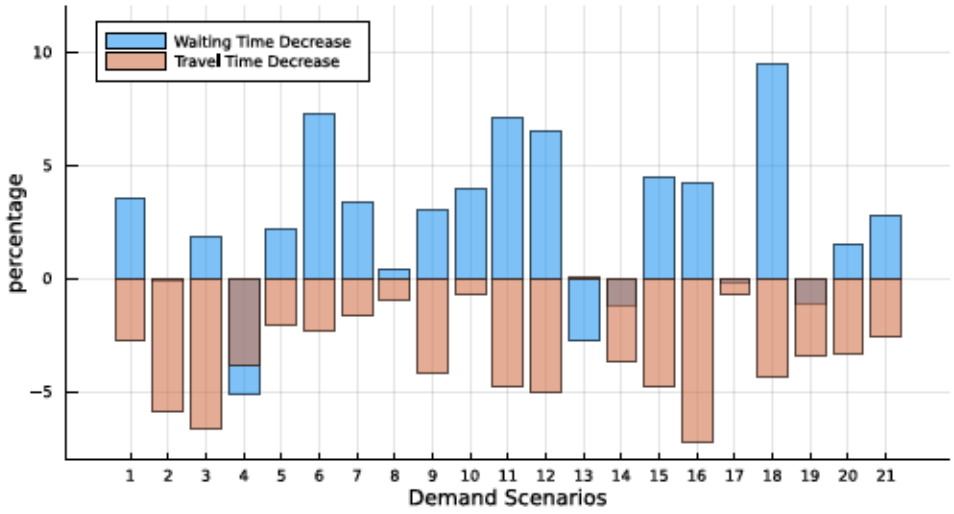


Figure 5-4: Performance comparisons between the current and the optimized transit schedules without considering demand uncertainty.

The performance comparison over 21 demand scenarios is shown in Figure 5-4. Bars indicate wait and travel time decreases for the optimal schedule compared to the current schedule. The wait time and travel time can be calculated for each passenger with the boarding variable $\lambda_{t,\tau}^{o,d,p,v}$ given the fixed pattern-specific travel times.

For the optimized transit schedule without considering demand uncertainty, passengers experience lower wait times in 15 out of 21 demand scenarios. However, passengers have higher in-vehicle travel times for almost all demand scenarios given the current transit schedule. In summary, a 2.43% wait time decrease and a 3.38% travel time increase are brought to passengers on average when switching from the current schedule to the optimized schedule without considering demand uncertainty. It works best for the input demand scenario of the optimization model. For other demand scenarios, it reduces passengers' wait times by sacrificing in-vehicle travel times.

The performance comparison indicates that demand uncertainty is crucial when generating transit schedules. The optimized transit schedule without considering

demand uncertainty does not have an edge over the existing transit schedule, which maintains a regular headway.

Crowding Extensions

Next, we will discuss the crowding extension of the nominal TFSP model (5.8). Existing demand scenarios from October 2020 lead to very few crowded transit vehicles. Therefore, model performances will be tested based on a synthetic demand scenario with an expanded demand level. In the following discussion, we generate synthetic demand scenarios with an expansion factor $\beta = 4$.

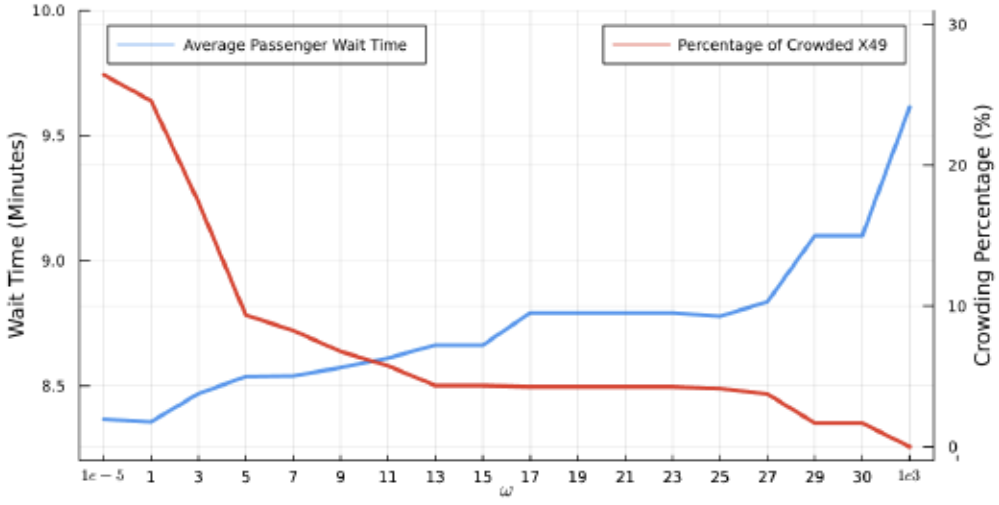


Figure 5-5: Trade-offs between average passenger wait times and crowding levels given different ω values.

In the crowding-extended model (5.8), parameter ω is utilized to control the level of penalty for crowded transit vehicles in the objective function. Figure 5-5 shows the average passenger wait time and percentage of crowded X49 given different values of ω. For the base case scenario ($\omega = 10^{-5}$) with the expanded demand scenario, 26.43% of operating time for transit vehicles on pattern X49 is crowded while 2.95% of pattern 49 operating time is crowded. The operating time indicates the time where a bus is on service. The average passenger wait time is 8.37 minutes, where the average passenger wait time is the average wait times among all passengers. When increasing the crowding penalty ω, the crowding level on pattern X49 decreases while

the average passenger wait time increases. When the value of ω exceeds a certain threshold, all passengers can have seats on buses and the average passenger wait time increases to 9.61 minutes, which is increased by 14.81%.

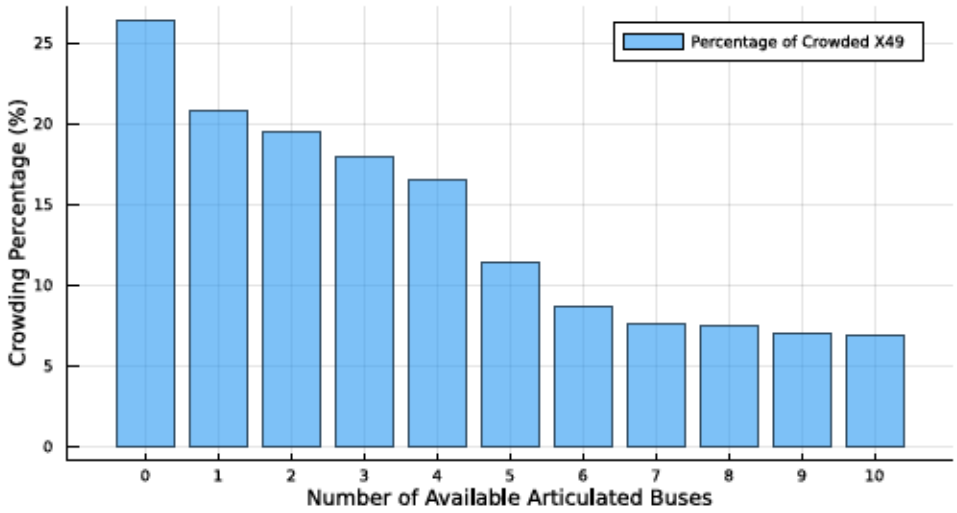


Figure 5-6: Crowding percentage of pattern X49 given different numbers of available articulated buses.

It is worth noting that the crowding level is reduced by the purposely left-behind behaviors of passengers. However, passengers will always board the first available transit vehicle in reality. One way to resolve this conflict is by introducing articulated buses with a larger seat capacity. Figure 5-6 displays the crowding percentage of pattern X49 given different numbers of available articulated buses. Introducing 6 additional articulated buses reduces the percentage of running time on pattern X49 with crowded transit vehicles to 8.72%. The optimized transit schedule with articulated buses is shown in Figure 5-7. Each colored dot represents a departure with a specific operation pattern from the terminal stop. Each pink dot indicates a departure of an articulated bus from the terminal stop. To better reduce the crowding on buses, articulated are dispatched within the first hour when more passengers are taking transit services.

The marginal benefit of bringing extra articulated buses drops significantly after having 6 articulated buses. For the scenario with 10 available articulated buses, the crowding percentage on pattern X49 is 7.04%. In summary, having a small fleet of

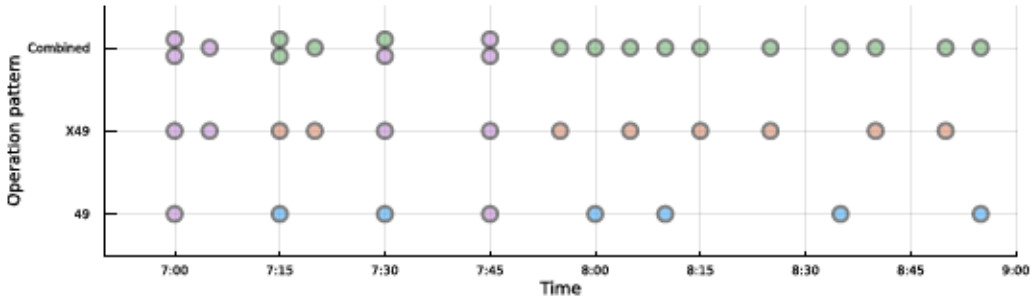


Figure 5-7: The optimal transit schedule with an expanded demand matrix and 6 available articulated buses.

articulated buses can reduce the crowding levels on buses significantly in bus operations.

Pattern Generation

In the initial phase, the optimization problem focused on two existing patterns within the studied transit line. To demonstrate the feasibility of incorporating multiple service patterns into the proposed TFSP model, we generated diverse patterns and assessed their performance across 50 randomly generated demand scenarios.

These additional service patterns were created based on a specified number of stops, denoted as k. For instance, when $k = 12$, the first and last stops were designated as terminals, and the remaining 10 stops were randomly selected using a weighted vector $\bar{\omega}$. This vector, $\bar{\omega}$, represented passenger counts at each bus stop, amalgamating both boarding and alighting actions as one utilization by passengers. Meanwhile, we assume a service pattern with fewer bus stops k has a shorter travel time between stops. In addition to the existing service patterns 49 and $X49$, we crafted 7 more patterns with varying stop numbers: $[12, 22, 32, 42, 52, 62, 72]$.

Building upon the baseline optimization problem comprising two service patterns, we sequentially introduced an additional pattern selected at random. This integration produced a new optimized transit schedule every time. Subsequently, we evaluated all these schedules across the spectrum of 50 randomly generated demand scenarios.

In real-world scenarios, transit lines usually do not operate multiple active service patterns simultaneously. To account for this practical constraint, we incorporate

Equation (5.1) into the TFSP model, where $P = 2$, signifying a maximum selection of two service patterns. The results of these evaluations are outlined in Table 5.2.

Table 5.2: Model Performances with Different Number of Patterns.

$\|\mathcal{P}\|$	n_{row}	n_{con}	n_{int}	*time*	*wait*	*travel*	*journey*
2	11842	38150	98	0.99	7.70	8.93	16.63
3	17866	57206	147	2.22	7.78	8.96	16.74
4	22930	75658	196	5.38	7.74	8.92	16.66
5	26074	93036	245	8.21	7.67	8.92	16.59
6	30178	110978	294	19.43	7.74	8.92	16.66
7	31402	127126	343	14.97	7.74	8.92	16.66
8	38386	146744	392	26.95	7.74	8.92	16.66
9	40570	163508	441	31.52	7.74	8.92	16.66

In Table 5.2, $|\mathcal{P}|$ represents the number of service patterns integrated into the TFSP model, n_{row} indicates the count of rows/constraints, and n_{con} and n_{int} denote the number of continuous and integer variables in the model, respectively. The variable *time* signifies the computational duration (in seconds) required to attain the optimal solution using Gurobi. *wait* represents the average passenger waiting time (in minutes), *travel* denotes the average in-vehicle travel time (in minutes) for passengers, and *journey* stands for the average overall journey time (in minutes) experienced by passengers.

Incorporating extra service patterns escalates the problem's complexity, causing a linear rise in both constraints and variables and consequently prolonging the optimization time. However, in terms of passenger service, introducing more patterns does not significantly change service performance. This is because all demand data is gathered from the existing two service patterns, and the distribution of demand for generating scenarios is based upon this existing data. Thus, it is challenging to generate better service patterns compared to the existing ones.

Sensitivity Analyses

Lastly, we test the sensitivity of the results when changing the weight parameter γ for in-vehicle travel times. In previous experiments, $\gamma = 1$ was used as a base case,

leading to a transit schedule that minimizes the total journey time. In this section, different values of γ ranging from 0 to 2 with a 0.1 step size are tested. Results are shown in Figure 5-8 and Figure 5-9.

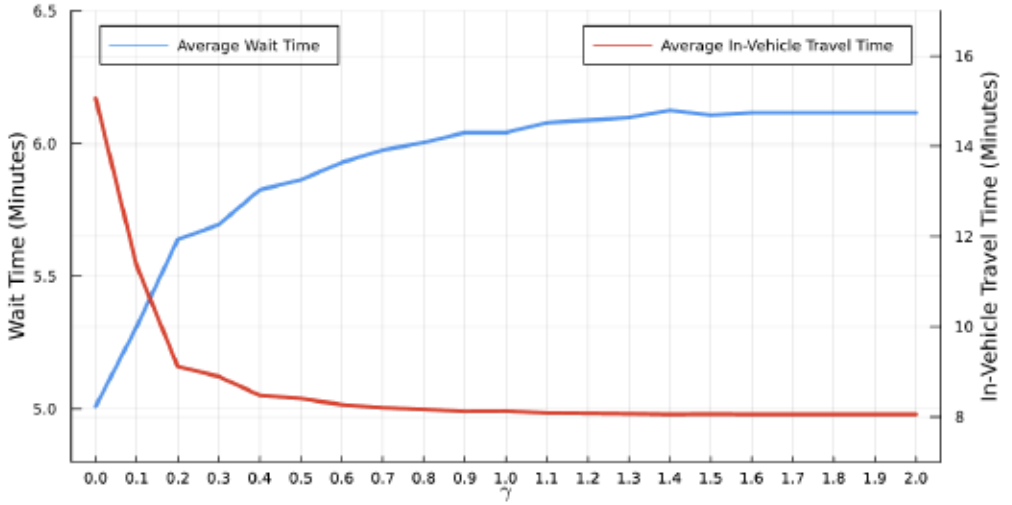


Figure 5-8: Sensitivity analyses results for the weight parameter γ with respect to average passenger wait time and average passenger in-vehicle travel time changes.

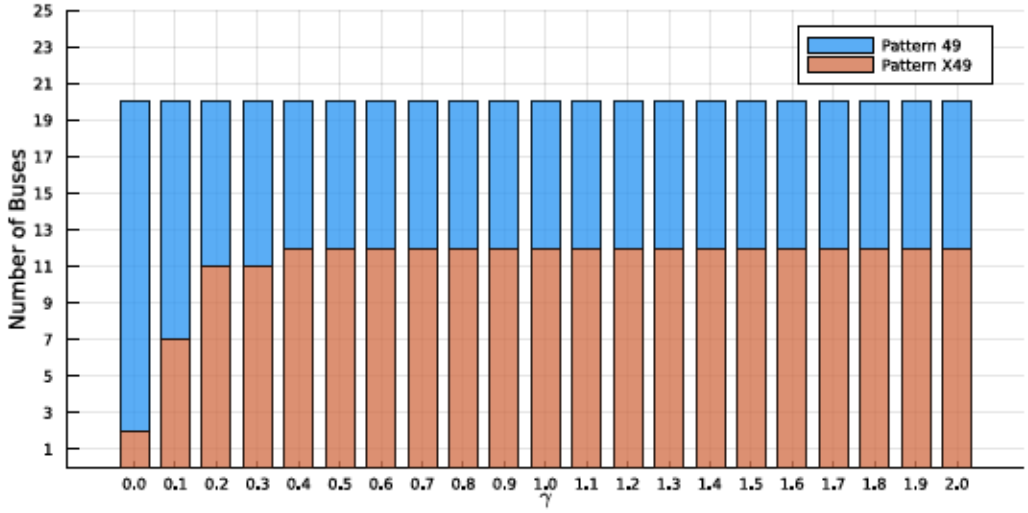


Figure 5-9: Sensitivity analyses results for the weight parameter γ with respect to the number of buses operated with pattern 49 and X49.

A smaller value of γ indicates that wait times are more important than in-vehicle travel times. For the scenario with $\gamma = 0$, where transit schedules solely minimize passengers' wait times, the average wait time is 5.01 minutes and the average in-

vehicle travel time is 15.07 minutes. The average wait time monotonically increases and the average in-vehicle travel time monotonically decreases when the value of γ increases, which is shown in Figure 5-8. For the scenario with $\gamma = 2$, where in-vehicle travel times are twice as important as wait times, the average wait time is 6.12 minutes and the average in-vehicle travel time is 8.05 minutes.

The average total travel time decreases from 20.08 minutes to 14.17 minutes when increasing γ from 0 to 2. This is intuitive; more vehicles will be operated with pattern X49 when increasing γ, and pattern X49 has a larger vehicle speed than pattern 49 given fewer bus stops. Figure 5-9 shows the number of buses running on each pattern given different values of γ. Only 2 bus with pattern X49 is operated when $\gamma = 0$, while 12 buses with pattern X49 are operated when γ becomes larger.

5.4.3 Robust Model Performances

To incorporate demand uncertainty into the TFSP, the robust TFSP model (5.11) and the benchmark stochastic TFSP model (5.15) are proposed. In this section, we will first show the performance of the benchmark stochastic TFSP model. The robust TFSP model is then compared with both the benchmark stochastic TFSP model and the current transit schedule over multiple synthetic demand scenarios. Two types of demand scenarios are generated following the method described in Section 5.4.1 to

Table 5.3: Performance evaluations for robust transit schedules under normal demand scenarios.

Γ	Wait Time	Compare	Improve	Travel Time	Compare	Improve	GAP	Time
0.0	7.749	-0.98%	3.79%	8.456	0.13%	0.93%	OPT	1910
1.0	7.706	-0.40%	4.34%	8.534	-0.79%	0.02%	8.51%	*Limit*
2.0	7.858	-2.42%	2.42%	8.443	0.29%	1.09%	3.56%	*Limit*
3.0	7.942	-3.50%	1.40%	8.425	0.49%	1.29%	2.40%	*Limit*
4.0	7.816	-1.87%	2.94%	8.435	0.38%	1.18%	1.38%	*Limit*
5.0	7.779	-1.39%	3.40%	8.425	0.50%	1.30%	1.08%	*Limit*
6.0	7.784	-1.46%	3.34%	8.429	0.46%	1.26%	0.75%	*Limit*
7.0	7.784	-1.46%	3.34%	8.429	0.46%	1.26%	0.51%	*Limit*
8.0	7.784	-1.46%	3.34%	8.429	0.46%	1.26%	OPT	7435
9.0	7.736	-0.84%	3.93%	8.443	0.29%	1.09%	OPT	9935
10.0	7.783	-1.44%	3.36%	8.433	0.40%	1.20%	OPT	3372

test the model performances: i) normal demand scenarios (without demand expansion $\beta = 1$), and ii) surge demand scenarios (w/ demand expansion $\beta = 4$).

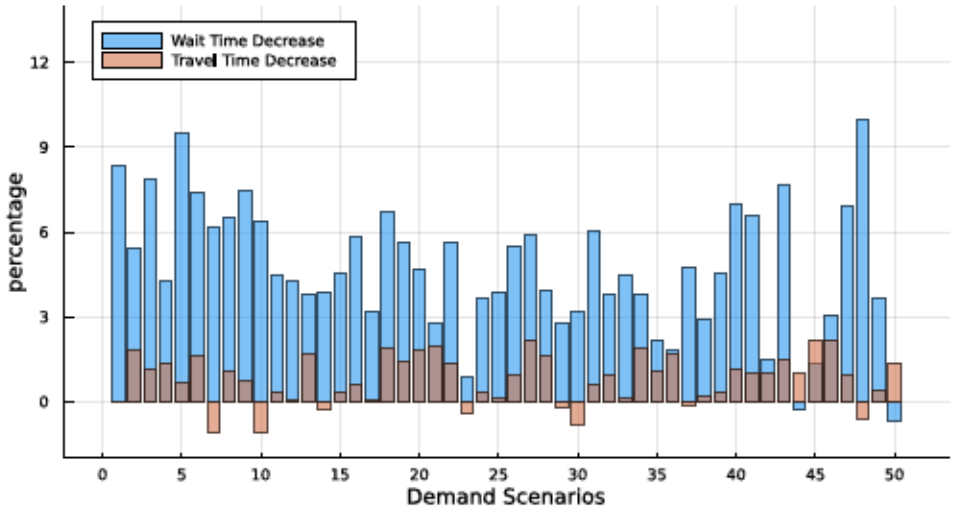


Figure 5-10: Performance comparisons between the current and the stochastic transit schedules over 50 randomly generated normal demand scenarios.

The benchmark stochastic transit schedule is generated by using the TD approach without the heuristic-based component and assuming equal probability for each demand scenario. Figure 5-10 shows the performance comparison between stochastic and current transit schedules over 50 randomly-generated normal demand scenarios. Blue bars represent wait time decrease for the stochastic transit schedule. Orange bars indicate travel time decrease for the stochastic transit schedule. On average, the stochastic schedule improves passengers' wait time by 4.71% and in-vehicle travel time by 0.80%. An optimized transit schedule over 22 demand scenarios is more robust than an optimized transit schedule with only one demand scenario. The stochastic transit schedule improves both wait and in-vehicle travel times in 41 out of 50 demand scenarios.

Figure 5-11 shows the stochastic transit schedule. Compared to the current transit schedule shown in Figure 5-2, it has fewer time intervals where buses are dispatched for both patterns. In the combined transit schedule, buses are spread more evenly during the two-hour decision time period. Meanwhile, one additional bus is operated with pattern X49. Compared to the optimal transit schedule with one-day demand

displayed in Figure 5-3, the stochastic transit schedule maintains a stable headway for both patterns, which is similar to the current schedule, where the headway-based transit operation strategy is utilized.

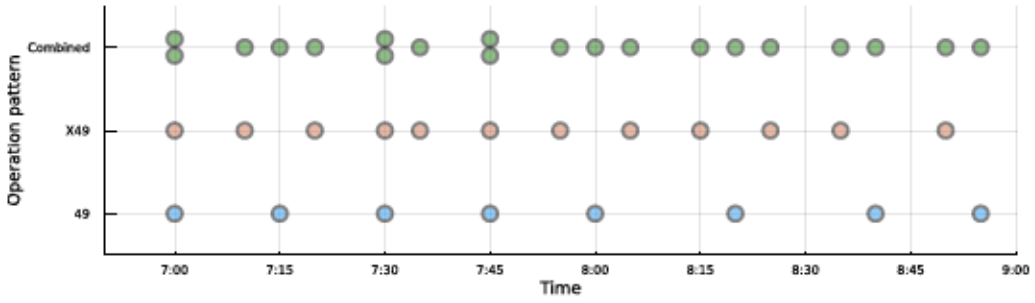


Figure 5-11: The stochastic transit schedule generated from 22 real-world demand scenarios.

For the robust transit schedule, it is generated by the TD approach with $\epsilon = 0.05$, meaning that a passenger flow (o, d, t) will be incorporated in the model only if it appears more than one time within 22 weekdays. The robust optimization model is solved by the off-the-shelf MIP (Mixed Integer Programming) solver Gurobi with a 3-hour time limit and an optimality gap of 0.5%. Results under normal demand scenarios are shown in Table 5.3. Γ indicates a parameter for controlling the size of budget uncertainty sets. *Wait Time* and *Travel Time* represent the average wait time and travel time for passengers over 50 randomly generated normal demand scenarios. *Compare* indicates the performance comparison with the stochastic transit schedule. *Improv* stands for the performance comparison with the current transit schedule. *GAP* is the optimality gap for the MIP solver and *Time* is the computational time (in seconds), where *Limit* indicates that the solver reaches the computational time budget 10800 seconds (3 hours).

Parameter Γ controls the level of demand uncertainty incorporated in the model. A higher value of Γ indicates that more demand uncertainty is considered when generating the robust transit schedule. When $\Gamma = 0$, the robust optimization is reduced to the nominal optimization model with the mean demand matrix $(\mu_t^{o,d})$ as the model input. For all uncertain scenarios, robust transit schedules outperform the current transit schedule by reducing both wait times and in-vehicle travel times.

Compared to the benchmark stochastic transit schedule, robust transit schedules have better in-vehicle travel times and worse wait times for passengers.

Table 5.4: Performance evaluations for robust transit schedules under surge demand scenarios.

Γ	Wait Time	Compare	Improve	Travel Time	Compare	Improve
0.0	8.219	-0.48%	3.68%	8.588	0.56%	1.49%
1.0	8.243	-0.75%	3.42%	8.691	-0.64%	0.31%
2.0	8.045	1.65%	5.72%	8.547	1.03%	1.95%
3.0	8.144	0.44%	4.56%	8.526	1.27%	2.19%
4.0	8.015	2.01%	6.07%	8.534	1.18%	2.11%
5.0	7.982	2.42%	6.46%	8.531	1.22%	2.14%
6.0	7.971	2.55%	6.58%	8.530	1.22%	2.15%
7.0	7.971	2.55%	6.58%	8.530	1.22%	2.15%
8.0	7.971	2.55%	6.58%	8.530	1.22%	2.15%
9.0	7.939	2.94%	6.96%	8.548	1.01%	1.94%
10.0	7.972	2.53%	6.57%	8.539	1.12%	2.05%

When increasing the value of Γ in the model, the robust optimization model becomes easier to be solved as the optimality gap becomes smaller. The model can be solved optimally when Γ is greater than 7. This can be explained as follows: a larger value of Γ leads to a less-restricted optimization problem; heuristic approaches implemented in Gurobi are more likely to produce feasible solutions; better heuristic solutions reduce the time for branch-and-bound significantly. With respect to the model performance, it does not have a pattern regarding the uncertain parameter Γ. The robust transit schedule with $\Gamma = 10$ is shown in Figure 5-12. Other robust transit schedules can be found in the Appendix C.2.

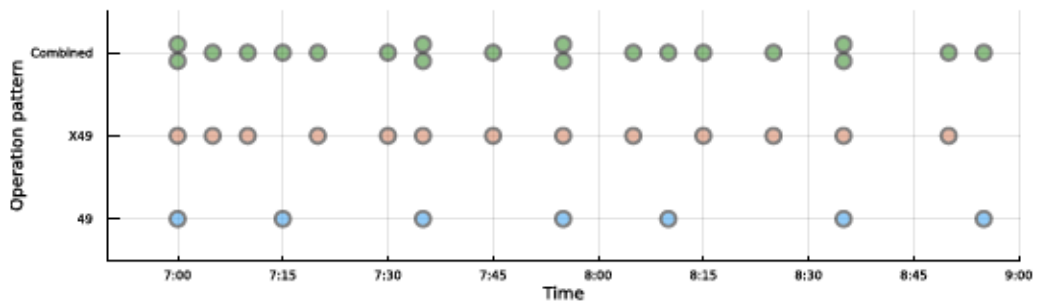


Figure 5-12: The robust transit schedule with $\Gamma = 10$.

Performance comparison results under surge demand scenarios are shown in Table 5.4. When considering more demand uncertainty (having a larger value of Γ), robust transit schedules reduce both passengers' wait times and travel times compared to the stochastic and the current transit schedules. The robust transit schedule has a performance edge over the benchmark stochastic transit schedule under surge demand scenarios. The largest improvement over the wait time is 2.94% when $\Gamma = 9$, while the largest improvement over the in-vehicle travel time is 1.27% when $\Gamma = 3$. As demand becomes more difficult to predict in transit systems given people's working arrangements become more flexible, using RO techniques to generate transit schedules improves the level of service for transit networks.

Compared to the stochastic transit schedule shown in Figure 5-11, the robust transit schedule utilizes one more bus over pattern X49. Meanwhile, more buses are dispatched during the first hour from the terminal. In summary, the robust transit schedule has a better performance than the benchmark stochastic transit schedule, especially under surge demand scenarios. Under normal demand scenarios, robust transit schedules can be adopted when vehicles are crowded and passengers prefer less in-vehicle travel times. The uncertain parameter Γ in the model needs to be selected carefully to reflect the actual demand uncertainty. Advanced data-driven robust optimization approach with the ability to automatically select uncertain parameter Γ can be further introduced [18].

Sensitivity analyses of the heuristic parameter ϵ are shown in Figure 5-13. Y-axis on the left represents the number of distinct passenger flows in $\mathcal{F}$, i.e., $|\mathcal{F}|$. Y-axis on the right indicates the percentage of unsatisfied passengers. When the value of ϵ increases, the number of passenger flows has an exponential decrease. With fewer passenger flows considered, the robust counterpart introduces fewer constraints and variables, therefore, robust transit frequency setting problems are easier to solve. On the other hand, fewer passenger flows lead to more unsatisfied passengers with the optimized transit schedule. Regarding the percentage of unsatisfied passengers for the optimized schedule, it indicates that some passengers are not able to board a transit vehicle which is departed from the terminal station during the studied time period.

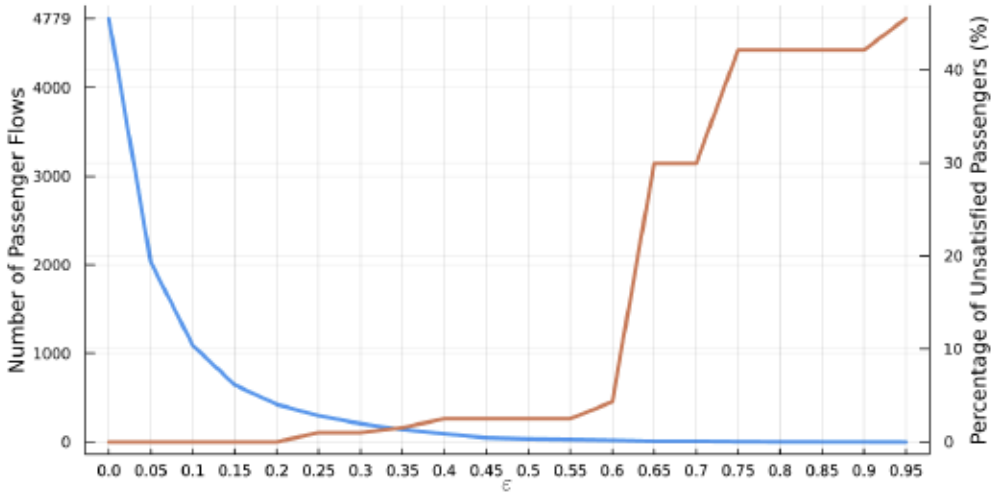


Figure 5-13: Sensitivity analyses for parameter ϵ in the heuristic-based dimensionality reduction approach.

In practice, unsatisfied passengers suffer longer wait times as they can board vehicles that depart from the terminal station later. In summary, robust transit schedules generated with a higher value of ϵ lead to excessive wait times by passengers. This sensitivity analysis echoes the Proposition 2 where the objective loss monotonically increases when ϵ increases.

5.4.4 Model Summary

Lastly, we summarize the computational efficiency of the proposed models with transit downsizing to illustrate the feasibility of using the TFSP in practice. The summary is shown in Table 5.5. n_{row} indicates the number of rows (or constraints), n_{con} represents the number of continuous variables, and n_{int} denotes the number of integer variables. *Time* represents the computational time for each model, where the time for robust TSFP is the average computational time of all robust models.

Table 5.5: Model Summary for Computation Efficiency.

Model	n_{row}	n_{con}	n_{int}	Time
Baseline TFSP	11745	38150	96	0.99
Baseline TFSP with Crowding	17361	38150	5712	1.43
Stochastic TFSP (22 scenarios)	352293	7393034	96	161.66
Robust TFSP	8222661	12462778	96	8932

The proposed model demonstrates computational efficiency in scenarios where demand uncertainty is not a factor. However, when accounting for demand uncertainty, the problem's complexity significantly rises, introducing numerous constraints and continuous variables into the model. When comparing the Stochastic TFSP and Robust TFSP models, the robust variant substantially amplifies the number of constraints, resulting in significantly extended computation times and more challenging problem-solving tasks. Fortunately, both stochastic and robust models do not augment the number of integer variables. This characteristic preserves the feasibility of solving the problem within a limited set of branching possibilities.

5.5 Conclusions and future work

In this chapter, two major issues are addressed when generating transit schedules: i) inherent demand uncertainties, and ii) gigantic OD matrices. To protect transit schedules against demand variations, a robust TFSP model is proposed. To the best of the authors' knowledge, this chapter is the first to apply RO technique for solving TFSPs. A nominal optimization model is formulated to solve the TFSPs under a single transit line setting, and an extended model considering crowding levels on transit vehicles is proposed. To solve optimization problems efficiently given real-world transit instances, the TD approach is proposed based on the observation where transit demand matrices are sparse. We theoretically prove that the optimal objective function of the problem after TD is close to that of the original problem (i.e., the difference is bounded from above). A benchmark stochastic TFSP model is formulated as well to demonstrate the robust TFSP model performance. Real-world transit lines operated by CTA are used to test the performances of transit schedules generated with proposed models compared to the current transit schedule. Both stochastic and robust transit schedules reduce wait times and in-vehicle travel times simultaneously for passengers over multiple demand scenarios. Compared to benchmark stochastic schedules, robust schedules further reduce passengers' wait times and in-vehicle travel times when the level of demand uncertainty is large.

The main limitation of this study is using heuristics to solve the robust TFSP model without proof of optimality. Meanwhile, the parameter controlling the size of the uncertainty set needs to be selected manually. Future studies could develop methodologies for decreasing problem sizes while maintaining a certain level of optimality loss. Data-driven approaches can be introduced to automatically select the value of uncertain parameter Γ. Also, our demand data only provides the time information when passengers actually board transit vehicles or enter subway stations, knowing more time information (e.g., the deadline for passengers to arrive at their destinations) could further introduce passengers' time preferences into the model.

Another interesting research direction is pattern generation. Our model has the ability to select an optimal set of patterns to operate on a single transit line. However, how to generate a set of potential patterns for a single transit line can be a challenging task. Performances of different pattern generation algorithms can be evaluated through our proposed TFSP model. Meanwhile, other sources of uncertainty in transit systems can be considered when generating robust transit schedules, e.g., supply uncertainty (last-minute driver absence) and travel time uncertainty. Lastly, the proposed TFSP model can be extended to solve a network-level frequency setting problem with multiple transit lines.

Chapter 6

Design of Transit-Centric Multimodal Urban Mobility System with Autonomous Mobility-on-Demand

6.1 Introduction

The global population is increasingly urban-centric, with 51%, or 3.5 billion people, currently residing in cities. The proportion is projected to rise to 70%, or 6.3 billion people, by the year 2050. It's anticipated that by 2050, the distances traveled within urban areas will triple [101]. Meanwhile, urban mobility demand patterns are undergoing changes in the post-pandemic world, influenced by the increased prevalence of remote working. One of the most significant challenges urban areas are confronting is urban mobility. With urban areas witnessing both a rise and transformation in travel demand, the necessity for an analytical framework capable of generating a high-capacity urban mobility system becomes crucial. Such a system should not only be efficient but also minimize environmental impact.

Emissions from urban mobility systems are one of the main sources of global

warming as transportation currently accounts for 29% of U.S. Greenhouse Gas Emissions [158]. Building a sustainable urban mobility system is an essential strategy to address climate change-related issues. Emerging Mobility-on-Demand (MoD) services, such as ride-hailing platforms like Uber and Lyft, have introduced responsive and reliable travel options for individuals. However, public transit (PT) systems continue to serve as the foundation of sustainable urban mobility, facilitating efficient city-wide travel for large populations. Both MoD and PT have their unique advantages and challenges. MoD offers flexible and direct services to a few passengers at a higher cost, while PT provides cost-effective transportation for large groups of people. However, the fixed schedules of public transit and concerns about accessibility are significant considerations for passengers.

Autonomous Vehicles (AV) signify a transformative future for urban mobility, offering prospects to enhance the quality of public transportation. Autonomous Mobility-on-Demand (AMoD) systems, such as Waymo and Cruise, have demonstrated significant potential in providing reliable and efficient services to passengers in major U.S. cities. While some researchers contend that AMoD and PT are in competition [124], there is a growing consensus that the two systems could actually complement each other. This chapter introduces a novel framework for creating a *Transit-Centric Multimodal Urban Mobility (TCMUM)* system, marking a first step towards the complete integration of AMoD systems and PT networks. Within the TCMUM system we envision, AMoD serves as the cornerstone for first-mile-last-mile (FMLM) services, linking passengers to transit stations. This proposed TCMUM-AMoD system merges the advantages of both transit and Mobility-on-Demand, enhancing accessibility while preserving high capacity and efficient passenger movement.

The TCMUM-AMoD system provides passengers with three transportation options: rail, bus, and AMoD. The design of the TCMUM-AMoD system involves several key components: (i) setting the operational frequency for the rail network, (ii) designing the bus network and determining its frequency, (iii) allocating vehicles for the AMoD fleet, and (iv) determining pricing structures using the TCMUM-AMoD system. Leveraging the existing transit infrastructure, the implementation of the

TCMUM-AMoD system could offer numerous enhancements to urban mobility, including: (i) the substitution of infrequent bus services with AMoD fleets, improving accessibility, (ii) the enhancement of bus service frequencies along key transit corridors, reducing passenger wait times and elevating service levels, (iii) improved coordination among different modes of transportation, and (iv) a decrease in long-haul MoD trips, potentially alleviating traffic congestion and lowering carbon emissions.

This chapter presents a tractable optimization framework that is designed to simultaneously configure networks and set service prices, with the goal of minimizing the total disutility experienced by passengers within the system. The contribution of this chapter can be summarized as follows:

1. To the best of authors' knowledge, this work is the first attempt to propose a tractable optimization framework for solving the joint transit network design, fleet sizing, and pricing in the multimodal mobility system while considering passengers' mode and route choices.

2. An optimization model for solving the design of Transit-Centric Multimodal Urban Mobility with Autonomous Mobility-on-Demand (TCMUM-AMoD) is proposed, where the sharing scenarios of AMoD services are modelled explicitly and passengers' mode and route choice behaviors are captured by discrete choice models.

3. The proposed optimization model is challenging to solve due to the non-linearity brought by discrete choice models and the growth of feasible combinations of modes and routes. A first-order approximation algorithm is introduced to solve the problem at scale.

4. The suggested optimization framework is assessed through a real-world case study in Chicago. It evaluates two types of demand – local and downtown – to demonstrate the optimal system design across various demand scenarios.

The remainder of the chapter is organized as follows. Section 6.2 reviews the relevant literature. Section 6.3 describes the problem and the optimization model for

TCMUM-AMoD, and proposes the first-order approximation algorithm for solving the problem at scale. Section 6.4 outlines experimental setups, data preparation, and displays experiment results and sensitivity analyses. Finally, Section 6.5 recaps the main contributions of this work, outlines the limitations, and provides future research directions.

6.2 Literature Review

6.2.1 Transit network design problem

The Transit Network Design Problem (TNDP) is first proposed by Ceder and Wilson [43], which can be separated into sub-problems ranging across tactical, strategical, and operational decisions, including Transit Network Design (TND), Frequency Setting (FS), Transit Network Timetabling (TNT), Vehicle Scheduling Problem (VSP), and Driver Scheduling Problem (DSP). Thorough reviews of TNDP and its sub-problems can be found in Ceder [42] and Ibarra-Rojas et al. [83].

This chapter addresses the challenge of transit network design by tackling the frequency setting problem. This involves determining the optimal number of trips for a specified set of transit lines to ensure a high level of service within a planning period. The Transit Frequency Setting Problem (TFSP) is first studied by Newell [129] using analytical models, where total passenger waiting time is minimized under a fixed passenger demand setting. Furth and Wilson [55] modelled the TFSP with a non-linear program that maximized the overall social welfare considering responsive demand. A heuristic-based algorithm was proposed to solve the non-linear program. Verbas and Mahmassani [161] extended the model proposed by Furth and Wilson [55] by considering multiple service patterns when determining transit frequencies, where a service pattern corresponds to a unique sequence of stops that are served by transit vehicles.

Recently, demand uncertainty, which is intensified by the remote work [10, 36, 37], has been considered when setting transit frequencies. Gkiotsalitis et al. [58] studied

the frequency setting problem for autonomous minibuses considering demand uncertainty. They utilized a traditional Stochastic Optimization (SO) method, Sample Average Approximation (SAA), to address the demand uncertainty. Guo et al. [69] utilized the Robust Optimization (RO) approach for solving the TFSP under a single transit line setting with demand uncertainty. A heuristic-based approach, namely Transit Downsizing (TD), was proposed to solve the large-scale real-world TFSP efficiently.

6.2.2 Mobility-on-Demand system

Ride-hailing platforms such as Uber and Lyft provide Mobility-on-Demand (MoD) services to millions of users globally every day. For an extensive overview of the ride-hailing system, one can refer to the review conducted by Wang and Yang [165]. Studies on the ride-hailing system consists of analyzing demand (customers) [186, 166], examining supply (drivers) [75, 40, 68], developing market structures [46, 190, 167, 70], and designing operational strategies for platforms. The operational strategies typically include dynamic pricing [8, 110], customer-driver matching [3, 19, 150], and vehicle rebalancing [172, 67, 71, 72].

Moreover, advancements in autonomous driving technology have introduced a new paradigm in transportation: Autonomous Mobility-on-Demand (AMoD) [188]. The AMoD system has the potential to increase driver supply within the ride-hailing system while reducing service costs. Furthermore, the complete compliance of AMoD vehicles eliminates scenarios of driver rejections and leads to more efficient vehicle allocations. Iglesias et al. [84] introduced a Model Predictive Control (MPC) algorithm designed to optimizing rebalancing strategies, capitalizing on short-term demand forecasts through LSTM neural networks. Their proposed approach could significantly reduce the average customer wait time. Tsao et al. [155] presented an MPC approach to optimize vehicle routes in AMoD system for both vacant and occupied vehicles. Their proposed algorithm has the potential to substantially decrease the distance traveled by mobility providers, thereby diminishing the impact of AMoD platforms on urban congestion.

6.2.3 Multimodal mobility system

In recent years, a substantial body of research has illustrated the potential benefits of integrating MoD with traditional transit services to enhance urban mobility system. Shen et al. [141] proposed an integrated AV-PT system where high-demand bus routes are maintained, low-demand bus routes are repurposed, and shared AVs are introduced as a complement for first-mile service during morning peak hours. An agent-based simulation was utilized to evaluate the integrated system performance. Their study revealed that the integrated system could potentially improve service quality, utilize road resources more efficiently, and being financially sustainable. Wen et al. [171] proposed a systematic approach for the integration of AV-PT, concentrating on the development of AV solutions that complement and enhance existing transit networks. An agent-based simulation platform was developed to evaluate service performance, complemented by a discrete choice model of demand. Their results showed that the integrated system can significantly enhance urban mobility systems by improved service availability, reduced operational costs, and enhanced accessibility. Salazar et al. [138] studied the integration and coordination of AMoD systems with public transit to enhance urban mobility. They proposed a network flow model that maximized social welfare by optimizing the allocation of autonomous vehicles and their interaction with existing transportation infrastructure. Their results showed that integrating AMoD fleets with PT can yield considerable advantages, including improved mobility, reduced congestion, lower emissions, and enhanced system efficiency.

Other studies have focus on the network design and system operation of such an integrated system. Luo et al. [114] proposed a framework for integrating micromobility services into transit network in order to connect packed urban centers to low-demand suburban areas, providing a low-cost, low-emission travel mode for short trips. They utilized a two-stage stochastic program to design the intermodal network considering demand uncertainty, with the first stage selecting transfer hub locations and the second stage optimizing system operations.

Steiner and Irnich [145] developed a strategic network planning optimization model that incorporates MoD offering first-and-last-mile services. Their model simultaneously optimized bus line configurations, identifies zones for MoD service deployment, establishes MoD interactions with fixed-route networks through transfer points, and optimizes passenger routes based on specified service levels. However, the proposed model did not generate detailed transit schedules, and passengers' route and mode choices are not fully modelled, where only route choice in the bus network is considered.

Pinto et al. [133] proposed a bi-level mathematical programming model for a joint system design problem with multimodal transit and shared AMoD. The upper level optimized transit network configurations and shared AMoD fleet sizes, while the lower level utilized an agent-based simulation to determine transit assignment and shared AMoD fleet operations with mode choice modeling. However, pricing for the integrated system is considered as an exogenous parameter.

Banerjee et al. [9] explored the development of efficient routing policies for smart transit systems. These systems integrated high-capacity vehicles like buses with a fleet of cars to optimize the routing of trip requests in real-time, aiming to maximize social welfare within a specified time window. Nonetheless, passengers' mode and route choice behaviors were not considered and only a line configuration of bus networks was generated.

Luo et al. [115] addressed the joint optimization problem of transit network design and pricing for multimodal mobility systems. They aimed to determine optimal settings for mass transit frequencies, flows of MoD services, and pricing for each trip to maximize social welfare. The solution method included a primal-dual approach, a decomposition framework, and an approximation algorithm to solve optimization of large-scale problem instances.

Wang et al. [169] proposed an analytical framework for designing a transit-oriented multi-modal transportation system with passengers' route choices. They introduced a system-state equilibrium model that accounts for travelers' rational choice behaviors across different transportation modes and the corresponding impact on service levels.

However, their study simplified the design of transit systems without generating a detailed network design and transit schedules.

Kumar and Khani [98] proposed a methodology framework for designing networks for an integrated system with MoD and transit. However, they did not allow shared rides, and transit lines are assumed to have unlimited capacity. Also, passengers' route and mode choice behavior is not considered. Their proposed method is demonstrated on small-sized networks (Sioux Falls).

Table 6.1: Research studies that solve the design of integrated MoD and PT system.

Paper	Transit decisions	MoD decisions	Pricing	Objectives	Demand modeling	Solution method
Luo et al. [114]	Location of transfer hubs	Movement of MoD vehicles	Yes	Maximize profit	Discrete choice model	Two-stage stochastic program with heuristic algorithm
Steiner and Irnich [145]	Transit line configuration	MoD zones and transfer points	No	Minimize total system cost	Route assignment in transit network	MILP + branch-and-price algorithm
Pinto et al. [133]	Transit network design and frequency setting	Fleet sizing	No	Minimize travelers' disutility	Mode choice and route assignment	Bi-level programming + simulation
Banerjee et al. [9]	Transit line configuration and frequency setting	Movement of MoD vehicles	No	Maximize system welfare	No	MILP + Approximation algorithm
Luo et al. [115]	Transit network design and frequency setting	Movement of MoD vehicles	Yes	Maximize system welfare	Route assignment	MILP + primal-dual approach + decomposition
Wang et al. [169]	Uniform stop distance and headway	Fleet sizing and allocation	Yes	Maximize social welfare	No	System-state equilibrium model + search algorithm
Kumar and Khani [98]	Transit line configuration and frequency setting	Fleet sizing and allocation	No	Minimize travelers' cost	No	MILP + Benders decomposition
This study	Transit network design and frequency setting	Fleet sizing and allocation	Yes	Minimize travelers' disutility	Discrete choice model	MINLP + first-order approximation

Table 6.1 provides a summary of previous studies on the development of integrated MoD and PT systems. To the authors' knowledge, this work represents the first attempt to jointly solve network design and frequency setting for the transit system, fleet sizing and allocation for the MoD system, and pricing strategies for the combined system, all while taking into account passengers' mode and route choices. The proposed framework is motivated by Bertsimas et al [21], where a joint frequency-setting and pricing optimization problem is solved on a multimodal transit networks at scale. We further extend their methodology into the context of integrated PT and MoD system.

6.3 Methodology

6.3.1 Problem description

Integrating the AMoD system into the existing transit network introduces a range of possible organizational structures. These structures can vary based on several factors, including the ownership of the transit and AMoD services, how these operators interact with each other, and the extent of regulation imposed by public authorities [141]. In this chapter, we assume that the public transit agency is responsible for managing both transit services and AMoD operations. This setup involves the transit agency either owning the AMoD fleet outright or contracting with AMoD service providers. The purpose of operating such a transit-centric multimodal system for agencies is to leverage the AMoD fleet to enhance service delivery for passengers. This includes substituting low-demand bus lines with AMoD services and providing connections for passengers to and from rail stations.

With the existing transit network, our study focuses on the design of the TCMUM-AMoD system under the morning commute setting. Within this framework, we identify two distinct categories of commuters: *local* commuters and *downtown* commuters. Local commuters primarily rely on the transit system for short-distance trips within their local area, generally utilizing bus services. On the other hand, downtown commuters require services for longer-distance travel from suburban areas to downtown, typically facilitated through rail services or express bus services, to accommodate their commute needs efficiently. In the TCMUM-AMoD system, local commuters have two available mode options:

1. **Bus**: local commuters take bus services to their destinations.
2. **AMoD**: local commuters take AMoD services directly from their origins to their destinations.

For downtown commuters, they have three available mode options:

1. **Rail**: downtown commuters take rail services to their destinations.

2. **Bus+Rail**: downtown commuters utilize local bus services to reach rail stations, from which they then board rail services to travel to their final destinations.

3. **AMoD+Rail**: downtown commuters utilize AMoD services to reach rail stations, from which they then board rail services to travel to their final destinations.

The rail services can also be replaced by express bus services for downtown commuters. Under this system setting, the AMoD only provides local trips to commuters and commuters have better accessibility to the transit network.

In the design of the TCMUM-AMoD system, our approach involves the optimization of the rail and bus networks, the sizing and distribution of the local AMoD fleet, and the pricing structure for utilizing the AMoD system. The overarching goal of this optimization is to minimize the total disutility experienced by commuters within the system, which are quantified by waiting times and walking times.

Figure 6-1: Two route options for a morning local commute from home to the company.

Let $G = (V, E)$ denote the road network and a *commute* in this problem is referred to as an origin-destination pair (u, v), where $u, v \in V$. Let $\mathcal{U}$ indicate the set of commutes in the problem. A commute $(u, v) \in \mathcal{U}$ could have multiple *route* choices, denoted by the set $\mathcal{R}^{u,v}$, and each route corresponds to a distinct sequence of mode and path choices. Each route $r \in \mathcal{R}^{u,v}$ contains at least one *leg*, indicating a trip stage along path r. Let $\mathcal{J}(r)$ indicate the set of legs in route r. Figure 6-1 and Figure 6-2 show instances with route options for a local commuter and a downtown commuter, respectively. Local commuters can: i) walk to the bus stop, take the bus and walk to the company, ii) take an AMoD service directly from home to the

company. Downtown commuters can: i) walk to the subway station, take the subway and walk to the company, ii) take a bus service to the subway station, take the subway and walk to the company, and iii) take an AMoD service to the subway station, take the subway and walk to the company.

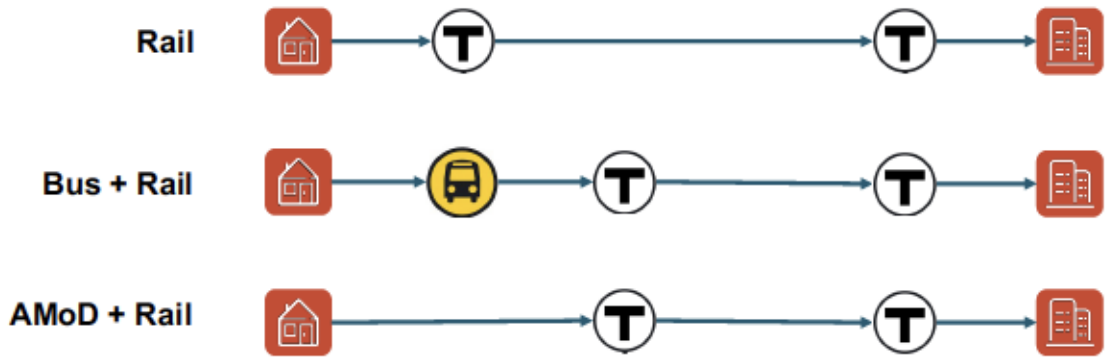


Figure 6-2: Three route options for a morning downtown commute from home to the company.

The full study time period $[T_{start}, T_{end}]$ is divided into T identical time intervals with length Δ_T. Let $\mathcal{T} = \{1, 2, ..., T\}$ stand for the set of time intervals. The commute demand data is indicated by the OD-matrix $\boldsymbol{d} = (d_t^{u,v})$ where commute (u, v) starts from origin u go to destination v at time t. Given the commute demand at time $d_t^{u,v}$, commuters make route choices $\boldsymbol{\theta} = (\theta_t^{u,v,r})$ among a set of routes, where $\theta_t^{u,v,r}$ represents the proportion of commuters for commute (u, v) at time t who chooses route r.

For the multimodal transit network design, let $\mathcal{L}^R$ and $\mathcal{L}^B$ denote the set of rail and bus *lines*, where a line indicates a sequence of stops. Let $\mathcal{L}$ represent the set of transit lines, i.e., $\mathcal{L} = \mathcal{L}^R \cup \mathcal{L}^B$. For any transit line $l \in \mathcal{L}$, the decision variable is $\boldsymbol{x} = (x_t^l)$, where $x_t^l \in \mathbb{Z}$ indicates the number of departures from the start of the line l at time t. A lower bound $\underline{B}_R$ and an upper bound $\overline{B}_R$ are imposed on the number of departures of rail lines to guarantee the minimum level of service and the minimum headway between two consecutive trains. Similarly, an upper bound $\overline{B}_B$ is enforced to the number of departures of bus lines for the purpose of maintaining the minimum headway condition. Setting decision variable x_t^l equal to 0 is equivalent to removing the bus line l from the transit network at time t. Given the total budget B_{rail} and

B_{bus} for the transit network, which denotes the total number of rail and bus services that can be offered, the set of feasible transit networks is denoted by

$$\mathcal{X} = \left\{ (\boldsymbol{x}^R, \boldsymbol{x}^B) \in \mathbb{Z}_+^{T \times |\mathcal{L}|} : \sum_{t \in \mathcal{T}} \sum_{l \in \mathcal{L}^B} c^l x_t^l \leq B_{bus}, \sum_{t \in \mathcal{T}} \sum_{l \in \mathcal{L}^R} c^l x_t^l \leq B_{rail}, \underline{B}_R \leq \boldsymbol{x}^R \leq \overline{B}_R, \boldsymbol{x}^B \leq \overline{B}_B \right\}, \tag{6.1}$$

where $\boldsymbol{x}^R$ and $\boldsymbol{x}^B$ are vectors of decision variables for rail and bus lines, and c^l stands for the cost for line l, which varies across different transit lines. Instead of solving for a detailed transit schedule, we relax the feasibility constraint (1) to generate a set of departure rates of transit services during each time interval:

$$\bar{\mathcal{X}} = \left\{ (\boldsymbol{x}^R, \boldsymbol{x}^B) \in \mathbb{R}_+^{T \times |\mathcal{L}|} : \sum_{t \in \mathcal{T}} \sum_{l \in \mathcal{L}^B} c^l x_t^l \leq B_{bus}, \sum_{t \in \mathcal{T}} \sum_{l \in \mathcal{L}^R} c^l x_t^l \leq B_{rail}, \underline{B} \leq \boldsymbol{x}^R \leq \overline{B}, \boldsymbol{x}^B \leq \overline{B} \right\}. \tag{6.2}$$

The relaxed feasible schedules $\bar{\mathcal{X}}$ are straightforward to interpret and more flexible for transit operators to follow. For instance, $x_t^l = 1.5$ indicates that line l departs twice every three time intervals. Also, the relaxed feasibility set only requires solving a Linear Programming (LP) problem.

The AMoD system offers both *local* and *first-mile-last-mile (FMLM)* trips to commuters. The local AMoD trip indicates that local commuters take an AMoD trip directly from their origins to their destinations. The FMLM trips connect downtown commuters' with rail stations. In this chapter, we assume the existence of a fleet of vehicles tasked with providing both direct and FMLM AMoD services. Our primary focus is on optimizing the sizing and allocation of this vehicle fleet and determining the pricing strategies for using these services.

Let $\mathcal{S}$ denote the set of rail stations in the transit system. The decision variable for the fleet size of AMoD vehicles is $\boldsymbol{N} = (N_t^s)$, where $N_t^s \in \mathbb{Z}_+$ indicates the number of AMoD vehicles within the *nearby* region of rail station $s \in S$ at time t. The nearby region of each rail station s is predefined and has area A_s. Given the maximum number of AMoD vehicles $\overline{N}$, the set of feasible AMoD vehicle allocations is indicated by

$$\mathcal{N} = \left\{ \boldsymbol{N} \in \mathbb{Z}_+^{|\mathcal{S}|} : \sum_{s \in \mathcal{S}} N_t^s \leq \overline{N}, \forall t \in \mathcal{T} \right\}. \tag{6.3}$$

We further relax the feasible vehicle allocation set $\mathcal{N}$ to reduce the computation complexity:

$$\bar{\mathcal{N}} = \left\{ \boldsymbol{N} \in \mathbb{R}_+^{|\mathcal{S}|} : \sum_{s \in \mathcal{S}} N_t^s \leq \overline{N}, \forall t \in \mathcal{T} \right\}. \tag{6.4}$$

The feasible allocation of vehicles around rail station s at time t can be obtained by rounding down $N_t^s \in \bar{\mathcal{N}}$.

6.3.2 Optimization with known path choices and non-shared local AMoD fleet

First, we formulate the optimization problem with known path choices $\boldsymbol{\theta}$ (satisfying $0 \leq \theta_t^{u,v,r} \leq 1$ and $\sum_{r \in \mathcal{R}^{u,v}} \theta_t^{u,v,r} = 1, \forall t \in \mathcal{T}$) and an AMoD fleet which only offers non-shared services to commuters. Both assumptions will be revisited and relaxed in the following sections.

Let (u, v, r) denote a *commute route* for route r which serves commute (u, v), and (u, v, r, i) indicate a *commute-route leg* corresponds to the i-th leg of the itinerary of commute route (u, v, r). Let $\boldsymbol{z} = (z_t^{u,v,r,i})$ denote the boarding variables for the commute-route leg (u, v, r, i), where $z_t^{u,v,r,i} \in \mathbb{R}_+$ indicates the number of commuters traveling from u to v, choosing route r, on the i-th leg of their itinerary, boarding a transit (or AMoD) service that starts (or is available) at time t. For the simplicity of the model formulation, we assume that the travel time on any trip leg is zero[1]. The system design of the TCMUM-AMoD system can be formulated as

$$\min_{\boldsymbol{x} \in \bar{\mathcal{X}}, \boldsymbol{N} \in \bar{\mathcal{N}}} Q(\boldsymbol{x}, \boldsymbol{N}, \boldsymbol{\theta}), \tag{6.5}$$

where

$$Q(\boldsymbol{x}, \boldsymbol{N}, \boldsymbol{\theta}) = \min_{\boldsymbol{z} \geq 0} \quad J_{Transit}(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{x}) + J_{AMoD}(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{N}) \tag{6.6a}$$

$$\text{s.t.} \quad O_{l,s,t}(\boldsymbol{z}) \leq K^l x_t^l, \quad \forall l \in \mathcal{L}, \forall s \in \mathcal{S}_l, \forall t \in \mathcal{T}; \tag{6.6b}$$

[1]Incorporating the travel time for trip legs only requires adjusting time indices in the model, which significantly complicates the model formulation. It is straightforward to incorporate travel times in experiments.

$$P_{s,t}(\boldsymbol{z}) \leq \frac{\Delta_T}{\mathbb{E}[T_l^s]} N_t^s, \quad \forall s \in \mathcal{S}, \forall t \in \mathcal{T}; \tag{6.6c}$$

$$\sum_{t=1}^{\tau} z_t^{u,v,r,1} \leq \sum_{t=1}^{\tau} d_t^{u,v} \theta_t^{u,v,r}, \quad \forall u, v \in \mathcal{U}, \forall r \in \mathcal{R}^{u,v}, \forall \tau \in \mathcal{T}; \tag{6.6d}$$

$$\sum_{t=1}^{\tau} z_t^{u,v,r,i} \leq \sum_{t=1}^{\tau} z_t^{u,v,r,i-1}, \quad \forall u, v \in \mathcal{U}, \forall r \in \mathcal{R}^{u,v}, \forall i = 2, ..., |\mathcal{J}(r)|, \forall \tau \in \mathcal{T}. \tag{6.6e}$$

The functions $J_{Transit}(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{x})$ and $J_{AMoD}(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{N})$ in Equation (6.6a) indicate the total waiting and walking time of the TCMUM-AMoD system. In our model, the waiting time for commuters in transit systems consist of both the expected waiting time given the line frequency and the excess waiting time. The expected waiting time for a commute route (u, v, r) traveling at time t can be calculated as

$$\sum_{i \in \mathcal{A}(u,v,r)} z_t^{u,v,r,i} \cdot \frac{\Delta_T}{2 \cdot x_t^{l(i)}},$$

where $\mathcal{A}(u, v, r)$ indicates the set of transit legs in the commute route (u, v, r) and $l(i)$ represents the transit line using by the transit leg $i \in \mathcal{A}(u, v, r)$. The total expected waiting time for the transit system is

$$J_{Transit}^{exp-wait}(\boldsymbol{z}, \boldsymbol{x}) = \sum_{(u,v) \in \mathcal{U}} \sum_{r \in \mathcal{R}^{u,v}} \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{A}(u,v,r)} z_t^{u,v,r,i} \cdot \frac{\Delta_T}{2 \cdot x_t^{l(i)}}. \tag{6.7}$$

The Equation (6.7) is ill-defined as $x_t^{l(i)}$ could be zeros. However, Constraints (6.6b) guarantee that $z_t^{u,v,r,i}$ will be zero if $x_t^{l(i)}$ is zero. Therefore, we assume $z_t^{u,v,r,i} \cdot \frac{\Delta_T}{2 \cdot x_t^{l(i)}} = 0$ whenever $x_t^{l(i)} = 0$.

For the excess waiting time, the total number of commuters waiting at a station (or stop) s on line l at time τ can be computed by

$$W_{Transit}^{s,l,\tau}(\boldsymbol{z}, \boldsymbol{\theta}) = AD^{s,l,\tau}(\boldsymbol{\theta}) + XD^{s,l,\tau}(\boldsymbol{z}) - BD^{s,l,\tau}(\boldsymbol{z}), \tag{6.8}$$

where $AD^{s,l,\tau}(\boldsymbol{\theta})$ indicates "*arrival demand*", $XD^{s,l,\tau}(\boldsymbol{z})$ represents "*transferring de-*

mand", and $BD^{s,l,\tau}(\boldsymbol{z})$ corresponds to "*boarding demand*", all cumulative up until time τ. $W^{s,l,\tau}_{Transit}(\boldsymbol{z},\boldsymbol{\theta})$ represents the number of commuters who are at station s but not able to board line l at time τ. Then the total excess waiting time for the transit system can be obtained by aggregating Equation (6.8) as

$$J^{exc-wait}_{Transit}(\boldsymbol{z},\boldsymbol{\theta}) = \sum_{l\in\mathcal{L}}\sum_{s\in\mathcal{S}_l}\sum_{\tau\in\mathcal{T}}\left[AD^{s,l,\tau}(\boldsymbol{\theta}) + XD^{s,l,\tau}(\boldsymbol{z}) - BD^{s,l,\tau}(\boldsymbol{z})\right]\cdot\Delta_T, \tag{6.9}$$

where $\mathcal{S}_l$ represents the set of stations for line l.

The arrival demand $AD^{s,l,\tau}(\boldsymbol{\theta})$ in Equations (6.8) and (6.9) indicates the total arrival demand that has arrived at station s on line l by time τ and it is computed by

$$AD^{s,l,\tau}(\boldsymbol{\theta}) = \sum_{(u,v,r)\in\mathcal{K}(s,l)}\sum_{t=1}^{\tau} d_t^{u,v}\theta_t^{u,v,r}, \tag{6.10}$$

where the set $\mathcal{K}(s,l)$ denotes the set of commute routes that board on station s on line l first.

The transferring demand $XD^{s,l,\tau}(\boldsymbol{z})$ represents the total number of passengers who have arrived at station s on line l by time τ, having transferred over from another transit line. It can be formulated as

$$XD^{s,l,\tau}(\boldsymbol{z}) = \sum_{(u,v,r,i)\in\mathcal{H}(s,l)}\sum_{t=1}^{\tau} z_t^{u,v,r,i-1}, \tag{6.11}$$

where the set $\mathcal{H}(s,l)$ denotes the set of commute-route legs that make a transfer through station s on line l. For a commute-route leg (u,v,r,i) to be considered in the set $\mathcal{H}(s,l)$, it must satisfy the following criteria: i) $i \geq 2$, ii) the transfer station s connects the $(i-1)$-th and i-th legs of the commute route (u,v,r), and iii) commute-route leg i utilizes the transit line l.

The boarding demand $BD^{s,l,\tau}(\boldsymbol{z})$ indicates the total number of passengers who have managed to board a transit vehicle at station s on line l by time τ, which can be formulated as

$$BD^{s,l,\tau}(\boldsymbol{z}) = \sum_{(u,v,r,i)\in\mathcal{U}(s,l)}\sum_{t=1}^{\tau} z_t^{u,v,r,i}, \tag{6.12}$$

where the set $\mathcal{U}(s,l)$ represents all of the commute-route legs that require boarding a transit vehicle at station s on line l. This includes commute routes that start from station s on line l or later transfer to line l through station s.

For the walking time of commuters in transit systems, it can be formulated as

$$J_{Transit}^{walk}(\boldsymbol{z}) = \sum_{(u,v)\in\mathcal{U}} \sum_{r\in\mathcal{R}^{u,v}} \sum_{t\in\mathcal{T}} z_t^{u,v,r,1} \cdot w(u,v,r), \tag{6.13}$$

where $w(u,v,r)$ denotes the walking time for a commute route (u,v,r). And the total disutility of using the transit system is

$$J_{Transit}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{x}) = J_{Transit}^{exp-wait}(\boldsymbol{z},\boldsymbol{x}) + J_{Transit}^{exc-wait}(\boldsymbol{z},\boldsymbol{\theta}) + J_{Transit}^{walk}(\boldsymbol{z}). \tag{6.14}$$

The AMoD system provides commuters with local or FMLM services. The AMoD system provides door-to-door services to commuters, therefore, commuters do not experience walking time when using AMoD services. For the waiting time, it consists of the expected waiting time given the number of AMoD vehicles nearby and the excess waiting time. The expected waiting time for a commute route (u,v,r) traveling at time t can be calculated as

$$\sum_{i\in\mathcal{B}(u,v,r)} z_t^{u,v,r,i} \cdot \frac{\alpha_s}{\bar{v}} \sqrt{A_s/N_t^s},$$

where $\mathcal{B}(u,v,r)$ indicates the set of AMoD legs in the commute route (u,v,r), and $\frac{\alpha_s}{\bar{v}}\sqrt{A_s/N_t^s}$ is the expected wait time for AMoD services around station s [99]. $\bar{v}$ indicates the average local AMoD vehicle speed and α_s is a parameter depending on the shape of nearby region and the location of station s. The total expected waiting time for the AMoD system is

$$J_{AMoD}^{exp-wait}(\boldsymbol{z},\boldsymbol{N}) = \sum_{(u,v)\in\mathcal{U}} \sum_{r\in\mathcal{R}^{u,v}} \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{B}(u,v,r)} z_t^{u,v,r,i} \cdot \frac{\alpha_s}{\bar{v}} \sqrt{A_s/N_t^s}. \tag{6.15}$$

Similarly, Equation (6.15) is ill-defined as N_t^s could be zero. Constraints (6.6c)

guarantee that $z_t^{u,v,r,i}$ will be zero if N_t^s is zero. Therefore, we assume the term $z_t^{u,v,r,i} \cdot \frac{\alpha_s}{\bar{v}} \sqrt{A_s / N_t^s} = 0$ whenever $N_t^s = 0$.

For the excess waiting time of the AMoD system, the local AMoD trips originating near station s, the number of local commuters waiting to get AMoD trips at time τ can be formulated as

$$W_{Direct}^{s,\tau}(\boldsymbol{z}, \boldsymbol{\theta}) = \sum_{(u,v,r) \in \mathcal{Y}(s)} \sum_{t=1}^{\tau} \left(d_t^{u,v} \theta_t^{u,v,r} - z_t^{u,v,r,1} \right), \tag{6.16}$$

where the set $\mathcal{Y}(s)$ denotes the set of local commutes that begin in close proximity to station s and utilize AMoD vehicles to reach their destinations directly. Similarly, for the first-mile trips, the number of downtown commuters waiting to get first-mile AMoD services to a rail station s at time τ can be computed by

$$W_{First}^{s,\tau}(\boldsymbol{z}, \boldsymbol{\theta}) = \sum_{(u,v,r) \in \mathcal{M}(s)} \sum_{t=1}^{\tau} \left(d_t^{u,v} \theta_t^{u,v,r} - z_t^{u,v,r,1} \right), \tag{6.17}$$

where the set $\mathcal{M}(s)$ indicates the set of downtown commute routes that take AMoD services from their origins to the rail station s. The number of downtown commuters waiting at a rail station s at time τ to get last-mile AMoD services to their destinations can be formulated as

$$W_{Last}^{s,\tau}(\boldsymbol{z}) = \sum_{(u,v,r,i) \in \mathcal{N}(s)} \sum_{t=1}^{\tau} \left(z_t^{u,v,r,i-1} - z_t^{u,v,r,i} \right), \tag{6.18}$$

where the set $\mathcal{N}(s)$ denotes the set of commute-route legs that take AMoD services from station s to their destinations. Overall, the total excess waiting time for the AMoD system is

$$J_{AMoD}^{exc-wait}(\boldsymbol{z}, \boldsymbol{\theta}) = \sum_{s \in \mathcal{S}} \sum_{\tau \in \mathcal{T}} \left[W_{Direct}^{s,\tau}(\boldsymbol{z}, \boldsymbol{\theta}) + W_{First}^{s,\tau}(\boldsymbol{z}, \boldsymbol{\theta}) + W_{Last}^{s,\tau}(\boldsymbol{z}) \right] \cdot \Delta_T, \tag{6.19}$$

and the total disutility for the AMoD system is

$$J_{AMoD}(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{N}) = J_{AMoD}^{exp-wait}(\boldsymbol{z}, \boldsymbol{N}) + J_{AMoD}^{exc-wait}(\boldsymbol{z}, \boldsymbol{\theta}). \tag{6.20}$$

Constraints (6.6b) guarantee that the number of commuters on board a transit vehicle does not exceed the capacity K^l. $O_{l,s,t}(\boldsymbol{z})$ represents the occupancy of a transit vehicle, which starts to operate at time t, as it passes the station (or stop) s on line l. It is formulated as

$$O_{l,s,t}(\boldsymbol{z}) = \sum_{(u,v,r,i) \in \mathcal{I}(s,l)} z_t^{u,v,r,i}, \tag{6.21}$$

where the set $\mathcal{I}(s,l)$ incorporates commute-route legs that pass through the station s on line l. For a commute-route leg (u,v,r,i) to be considered in the set $\mathcal{I}(s,l)$, it must satisfy the following criteria: i) the i-th leg utilizes the transit line l, ii) the transfer station connecting the $(i-1)$-th and i-th legs is *at or before* station s on line l, and iii) the transfer station connecting i-th and $(i+1)$-th legs is *after* station s on line l.

Constraints (6.6c) ensure that the number of vehicles providing AMoD services does not exceed the number of available AMoD vehicles $\frac{\Delta_T}{\mathbb{E}[T_l^s]} N_t^s$ in the nearby region of station s at time t. Given that not every AMoD vehicle will be accessible to commuters at each time interval due to some vehicles being occupied with serving existing demand, we employ the expression $\frac{\Delta_T}{\mathbb{E}[T_l^s]}$ to approximate the average availability rate of AMoD vehicles near the station s. The ratio is constructed as follows: i) let the average local trip distance with both origins and destinations within the nearby region of station s be $\mathbb{E}[D_l^s] = \alpha_s \sqrt{A_s}$, while the value of α_s depends on the shape of nearby region and the location of station s [99]; ii) assume an average local AMoD vehicle speed of $\bar{v}$; iii) the average local AMoD trip time is $\mathbb{E}[T_l^s] = \frac{\mathbb{E}[D_l^s]}{\bar{v}}$; iv) the average availability rate of AMoD vehicles near the station s can then be formulated as $\frac{\Delta_T}{\mathbb{E}[T_l^s]}$.

$P_{s,t}(\boldsymbol{z})$ denotes the number of AMoD trips near the rail station s at time t, and

it can be computed by

$$P_{s,t}(\boldsymbol{z}) = \sum_{(u,v,r)\in\mathcal{Y}(s)\bigcup\mathcal{M}(s)} z_t^{u,v,r,1} + \sum_{(u,v,r,i)\in\mathcal{N}(s)} z_t^{u,v,r,i}. \tag{6.22}$$

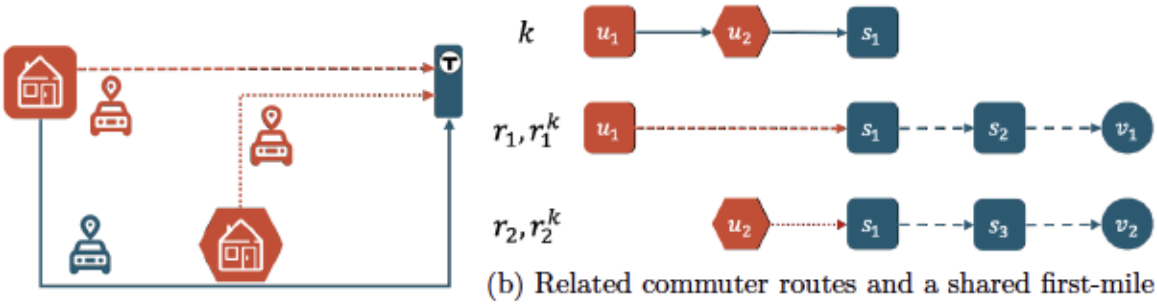
Constraints (6.6d) make sure that the number of commuters boarding the first leg of a commute route (u, v, r) up until time τ should not exceed the total demand for the commute route (u, v, r) up until time τ. Constraints (6.6e) impose that commuters with a commute route (u, v, r) can board i-th trip leg only if they complete $(i-1)$-th trip leg in the itinerary.

Both functions $J_{Transit}(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{x})$ and $J_{AMoD}(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{N})$ are nonlinear with respect to decision variables $\boldsymbol{x}$ and $\boldsymbol{N}$. Overall, the problem (6.5) is a nonlinear program that is intractable for large-scale instances. We will introduce a heuristic to linearize and solve the problem later.

6.3.3 Serving commuters with shared AMoD fleet

In this section, we consider scenarios with a shared AMoD fleet. After incorporating shared AMoD services into the problem, the same commute route including AMoD trips should be separated into multiple commute routes indicating different sharing scenarios. Figure 6-3 provides an instance for explaining the route separation. Trip k indicates a first-mile shared AMoD trip which is shared by commutes (u_1, v_1) and (u_2, v_2). Each commute contains both routes with non-shared AMoD trips r_1, r_2 and routes with shared AMoD trips r_1^k, r_2^k. Routes r_1 (r_2) and r_1^k (r_2^k) share the same itinerary but different first-mile AMoD services.

Let $\mathcal{P}$ denote the set of shared AMoD trips. For a shared trip $p \in \mathcal{P}$, let $\mathcal{R}(p)$ represent the set of commute routes that incorporates the shared trip p. Meanwhile, let $\mathcal{Q}(p)$ indicate the set of commutes that get involved in the shared trip p. For a commute $(u, v) \in \mathcal{Q}(p)$, let $\mathcal{R}^{u,v}(p)$ denote the set of routes for the commute (u, v) that includes the shared trip p. Then we introduce additional constraints to the



(a) First-mile AMoD scenarios

(b) Related commuter routes and a shared first-mile AMoD trip

Figure 6-3: Example explaining the route separation after introducing shared AMoD services.

optimization problem:

$$\sum_{r \in \mathcal{R}^{u,v}(p)} z_t^{u,v,r,i} = \sum_{r' \in \mathcal{R}^{u',v'}(p)} z_t^{u',v',r',i'} \quad \forall p \in \mathcal{P}, \forall (u,v), (u',v') \in \mathcal{Q}(p), \forall t \in \mathcal{T}, \quad (6.23)$$

where i and i' are indices of commute-route legs in routes r and r' that correspond to the shared trip p, respectively. Constraint (6.23) maintains the consistency of boarding variables among commute-route legs that correspond to the same shared trip.

When allowing shared AMoD services, the number of AMoD trips does not equal to the number of vehicles utilized for providing AMoD services. Therefore, we define $\tilde{P}_{s,t}(z)$ as the number of vehicles used for providing AMoD trips around rail station s at time t with sharing, which is formulated as

$$\tilde{P}_{s,t}(z) = \sum_{(u,v,r) \in \mathcal{Y}(s) \bigcup \mathcal{M}(s)} \xi^{u,v,r,1} z_t^{u,v,r,1} + \sum_{(u,v,r,i) \in \mathcal{N}(s)} \xi^{u,v,r,i} z_t^{u,v,r,i}, \quad (6.24)$$

where $\xi^{u,v,r,i}$ indicates the vehicle discount factor for a commute-route leg (u,v,r,i) which corresponds to a shared AMoD trip. The value of $\xi^{u,v,r,i}$ depends on the number of commutes sharing the same AMoD trip. For instance, $\xi^{u,v,r,i} = 0.5$ if the commute-route leg i corresponds to a shared AMoD trip with another commute (u',v'). In general, $\xi^{u,v,r,i} = \frac{1}{n}$ where n is the number of commutes included in the shared AMoD trip.

The optimization problem with shared AMoD fleet can then be formulated as

$$\min_{\boldsymbol{x}\in\bar{\mathcal{X}},\boldsymbol{N}\in\bar{\mathcal{N}}} \tilde{Q}(\boldsymbol{x},\boldsymbol{N},\boldsymbol{\theta}), \tag{6.25}$$

where

$$\begin{aligned}
\tilde{Q}(\boldsymbol{x},\boldsymbol{N},\boldsymbol{\theta}) = \min_{\boldsymbol{z}\geq 0} \quad & J_{Transit}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{x}) + J_{AMoD}(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{N}) && \text{(6.26a)}\\
\text{s.t.} \quad & \text{Constraints (6.6}b\text{), (6.6}d\text{), (6.6}e\text{), (6.23),} && \text{(6.26b)}\\
& \tilde{P}_{s,t}(\boldsymbol{z}) \leq \frac{\Delta_T}{\mathbb{E}[T_l^s]} N_t^s, \quad \forall s \in \mathcal{S}, \forall t \in \mathcal{T}. && \text{(6.26c)}
\end{aligned}$$

6.3.4 Optimization with design-dependent choices

In this section, we introduce a framework that not only facilitates the generation of the optimal design for a multimodal system but also enables the service operator to adjust pricing structures, denoted as $\boldsymbol{p}(\lambda) = (p^{u,v,r}(\lambda))$. Here, $p^{u,v,r}(\lambda)$ represents the price levied for a commute between points u and v via route r. This chapter primarily concentrates on the modification of a discount factor $\lambda \in [\underline{\lambda}, \bar{\lambda}]$, which is utilized for the AMoD services. Regarding the pricing of transit services and baseline AMoD services, it is presumed that their fares are predetermined and treated as exogenous parameters within our analysis.

For transit services, we assume a flat fare f^l for each transit line $l \in \mathcal{L}$ and a discount factor ν for transfers within the multimodal transit system. In practice, a transfer from the rail system to the bus system is typically free within a time period, i.e., $\nu = 0$ for taking the bus line.

For the baseline AMoD fare, we assume it follows the standard fare structure of Transportation Network Companies (TNCs), e.g., Uber or Lyft. The fare structure consists of a base fare f^{base}, a book fare f^{book}, a minimum fare f^{min}, a distance rate π_d and a time rate π_t. Given an AMoD trip with distance d and travel time τ, the price p will be

$$p^{AMoD}(d,\tau) = \max(f^{base} + f^{book} + \pi_d \cdot d + \pi_t \cdot \tau, f^{min}).$$

For a route $r \in \mathcal{R}^{u,v}$ of a commute (u,v), let $\mathcal{J}^{transit}(r)$ and $\mathcal{J}^{AMoD}(r)$ denote the set of legs in r which corresponds to commuters taking transit lines and AMoD vehicles, respectively. For each leg $j \in \mathcal{J}(r)$, let d_j represent the travel distance and τ_j indicate the travel time. For a transit leg $j \in \mathcal{J}^{transit}(r)$, let $l(j)$ stand for the transit line corresponding to the leg j. Let $\hat{l}(r)$ represent the first transit line in route r if the route r does not contain any AMoD legs. For a route r including AMoD legs, $\hat{l}(r) = \emptyset$ is an empty set. The price $p^{u,v,r}$ can then be formulated as

$$p^{u,v,r}(\lambda) = f^{\hat{l}(r)} + \sum_{j \in \mathcal{J}^{transit}(r) \setminus \{\hat{l}(r)\}} \nu f^{l(j)} + \sum_{j \in \mathcal{J}^{AMoD}(r)} \lambda \cdot p^{AMoD}(d_j, \tau_j). \tag{6.27}$$

It is worth noting that the pricing structures for both transit and AMoD operators can be adjusted seamlessly without changing the overall optimization model. Future extensions with more complicated and practical pricing structures can be considered.

By allowing the design-dependent choices for commuters, the path choice parameter will be modified as $\boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{N}, \lambda)$, indicating that commuters' path choices depend on transit frequencies $\boldsymbol{x}$, AMoD fleet size $\boldsymbol{N}$ and the discount factor λ. And the optimization with design-dependent choices can be formulated as

$$\min_{\boldsymbol{x} \in \bar{\mathcal{X}}, \boldsymbol{N} \in \bar{\mathcal{N}}, \lambda \in [\underline{\lambda}, \bar{\lambda}]} \tilde{Q}(\boldsymbol{x}, \boldsymbol{N}, \boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{N}, \lambda)). \tag{6.28}$$

The probability of selecting a specific route, denoted by $\boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{N}, \lambda)$, is influenced by the utility that a commuter derives from each available route option. This utility, represented by $\mu_t^{u,v,r}(\boldsymbol{x}, \boldsymbol{N}, \lambda)$, pertains to a commute (u,v), departing at time t and choosing route r, within the context of transit schedules $\boldsymbol{x}$, AMoD fleet allocations $\boldsymbol{N}$, and discount rate λ for AMoD services. It is postulated that the utilities associated with different route options are predominantly affected by two factors: i) *journey time*, and ii) *monetary cost*.

For a route $r \in \mathcal{R}^{u,v}$, the journey time consists of in-vehicle travel time, waiting time, and walking time. The in-vehicle travel time can be formulated as $\sum_{j \in \mathcal{J}(r)} \tau_j$. For a transit line l, the waiting time can be denoted as $\Delta_T / 2x_t^l$ assuming a uniformly distributed headway. The walking time τ_r^{walk} is predetermined for any route r as-

suming a constant walking speed $\bar{v}_{walking}$. For AMoD services, the waiting time for AMoD services around station s is $\frac{\alpha_s}{\bar{v}}\sqrt{A_s/N_t^s}$ [99].

For the monetary cost of route $r \in \mathcal{R}^{u,v}$, it is previously defined by $p^{u,v,r}(\lambda)$. Assuming that the commuter utility function is linear in time and cost attributes, the formulation of $\mu_t^{u,v,r}(\boldsymbol{x}, \boldsymbol{N}, \lambda)$ is given by

$$\begin{aligned}\mu_t^{u,v,r}(\boldsymbol{x}, \boldsymbol{N}, \lambda) &= -\beta_2 \cdot p^{u,v,r}(\lambda) \\ &- \beta_1 \left[\sum_{j \in \mathcal{J}^{transit}(r)} \left(\frac{\Delta_T}{2x_t^{l(j)}} + \tau_j \right) + \sum_{j \in \mathcal{J}^{AMoD}(r)} \left(\frac{\alpha_{s(j)}}{\bar{v}} \sqrt{\frac{A_{s(j)}}{N_t^{s(j)}}} + \tau_j \right) + \tau_r^{walk} \right], \quad (6.29)\end{aligned}$$

where β_1 stands for the marginal utility of time, β_2 represents the marginal utility of money, $l(j)$ denotes the transit line corresponding to transit leg j, and $s(j)$ corresponds to the station near the origin of the AMoD leg j.

Discrete choice models

With the systematic utility function for different route choices, we can establish a discrete choice model to calculate the design-dependent choices $\boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{N}, \lambda)$ based on utilities $\mu_t^{u,v,r}(\boldsymbol{x}, \boldsymbol{N}, \lambda)$. For the standard multinomial logit model [119, 11], a Gumbel-distributed random noise component $\varepsilon_t^{u,v,r}$ is attached to the utility function, i.e.,

$$\tilde{\mu}_t^{u,v,r}(\boldsymbol{x}, \boldsymbol{N}, \lambda) = \mu_t^{u,v,r}(\boldsymbol{x}, \boldsymbol{N}, \lambda) + \varepsilon_t^{u,v,r}, \quad (6.30)$$

and the probability for commuters to choose routes with the choice probabilities are given by

$$\boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{N}, \lambda) = \frac{\exp\left(\mu_t^{u,v,r}(\boldsymbol{x}, \boldsymbol{N}, \lambda)\right)}{\sum_{r' \in \mathcal{R}^{u,v}} \exp\left(\mu_t^{u,v,r'}(\boldsymbol{x}, \boldsymbol{N}, \lambda)\right)}. \quad (6.31)$$

The multinomial logit model (6.31) suffers from a property known as the independence from irrelevant alternatives (IIA) as utilities of different routes with identical transportation modes could share similar attributes. To address this issue, we propose a nested logit model, which consists of a two-level choice model: mode choice and route choice. Figure 6-4 displays the two-level decisions that commuters has to

make in the nested logit model. Commuters first choose the mode between AMoD-only mode, PT-only mode, and PT-AMoD mode. Under each mode, commuters then make a route choice given a set of available routes with the selective mode.

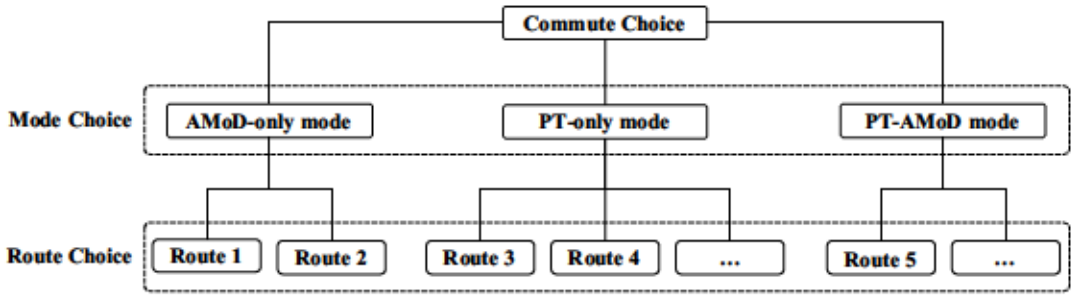


Figure 6-4: Multidimensional choices of the nested logit model.

For a set of route options $\mathcal{R}^{u,v}$ for a commute (u,v), let $\mathcal{R}_P^{u,v}$, $\mathcal{R}_A^{u,v}$ and $\mathcal{R}_{PA}^{u,v}$ represent the set of route options corresponds to PT-only mode, AMoD-only mode, and PT-AMoD mode, respectively. Then the route choice probability can be formulated as

$$\theta_t^{u,v,r}(\boldsymbol{x},\boldsymbol{N},\lambda) = \left[\frac{\exp(\phi_m \mu_t^{u,v,r}(\boldsymbol{x},\boldsymbol{N},\lambda))}{\sum_{r\in\mathcal{R}_m^{u,v}} \exp(\phi_m \mu_t^{u,v,r}(\boldsymbol{x},\boldsymbol{N},\lambda))}\right] \cdot \left[\frac{\exp(\phi I_m)}{\exp(\phi I_P)+\exp(\phi I_A)+\exp(\phi I_{PA})}\right], \tag{6.32}$$

where m corresponds to the travel mode of route r, i.e., $r \in \mathcal{R}_m^{u,v}$, and

$$I_{m'} = \frac{1}{\phi_{m'}} \ln\left(\sum_{r\in\mathcal{R}_{m'}^{u,v}} \phi_{m'} \mu_t^{u,v,r}(\boldsymbol{x},\boldsymbol{N},\lambda)\right), \quad \forall m' \in \{\mathrm{P},\mathrm{A},\mathrm{PA}\}, \tag{6.33}$$

where ϕ, ϕ_P, ϕ_A and ϕ_{PA} are parameters for the nested logit model that have to be estimated from commuter survey data in practice.

Solution algorithm

The problem (6.28) which solves the system design of the TCMUM-AMoD with design-dependent path choices is a nonlinear optimization problem given the nonlinear objective function (6.26a) and the nonlinear discrete choice model formulations (6.31) and (6.32). To solve this problem, we utilize a first-order approximation method proposed by Bertsimas et al. [21], where the nonlinear function is replaced with a

series of locally linear approximations. Given a feasible system design point $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{N}}, \tilde{\lambda})$, the nonlinear route choice probabilities can be approximated as

$$\hat{\theta}_t^{u,v,r}(\boldsymbol{x}, \boldsymbol{N}, \lambda; \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{N}}, \tilde{\lambda}) \approx \theta_t^{u,v,r}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{N}}, \tilde{\lambda}) + \bigtriangledown \theta_t^{u,v,r}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{N}}, \tilde{\lambda})'(\boldsymbol{x} - \tilde{\boldsymbol{x}}, \boldsymbol{N} - \tilde{\boldsymbol{N}}, \lambda - \tilde{\lambda}), \quad (6.34)$$

where the new system design point $(\boldsymbol{x}, \boldsymbol{N}, \lambda)$ is close to the initial point $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{N}}, \tilde{\lambda})$. By substituting the approximated path choice probability (6.34) into the optimization (6.28), we obtain a linear optimization problem. Given a feasible system design point $(\boldsymbol{x}^{(i-1)}, \boldsymbol{N}^{(i-1)}, \lambda^{(i-1)})$, a locally better system design can be solved with the problem

$$\min_{\boldsymbol{x}, \boldsymbol{N}, \lambda} \quad \tilde{Q}\left(\boldsymbol{x}, \boldsymbol{N}, \hat{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{N}, \lambda; \boldsymbol{x}^{(i-1)}, \boldsymbol{N}^{(i-1)}, \lambda^{(i-1)})\right) \quad (6.35a)$$

$$\text{s.t.} \quad \boldsymbol{x}^{(i-1)} - \boldsymbol{\rho} \leq \boldsymbol{x} \leq \boldsymbol{x}^{(i-1)} + \boldsymbol{\rho}, \quad (6.35b)$$

$$\boldsymbol{N}^{(i-1)} - \boldsymbol{\eta} \leq \boldsymbol{N} \leq \boldsymbol{N}^{(i-1)} + \boldsymbol{\eta}, \quad (6.35c)$$

$$\lambda^{(i-1)} - \sigma \leq \lambda \leq \lambda^{(i-1)} + \sigma, \quad (6.35d)$$

$$\boldsymbol{x} \in \bar{\mathcal{X}}, \quad (6.35e)$$

$$\boldsymbol{N} \in \bar{\mathcal{N}}, \quad (6.35f)$$

$$\underline{\lambda} \leq \lambda \leq \bar{\lambda}, \quad (6.35g)$$

where constant vectors $\boldsymbol{\rho}$, $\boldsymbol{\eta}$ and σ specify step sizes for line frequencies $\boldsymbol{x}$, local AMoD fleet allocations $\boldsymbol{N}$ and discount rate λ, respectively. For the nonlinear objective function (6.26a), given we are solving the problem iteratively, we approximate the objective function of iteration i by using system design decisions $(\boldsymbol{x}^{(i-1)}, \boldsymbol{N}^{(i-1)}, \lambda^{(i-1)})$ in iteration $i-1$. Therefore, the problem (6.35) becomes a linear program that can be solved efficiently.

The heuristic algorithm for solving the network design of the TCMUM-AMoD system with design-dependent path choices is described in Algorithm 1.

It is worth noting that the Algorithm 1 is guaranteed to output a locally optimal solution instead of the global optimum. To improve the quality of solutions, we generate multiple starting points $(\boldsymbol{x}^{(0)}, \boldsymbol{N}^{(0)}, \lambda^{(0)})$, run the approximated algorithm multiple times and select the best system design solution. Regarding the step size

Algorithm 1 First-order approximation algorithm for solving the system design of the TCMUM-AMoD with design-dependent path choices. **Input**: initial feasible system design point $(\boldsymbol{x}^{(0)}, \boldsymbol{N}^{(0)}, \lambda^{(0)})$, step size vectors $\boldsymbol{\rho}$, $\boldsymbol{\eta}$ and σ, termination threshold ϵ, maximum iteration κ. **Output**: locally optimal system design $(\boldsymbol{x}^*, \boldsymbol{N}^*, \lambda^*)$.

function FIRST-ORDER-APPROXIMATION($(\boldsymbol{x}^{(0)}, \boldsymbol{N}^{(0)}, \lambda^{(0)})$, $\boldsymbol{\rho}$, $\boldsymbol{\eta}$, σ, ϵ)
  $i \leftarrow 1$
  $\tilde{Q}_{prev} \leftarrow 0$
  while $i \leq \kappa$ **do**
    Solve problem (6.35) with a feasible point $(\boldsymbol{x}^{(i-1)}, \boldsymbol{N}^{(i-1)}, \lambda^{(i-1)})$, approximated objective functions, step sizes $\boldsymbol{\rho}, \boldsymbol{\eta}, \sigma$ and get the optimal solution $(\boldsymbol{x}^{(i)}, \boldsymbol{N}^{(i)}, \lambda^{(i)})$ and the objective value $\tilde{Q}^*$
    $threshold \leftarrow |\tilde{Q}^* - \tilde{Q}_{prev}|$
    if $threshold \leq \epsilon$ **then**
      break
    else
      $i \leftarrow i + 1$
      $\tilde{Q}_{prev} \leftarrow \tilde{Q}^*$
  return $(\boldsymbol{x}^{(i)}, \boldsymbol{N}^{(i)}, \lambda^{(i)})$

vectors $\boldsymbol{\rho}$, $\boldsymbol{\eta}$ and σ, they have to be chosen to balance the computation complexity and algorithm accuracy. A smaller step size leads to a more accurate local optimal solution, but it will take more iterations for the algorithm to converge. For the gradient $\bigtriangledown\theta_t^{u,v,r}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{N}}, \tilde{\lambda})$ in the equation (6.34), it can be computed using the automatic differentiation approach.

6.4 Numerical Experiments

In this section, we conduct numerical experiments on the Chicago transit network operated by the Chicago Transit Authority (CTA), which is one of the largest transit system in the North America. All models are implemented in Python programming language [159] and solved using Gurobi 10.0.2 [74]. All experimental results were generated on a machine with a 3.0 GHz AMD Threadripper 2970WX Processor and 128 GB Memory.

Before delving into the details of our numerical experiments, it is important to clarify that the aim of these experiments is not to offer policy recommendations

to transit authorities. Instead, the objective is to demonstrate the applicability of our proposed methodology using realistic data. Given the constraints of the data available to us, we employed the multinomial logit model (6.31) to simulate the route choice behavior of commuters. It should be noted that while the parameters used in this study may not be precise, they are considered sufficiently reasonable to provide valuable insights.

6.4.1 Data description

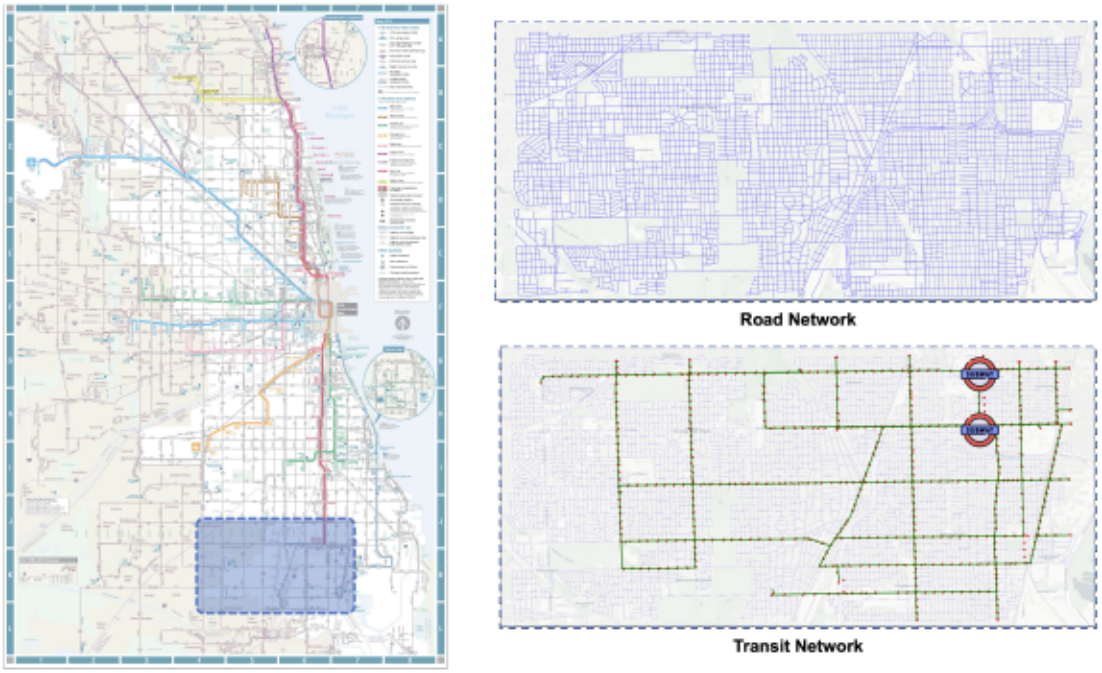


Figure 6-5: Road and transit networks for the study region.

Transit network data

The parameter values utilized in the experiments are presented in Table 6.2. The geographical focus of the experiments is the southern region of Chicago, characterized by several bus routes that interface and integrate with the Red line of the CTA's rail network. The road and transit networks under consideration are depicted in Figure 6-5. Blue region indicates the study region within the CTA network, blue lines represent

Table 6.2: Model parameters and values.

Parameter	Explanation	Value
Network Design Model Parameters		
T_{start}	Start time of planning period	06:00
T_{end}	End time of planning period	10:00
T	Number of time periods	48
Δ_T	Length of each time interval	5 (min)
$\|\mathcal{L}^B\|$	Number of bus lines	40
$\|\mathcal{L}^R\|$	Number of rail lines	1
$\|\mathcal{U}\|$	Number of morning commutes	2276
$\underline{B}_R$	Minimum number of departures for rail	0.5
B_R	Maximum number of departures for rail	2.5
B_B	Maximum number of departures for bus	1
B_{bus}	Number of available bus vehicles	814
B_{rail}	Number of available rail vehicles	94
$K^l, l \in \mathcal{L}^B$	Bus vehicle capacity	70
$K^l, l \in \mathcal{L}^R$	Rail vehicle capacity	640
$\|\mathcal{S}\|$	Number of rail stations	1
A	Area of the nearby region for rail station	90 (km^2)
α	Coefficient for approximating local trip distance	0.667
$\bar{v}$	Average vehicle speed	20 (mph)
δ_w	Maximum wait time for FMLM sharing AMoD trips	60 (seconds)
δ_d	Maximum delay time for FMLM sharing AMoD trips	60 (seconds)
Discrete Choice Model Parameters		
$\underline{\lambda}$	Minimum discount rate for AMoD services	0.1
$\bar{\lambda}$	Maximum discount rate for AMoD services	1
f^l	Fare for transit system	2.5 (dollars)
ν	Discount factor for transfers in transit	0
f^{base}	Base fare for AMoD services	1.87 (dollars)
f^{book}	Booking fare for AMoD services	1.85 (dollars)
f^{min}	Minimum fare for AMoD services	4.98 (dollars)
π_d	Distance fare rate for AMoD services	0.85 (dollars/mile)
π_t	Time fare rate for AMoD services	0.30 (dollars/minute)
$\bar{v}_{walking}$	Average walking speed	3 (mph)
$\beta_1(AMoD)$	Marginal utility of time in AMoD	16.3 (dollars/hour)
$\beta_1(transit)$	Marginal utility of time in transit	21.1 (dollars/hour)
β_2	Marginal utility of money	1
First-Order Approximation Algorithm Parameters		
ϵ	Termination threshold	0.1
κ	Maximum iteration	15
ρ $(rail)$	Step size for rail frequencies	0.1
η	Step size for AMoD allocations	10
σ	Step size for AMoD discount rate	0.1

road network, green lines denote bus network, red dots stand for bus stops, and rail symbols are rail stations.

Within the specified study area, there are 20 distinct bus routes: Routes 3, 4, 8A, 9, X9, 29, 34, 52A, 53A, 87, 95, 100, 103, 106, 108, 111, 111A, 112, 115, and 119. Given that each bus route operates in two directions, our analysis encompasses a total of 40 bus routes, i.e., $|\mathcal{L}^B| = 40$. The rail service included in this study is the Red line of the CTA transit network. Considering the focus on morning commute patterns, only the inbound direction of the Red line is taken into account for this analysis.

The analysis covers the morning hours from 6 AM to 10 AM, segmenting this time frame into 5-minute intervals, resulting in a total of 48 time periods within the model. For each bus route, it is assumed that there is a maximum of one departure per 5-minute interval. Regarding the rail service, the frequency of departures varies, with headways ranging from 2 to 10 minutes. This variation is designed to ensure both a minimum gap between consecutive vehicles and adherence to a baseline level of service.

The experiments leverage data collected over 20 weekdays in September 2019. The information pertaining to the current transit schedules is obtained from an open-source Generalized Transit Feed Specification (GTFS) dataset, which is regularly updated and released by the CTA on a monthly basis. The available number of bus and rail vehicles are calculated based on the transit schedule information. There are $B_{bus} = 814$ bus trips and $B_{rail} = 94$ rail trips during the 4-hour study period for transit lines in the model.

The travel times between any two stops, accommodating various service patterns, are computed using data from the Automatic Vehicle Location (AVL) dataset for September 2019, also provided by CTA. For transit vehicle capacity, it is posited that each bus is capable of accommodating up to 70 passengers. Similarly, for the rail system, each 8-car rail vehicle is designed to carry a maximum of 640 passengers.

Demand data

Given our study's focus on a specific subregion within the CTA's network, we categorize the passenger demand data into two distinct types based on their travel patterns. The local demand is defined as passengers whose trip origins and destinations are both located within the boundaries of the study area. Conversely, the downtown demand refers to passengers who begin their journeys within the study area and then utilize the Red Line service for traveling to their destinations outside the study area.

The demand data employed in the analysis is sourced from CTA's ODX dataset for September 2019. The "ODX" stands for the "origin, destination, and transfer inference algorithm," an advanced algorithm developed by Gabriel et al. [148] and currently implemented by the CTA. The CTA utilizes a fare collection system that requires passengers to "tap-on" but does not record their exit points, thereby not capturing alighting data. To bridge this gap, the ODX algorithm is employed to infer passengers' alighting points. Detailed insights into the creation and implementation of the ODX algorithm can be found in Sánchez-Martínez [148], Zhao et al. [198], Caros et al. [35].

In this chapter, we assume there exists a total commuter demand of 12,400 individuals throughout the studied period. This analysis incorporates the demand data for both local and downtown commuters as sourced from the CTA, to establish demand seeding matrices. These matrices serve as the foundation for generating the demand data used in the optimization model. In the process of constructing these demand seeding matrices, we calculate the demand for the rail system based on an average over 20 workdays within the month of September 2019. Conversely, the bus system's demand is derived from the data of a specific workday, namely September 5th, chosen due to the bus system's sparse demand patterns. Altogether, the dataset encompasses 2276 distinct commutes within the analysis, denoted as $|\mathcal{U}| = 2276$.

AMoD parameters

In the context of first-mile AMoD services, our study area includes two rail stations. Due to their proximity, these stations are consolidated into a single entity within our analysis. We determine the optimal fleet size for AMoD in the study region across each time interval. The study region is defined as a rectangular area, approximately $A = 90\ km^2$ in size. Following the methodology outlined in Larson and Odoni [99], we adopt a coefficient of $\alpha = 0.667$ for estimating the distance of local trips. Furthermore, the model assumes an average vehicle speed of $\bar{v} = 20\ mph$.

To reduce the computational complexity, we only allow sharing within FMLM AMoD trips. To generate sharing scenarios for the first-mile AMoD services, we assume that each vehicle is limited to being shared by two separate commuters. Additionally, it is crucial that each commuter within a shared route adheres to specified constraints regarding the maximum wait time (δ_w) and the maximum delay time (δ_d). For the purposes of this study, both the maximum wait time and the maximum delay time are established at 60 seconds, i.e., $\delta_w = \delta_d = 60$.

Discrete choice model parameters

For the transit system pricing, a uniform fare structure is implemented, with a flat fare of $f^l = 2.5$ dollars for using any transit line $l \in \mathcal{L}$. Additionally, transfers between transit lines are offered at no additional cost, i.e., $\nu = 0$.

In the context of the AMoD services pricing mechanism [31], the pricing model includes a base fare of $f^{base} = 1.87$ dollars, a booking fee of $f^{book} = 1.85$ dollars, and a minimum fare of $f^{min} = 4.98$ dollars. The fare structure also incorporates a distance-based rate of $\pi_d = 0.85$ dollars per mile and a time-based rate of $\pi_t = 0.30$ dollars per minute. For commuters who take AMoD services, a discount rate λ is applied, varying within a range from $\underline{\lambda} = 0.1$ to $\bar{\lambda} = 1$.

Regarding the parameters in the utility function as specified in Equation (6.29), with β_2 set to 1, β_1 represents the value of time in vehicles. For this analysis, the value of time spent in transit vehicles is set at 21.1 dollars per hour, whereas the

value of time in AMoD vehicles is determined to be 16.3 dollars per hour, based on findings from Hyland et al. [82]. The average walking time for commuters is set to be $\bar{v}_{walking} = 3$ mph.

Algorithm parameters

In the application of the first-order approximation algorithm to solve the TCMUM-AMoD system design problem, we initiate the process with 15 randomly selected starting solutions. The algorithm is set to terminate under one of two conditions: either after $\kappa = 15$ iterations or when the objective values of two successive iterations are within a tolerance of $\epsilon = 0.1$.

Given the relatively low demand observed for bus routes, the decision variables related to bus schedules are treated as integers. This approach is adopted because employing continuous variables for these schedules often results in very small, practically negligible values, essentially equivalent to discontinuing the bus route. Consequently, continuous step sizes are reserved exclusively for rail decision variables, with a specified step size of $\rho(rail) = 0.1$.

For the allocation of AMoD fleet, the step size is set at $\eta = 10$, reflecting adjustments in the number of AMoD vehicles allocated. Similarly, the step size for modifying the AMoD service discount rate is $\sigma = 0.1$.

6.4.2 Model results

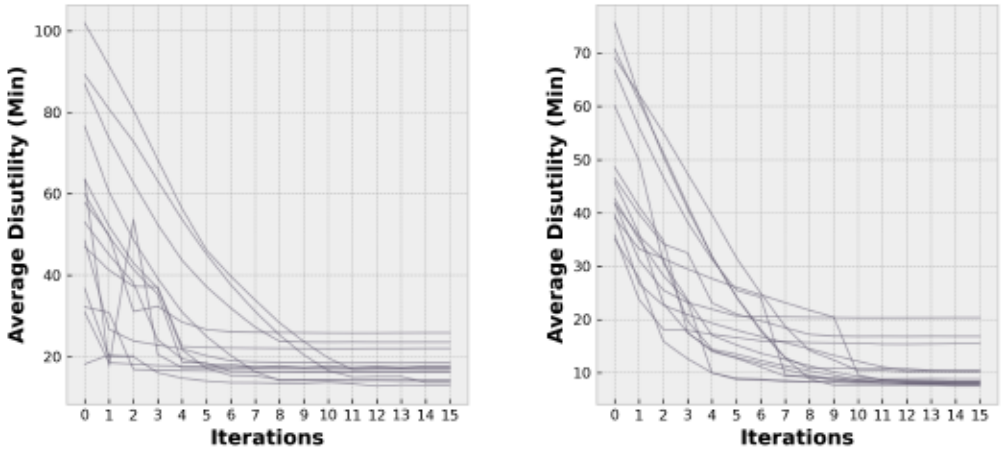
In the experiments, we would like to understand the trade-offs between different number of buses and AMoD vehicles under different demand profiles. Therefore, we adjust three parameters in the experiments: i) the proportion of available bus runs γ, ii) the number of available AMoD vehicles $\bar{N}$, and iii) the proportion of downtown commuters ψ.

In the baseline scenario, we assume that the percentage of available bus runs γ and the number of available AMoD vehicles $\bar{N}$ follows two relationships: i) passenger car equivalence (PCE), and ii) capital cost equivalence (CCE). For instance, if we

are removing 20% of bus runs within $B_{bus} = 814$ total bus runs during the four-hour study period, it is equivalent to remove around $\frac{814 \times 20\%}{4} \approx 41$ buses assuming each bus makes one run every hour. Then, the passenger car equivalence leads to 82 AMoD vehicles as the passenger car unit (PCU) for bus is 2.0 [153]. Regarding the capital cost equivalence, 41 buses is equivalent to 164 AMoD vehicles, given the cost for a E-bus around $\$800,000$ [118] the cost for an AMoD vehicle around $\$200,000$ [123].

For the proportion of downtown commuters, we assume the baseline number to be $\psi = 80\%$, reflecting commuters' demand pattern in CTA. The demand profile used in the optimization model is generated as follows: i) generate the downtown demand with a demand level of $12,400 \cdot \psi$, ii) generate the local demand with a demand level of $12,400 \cdot (1 - \psi)$, iii) combine local and downtown demand as the final demand profile.

Figure 6-6 shows the convergence performance for the proposed first-order approximation algorithm under two scenarios. Although some randomly-generated initial starting points lead to local optimal solutions, the proposed algorithm with 15 initial starts is capable of generating satisfying system design solutions. Also, 15 iterations is enough for the proposed algorithm to converge.



(a) Scenarios with 80% buses, 82 AMoD vehicles, and 80% downtown commuters. (b) Scenarios with 80% buses, 164 AMoD vehicles, and 80% downtown commuters.

Figure 6-6: Convergence of the first-order approximation algorithm.

Table 6.3 presents the findings from the baseline experiment, focusing on a scenario where 80% of commuters are heading downtown ($\psi = 0.8$), within the context of passenger car equivalence. The term *Avg disutility* refers to the average disutility for commuters, which is a composite measure including walking time, expected waiting time, and excess waiting time. *Avg walking* provides the average walking duration for each commuter, while *Avg waiting* quantifies the average expected wait time. *Avg utility* reflects the average utility for chosen paths by commuters. λ^* represents the optimal discount rate applied to AMoD service usage. The table also shows the mode share among downtown and local commuters, with L and DT signifying local and downtown commuters, respectively. *Unserved%* highlights the proportion of commuters who were not served by the end of the study period, remaining in wait for the next available vehicle.

The percentage of selected bus routes in the optimally designed system is captured by *Line utilization*, whereas *AMoD activation* indicates the average activation rate of AMoD vehicles throughout the studied period. Given the availability of an AMoD fleet, system operators have the strategic option to activate only a portion of the fleet to prevent system-wide congestion associated with the use of AMoD services. Activating more AMoD vehicles can decrease waiting times for these services, making them more attractive to commuters according to their behavior choice models.

Table 6.3: Experimental results for baseline scenario with 80% downtown commuters ($\psi = 0.8$) under PCE.

		System Performance Indicators						
γ	$\bar{N}$	Avg disutility	Avg walking	Avg waiting	Avg utility	Line utilization	AMoD activation	λ^*
100%	0	8.88	1.97	2.15	-6.00	90%	-	-
80%	82	14.15	1.48	1.96	-6.20	83%	99%	1.0
60%	164	9.46	1.32	1.85	-6.41	76%	94%	1.0
40%	246	10.06	1.17	1.78	-6.53	73%	95%	1.0
20%	328	11.32	1.03	1.60	-6.60	49%	100%	1.0
0%	410	13.08	0.82	1.35	-6.74	-	100%	1.0
		Mode Share for Downtown and Local Commuters						
γ	$\bar{N}$	AMoD% (L)	Bus% (L)	Unserved% (L)	AMoD+rail (DT)	Bus+rail (DT)	Rail (DT)	Unserved% (DT)
100%	0	0%	100%	22.4%	0%	26%	74%	0.3%
80%	82	53%	47%	10.6%	9%	22%	69%	7.3%
60%	164	61%	39%	0%	16%	18%	66%	4.1%
40%	246	74%	26%	0%	21%	13%	66%	5.0%
20%	328	88%	12%	0%	25%	8%	67%	6.3%
0%	410	100%	0%	0%	33%	0%	67%	7.9%

Under the passenger car equivalence setting, the replacement of buses with AMoD

Table 6.4: Experimental results for baseline scenario with 80% downtown commuters ($\psi = 0.8$) under CCE.

		System Performance Indicators						
γ	$\bar{N}$	Avg disutility	Avg walking	Avg waiting	Avg utility	Line utilization	AMoD activation	λ^*
100%	0	8.88	1.97	2.15	-6.00	90%	-	-
80%	162	7.48	1.39	1.98	-6.33	66%	88%	1.0
60%	328	8.30	1.31	1.94	-6.41	56%	61%	1.0
40%	492	9.05	1.19	1.93	-6.55	71%	100%	1.0
20%	656	11.26	1	1.56	-6.64	49%	66%	1.0
0%	820	13.03	0.82	1.28	-6.73	-	100%	1.0
		Mode Share for Downtown and Local Commuters						
γ	N	AMoD% (L)	Bus% (L)	Unserved% (L)	AMoD+rail (DT)	Bus+rail (DT)	Rail (DT)	Unserved% (DT)
100%	0	0%	100%	22.4%	0%	26%	74%	0.3%
80%	162	55%	45%	0.3%	14%	21%	66%	3.0%
60%	328	64%	36%	0.1%	16%	18%	66%	3.6%
40%	492	75%	25%	0%	19%	16%	65%	4.2%
20%	656	89%	11%	0%	26%	8%	67%	6.4%
0%	820	100%	0%	0%	33%	0%	67%	7.8%

vehicles does not influence the traffic condition. Therefore, the fixed travel time assumption used in this paper holds.

Table 6.4 presents the results for scenarios featuring 80% downtown commuters ($\psi = 0.8$) under the CCE setting. While the CCE does not preserve the fixed travel time assumption, it offers transit agencies a financially viable pathway for deploying the integrated system. Insights from numerical results under PCE and CCE scenarios can be summarized as follows.

1. Replacing 20% buses with 162 AMoD vehicles can reduce the average disutility for commuters by 15.8%. While substituting buses with AMoD vehicles in the PCE setting does not yield benefits at the system level, a decrease in average disutility is observed within the CCE scenarios. This suggests that a sufficient number of AMoD vehicles can effectively compliment the transit network. Specifically, replacing 20% of buses with 82 AMoD vehicles leads to a 59.3% increase in commuter disutility, while replacing with 162 AMoD vehicles results in a 15.8% reduction in commuter disutility.

Nonetheless, the benefits of such replacements exhibit a non-monotonic pattern; substituting 40% of buses with 328 AMoD vehicles leads to a lesser improvement, reducing commuter disutility by only 6.5%. Further replacements diminish system performance, underscoring that while AMoD vehicles can enhance transit networks, maintaining an essential level of transit services is crucial for their capacity to trans-

port large numbers of people efficiently.

2. Buses can better serve global demand and AMoD vehicles can better serve local demand. With an increase in the number of AMoD vehicles, there's a noticeable decrease in the percentage of unserved local commuters, whereas the percentage of downtown commuters not served goes up. AMoD vehicles are particularly adept at serving local commuters, owing to the less dense demand patterns in local areas. On the other hand, downtown commuter demand is more destination-focused, making bus transportation exceptionally efficient for serving these concentrated demand patterns.

3. Optimal operation strategy is to not discounting AMoD services. Across various scenarios, the optimal discount factor for utilizing AMoD services consistently stands at 1.0, suggesting that no discounts are applied to AMoD services. This decision is intuitive given the limited AMoD vehicles; introducing discounts on AMoD fares would render it the most attractive option for commuters, subsequently causing delays across the entire system. Notably, a noticeable increase in disutility is observed when 20% of buses are substituted with 82 AMoD vehicles. Lowering AMoD fares under these conditions would only exacerbate the issue, leading to further increases in excess waiting times.

4. Activation rate of AMoD and utilization of bus routes are non-monotonic with respect to AMoD-bus configurations. The utilization patterns of transit line and AMoD services, as shown in Table 6.4, demonstrate non-linear or non-monotonic changes when replacing transit vehicles with AMoD vehicles. At first glance, increasing the number of AMoD vehicles might seem like a straightforward solution to replace bus routes serving local commuters. However, as the fleet of AMoD vehicles increases, the reduced waiting times make these services more appealing, thereby attracting more commuters to use. Concurrently, reducing bus routes leaves some local commuters with no option but to use AMoD, leading to potential system congestion as more commuters waiting for the available AMoD vehicles.

To mitigate system-level congestion when utilizing AMoD services, optimal system design might involve limiting the availability of AMoD vehicles during peak times,

despite having a large fleet. This lead to 61% and 66% AMoD activation for scenarios with 328 and 656 AMoD vehicles. For the scenario with 40% buses and 492 AMoD vehicles, AMoD vehicles can operate at 100% activation rate due to having more bus lines in operation. This is feasible because the existing bus services can accommodate most local commuters, reducing the likelihood of a surge in AMoD demand and thus avoiding congestion.

Figure 6-7: Optimal bus network for 100% buses, 0 AMoD vehicles, and 80% downtown commuters.

Figure 6-8: Optimal bus network for 20% buses, 328 AMoD vehicles, and 80% downtown commuters under the PCE setting.

Figure 6-7 to 6-9 illustrate the bus network configuration in a scenario with 80% downtown commuters. Figure 6-7 details the network design when it's composed entirely of buses (100%) without any AMoD vehicles. Figure 6-8 outlines the structure with a significant reduction in buses to 20%, incorporating 328 AMoD vehicles into the system under the PCE setting. In this transformation, a noticeable shift occurs with the majority of bus routes in the north-south direction being eliminated. However,

Figure 6-9: Optimal bus network for 20% buses, 656 AMoD vehicles, and 80% downtown commuters under the CCE setting.

routes that connect to rail stations are preserved, and these retained bus services enjoy higher frequencies compared to other routes.

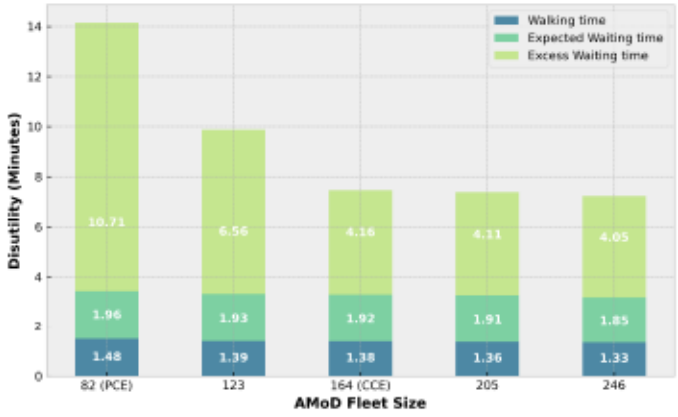
Figure 6-9 shows the bus network configuration in a scenario featuring 20% buses and 656 AMoD vehicles under the CEE setting. While the same bus routes are preserved as in the PCE scenario, the distribution of buses across these routes is more equitable in the CCE scenario. The increased presence of AMoD vehicles facilitates better service to areas with higher demand, thereby reducing the necessity for bus routes to operate at high frequencies.

6.4.3 Sensitivity Analyses

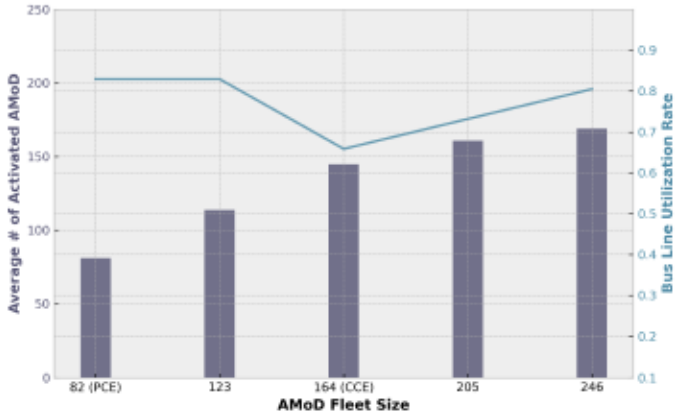
In this section, we conduct sensitivity analyses over two critical model parameters: i) AMoD fleet size, and ii) percentage of downtown commuters.

Figure 6-10 illustrates a sensitivity analysis of varying AMoD fleet sizes, from 82 to 246 vehicles, increasing incrementally by 41, within a scenario that maintains 80% buses and 80% downtown commuters. Figure 6-10a details the average disutility faced by commuters and its breakdown across different AMoD fleet sizes. An increase in the number of AMoD vehicles leads to a decrease in average disutility, attributed to improvements in walking time, expected waiting time, and reduction in excess waiting time. The expansion of the AMoD fleet enhances the provision of door-to-door services, thereby reducing the waiting time for accessing AMoD services.

Figure 6-10b presents the activation rates of AMoD vehicles and utilization of



(a) Average disutility and its breakdown.



(b) Utilization of AMoD vehicles and bus lines.

Figure 6-10: Sensitivity analysis for different AMoD fleet size.

bus lines, highlighting that the usage of AMoD vehicles rises with an increase in fleet size. Surprisingly, the utilization rate of bus lines initially decreases but then shows an increase. This unexpected trend suggests that while a larger fleet of AMoD vehicles might theoretically replace more bus services—especially those serving local commuters—the decreased waiting time for AMoD services actually encourages more commuters to choose AMoD. Consequently, certain bus routes remain necessary to accommodate commuters who would benefit from AMoD services but cannot access them due to the high demand from others also switching to AMoD.

Figure 6-11 shows how the average commuter disutility varies across different ratios of available bus runs (γ) and the percentage of downtown commuters (ψ) under the PCE setting. In scenarios with entirely downtown commuters (100%), an increase

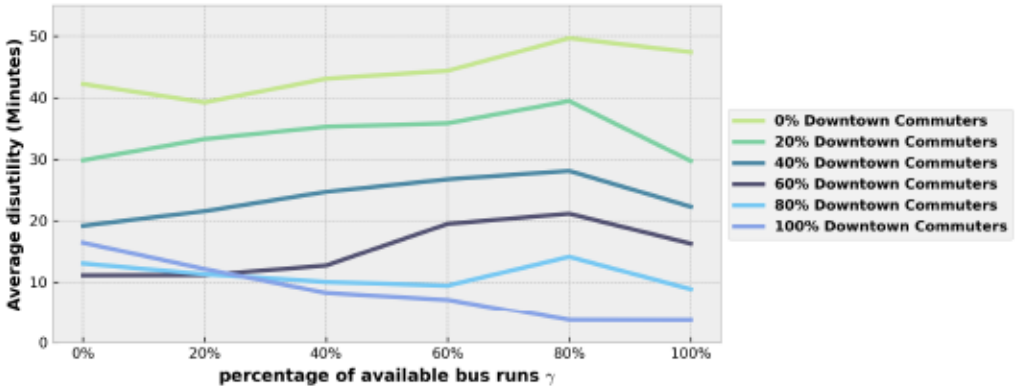


Figure 6-11: Sensitivity analysis for different percentage of downtown commuters.

in the proportion of bus services (γ) correlates with reduced average disutility, as buses are more adept at serving the needs of downtown commuters. Conversely, in scenarios with exclusively local commuters (0% downtown commuters), a reduction in bus availability generally leads to lower disutility levels. The relationship is not straightforward, however, as a balanced mix of buses and AMoD vehicles can better serve local commuters.

Interestingly, in scenarios without bus services (0% available bus runs), the situation with 60% downtown commuters has the lowest level of commuter disutility. In these instances, AMoD vehicles are more efficiently matched to the combined needs of local commuters and downtown commuters traveling to rail stations. Scenarios with 100% downtown commuters experience surge demand periods that AMoD vehicles alone cannot immediately accommodate, resulting in increased wait times for commuters within the system.

In conclusion, AMoD vehicles are particularly effective at accommodating local demand patterns, while traditional transit systems excel in serving downtown commuters. Nevertheless, maintaining a synergy between both systems is crucial for addressing the full spectrum of demand patterns. Identifying the sweet spot between the size of the AMoD fleet and the level of transit services emerges as a key strategy for maximizing efficiency across all commuter types.

6.5 Conclusions and Future Work

In this chapter, we introduce a comprehensive optimization framework designed to jointly optimize the transit networks and frequencies, specify the size and distribution of AMoD fleets, and determine service pricing, all with the objective of minimizing commuters' total disutility. We propose an optimization model for the integrated design of Transit-Centric Multimodal Urban Mobility with Autonomous Mobility-on-Demand (TCMUM-AMoD) systems, which incorporates commuters' mode and route choice behavior via discrete choice models. The proposed optimization model is a mixed integer non-linear program (MINLP) which is intractable to solve at a large scale. Therefore, a first-order approximation algorithm is employed which can solve the problem efficiently. This framework has been tested through a real-world case study in Chicago, encompassing a variety of demand scenarios. The outcomes validate the effectiveness of our model in generating system design solutions. Moreover, the findings reveal the efficiency of AMoD vehicles in meeting local demand patterns and the efficacy of transit vehicles in accommodating downtown and long-distance commuting needs. Meanwhile, the study highlights the importance of striking an optimal balance between AMoD vehicle availability and transit service levels when designing the integrated urban mobility systems.

There are several limitations in this work. Firstly, it does not account for dynamic information within the commuter decision-making process. The discrete choice model applied assumes that commuter preferences for modes and routes are based on static information, an assumption that may not hold true in real-world scenarios. While real-time travel data for traditional transit systems pose a challenge, AMoD systems offer up-to-the-minute information on waiting times and estimated arrivals. Secondly, the model overlooks real-time traffic conditions, which can be significantly impacted by the introduction of AMoD vehicles, particularly in congested areas near subway stations. Lastly, the operational dynamics of AMoD systems, including vehicle rebalancing to redistribute vacant vehicles across different areas, are not considered. This aspect could notably enhance system efficiency. Addressing these limitations presents

valuable avenues for future research.

Chapter 7

Conclusion

7.1 Summary: Results and Contributions

This dissertation outlines a framework for creating a robust and integrated urban mobility system. It implements and evaluates multiple decision-making approaches under uncertainty, using real-world transportation scenarios. In this section, we summarize the key empirical results and methodological contributions of the dissertation.

7.1.1 Empirical Results

From an empirical perspective, this dissertation assesses the performance of multiple models within a real-world context. These empirical findings offer valuable insights into handling demand uncertainty in practical operations and demonstrate how integrated system designs perform under various demand scenarios. In this section, we will summarize the main empirical findings of the entire dissertation.

Vehicle Rebalancing.

In Chapter 2, we evaluate three models designed to solve the vehicle rebalancing problem: 1) matching integrated vehicle rebalancing (MIVR), 2) independent vehicle rebalancing (VR) and 3) fluid-based empty car routing policy (FERP). Our findings demonstrate that MIVR consistently outperforms both VR and FERP in terms of

reducing customer wait times.

When comparing MIVR with VR, performance differences emerge based on demand-to-supply ratio and demand prediction error. Under conditions of a low demand-to-supply ratio, MIVR excels by achieving shorter wait times and requiring fewer vehicle rebalances. However, as the demand-to-supply ratio increases, MIVR necessitates rebalancing more vehicles than VR. This increased rebalancing makes MIVR more vulnerable to demand uncertainty, leading to poorer performance compared to VR when both the demand-to-supply ratio and the demand prediction error are high.

In contrast, when MIVR is compared with FERP, it is observed that MIVR effectively reduces customer wait times by taking more frequent vehicle rebalancing actions. FERP, while quicker in decision-making due to its consistent strategy over fixed time periods (e.g., 2 hours), lacks the responsiveness of MIVR. MIVR updates its strategy based on the system's state at the beginning of every decision interval (every 5 minutes in the experiments), solving an optimization problem to generate new rebalancing decisions. Despite requiring more computational resources, MIVR's process can be executed within seconds for large-scale networks, offering a more dynamic and responsive solution to the vehicle rebalancing challenge.

Handling Demand Uncertainty in Vehicle Rebalancing.

In Chapters 2 and 3, we explore various methods for handling demand uncertainty within the framework of the MIVR model. These methods include robust optimization, predict-then-optimize framework, and data-driven optimization. Our simulation results demonstrate that the robust MIVR approach can significantly reduce customer wait times while requiring fewer vehicle rebalancing. By adopting a conservative position in the face of uncertainty in demand, robust MIVR improves overall system performance.

The predict-then-optimize framework performs well when demand forecasts are accurate. However, in scenarios with large prediction errors, underestimating demand leads to better system outcomes by minimizing customer wait times. Demand underestimation introduces a level of conservativeness in vehicle rebalancing decisions,

which is a key factor in the success of robust MIVR.

Data-driven approaches outperform the predict-then-optimize framework when faced with high-demand prediction errors. Moreover, these approaches maintain a computational advantage over robust optimization methods while delivering comparable results. It is also important to note that data-driven MIVR models benefit from the sparsity property, meaning that considering only a small proportion of historical demand data can yield excellent performance outcomes.

Quantification and Reduction of Disparity in Vehicle Rebalancing.

In Chapter 4, we delve into quantifying and reducing disparity within the vehicle rebalancing problem, which consists of two main components: upstream demand prediction and downstream vehicle rebalancing.

Upstream Demand Prediction: The disparity here refers to the variance in prediction errors across different regions. We quantify this disparity using two metrics: the Mean-Variance of the Percentage Error (MVPE) and the Generalized Entropy Index (GEI). Observing significant error disparities within the existing demand prediction framework, this dissertation introduces a Social Aware Spatio-Temporal Graph Convolutional Network (SA-STGCN) framework. The SA-STGCN framework is able to reduce MVPE and GEI by 11.6% and 12.1%, respectively, without compromising the accuracy of the predictions.

Downstream Vehicle Rebalancing: In this component, disparity is understood as the service disparity experienced by customers from different regions, quantified by the standard deviation of customer wait times across regions. To address this issue, we propose an equity-enhanced version of the MIVR model, which incorporates equity weights derived from the SA-STGCN framework. This model effectively reduces the variability in customer wait times by 8.43%, while maintaining a similar average wait time, thereby enhancing service equity across different geographic areas.

Handling Demand Uncertainty in Transit Scheduling.

In Chapter 5, we address the issue of demand uncertainty within transit systems by proposing a robust transit frequency setting model. This model is designed to generate transit schedules that are resilient against fluctuations in demand. Additionally, a stochastic transit frequency setting model using stochastic optimization methods is introduced as a benchmark to compare against the robust model. Multiple demand scenarios were created to evaluate the performance of various transit schedules along a specific transit corridor in Chicago.

Under typical demand scenarios, both the robust and stochastic transit schedules demonstrated superior performance compared to the current operational schedules and those generated using static optimization models. When comparing the robust and stochastic schedules directly, the robust schedules were found to reduce in-vehicle travel time but increase wait time for passengers.

In scenarios of surge demand, the robust schedules clearly outperformed the stochastic schedules by reducing both in-vehicle travel and wait times. This improvement indicates that robust scheduling is particularly effective under high-demand conditions, ensuring a more reliable and efficient transit service for passengers. This chapter highlights the importance of incorporating robustness into transit planning to better accommodate variable and unpredictable demand patterns, enhancing overall system performance.

Integrated System Design.

In Chapter 6, we introduce a framework designed to develop a transit-centric multimodal system that incorporates Autonomous Mobility-on-Demand (AMoD) services. This framework was applied to the southern side of Chicago under various demand scenarios to assess the impacts of integrating AMoD with traditional bus services.

The results of our analysis reveal nuanced outcomes when buses are partially replaced with AMoD vehicles. Specifically, replacing 20% of buses with 82 AMoD vehicles resulted in a significant increase in average disutility for commuters—up by

59.3%. In contrast, replacing the same percentage of buses with 164 AMoD vehicles led to a 15.8% reduction in average disutility, demonstrating improved commuter experiences.

These numerical findings indicate that integrating AMoD services into existing transit systems is not merely a straightforward augmentation. Having only a small fleet of AMoD vehicles can lead to excessive delays across the system. It is therefore essential to carefully co-design the transit and AMoD systems, ensuring an adequate number of AMoD vehicles are deployed to effectively replace existing low-frequency bus services.

Moreover, the study suggests that there is an optimal configuration of buses and AMoD vehicles for any given demand pattern, which can maximize efficiency and minimize commuter disutility.

7.1.2 Methodological Contributions

From a methodological standpoint, this dissertation introduces several innovative approaches designed to tackle diverse challenges within urban mobility systems. These methods are not only applicable to the specific scenarios discussed in this work but also hold potential for broader application in other research areas.

Matching-Integrated Vehicle Rebalancing Model.

In Chapter 2, we propose the MIVR model, and it is used as the foundation model for Chapters 3 and 4. The MIVR model, for the first time, introduce the matching component into the vehicle rebalancing problem to generate a forward-looking rebalancing decisions. As both problems are generating vehicle movement decisions, it provides the opportunity to fully integrate two problems under the traditional optimization framework. Meanwhile, the MIVR model is also more robust to demand uncertainty in the vehicle rebalancing problem.

Problem-Specific Uncertainty Set for Robust Optimization.

In Chapter 2, we further utilize the robust optimization technique to handle the demand uncertainty in the vehicle rebalancing problem. Unlike the standard robust optimization studies, we propose the problem-specific uncertainty set for the robust MIVR model. The designed uncertainty set describes what level of fluctuations of ride-hailing demand should be considered when generating rebalancing decisions. Other transportation-related problems should also consider formulating uncertainty sets with real-world meanings when applying robust optimization techniques.

Socio-Aware Spatial-Temporal Graph Convolutional Network.

In Chapter 4, we propose the Socio-Aware Spatial-Temporal Graph Convolutional Network (SA-STGCN) framework for the ride-hailing demand prediction task with the objective of reducing error disparity. The sociodemographic information is embedded in the framework by integrating with the adjacency matrix, which incorporates the spatial locality information. Meanwhile, disparity-reduced loss regularization terms are added in the training loss function to further reduce the error disparity in the demand prediction. The proposed SA-STGCN framework can be used in other transportation-related problems to get a "fairer" prediction.

Transit Downsizing Approach.

In Chapter 5, the Transit Downsizing (TD) approach is introduced to reduce the problem size in transit frequency setting when employing robust optimization techniques. This approach reduces the size of the demand matrix utilized in the optimization model. The TD method is intuitive because transit planning should focus on the demand profiles of frequent users while disregarding less frequent demand patterns. This method is universally applicable to any transit-related problems that incorporate a demand matrix. Given that transit demand matrices are typically sparse and large-scale, the TD approach proves to be an effective strategy for reducing problem complexity.

7.1.3 Results and Contributions of Each Chapter

Robust Matching-Integrated Vehicle Rebalancing in Ride-Hailing System with Uncertain Demand

In this chapter, we formulate the MIVR model, which incorporates the driver-customer matching component into the consideration of vehicle rebalancing decisions made by ride-hailing operators, to protect rebalancing decisions against future demand uncertainty induced by inaccurate demand estimates. We evaluate the performance of our model by comparing against a benchmark VR model and a state-of-the-art model, named fluid-based empty-car routing policy (FERP), using real-world ride-hailing trip data from the New York City. Comparing to the VR model, the MIVR model reduces the average customer wait time by 4.4% and the total non-occupied VMT by 8.5%. The MIVR and VR perform differently under different scenarios. When having more vacant vehicles compared to future demand, the MIVR model conducts less vehicle rebalancing operations compared to the VR model. When a large fleet is available, a *Pareto* improvement can be found regarding the overall non-occupied VMT, the average vehicle rebalancing trips, the average customer wait time and the number of unsatisfied requests. Comparing to the FERP, the MIVR model reduces the average customer wait time by 18% while generating a more proactive rebalancing strategy with 24% more non-occupied VMT.

To further immunize solutions against demand uncertainty, we propose the robust MIVR model by introducing RO techniques. The robust MIVR is especially effective when the supply of ride-hailing vehicles is sufficient and most requests can be satisfied, reducing average customer wait time by up to 41%. Under both sufficient-supply and insufficient-supply cases, the robust MIVR model prevents rebalancing decisions from inaccurate demand estimation by rebalancing fewer vehicles. Additionally, introducing robustness into the MIVR model generates rebalancing decisions that performs better than decisions produced by the nominal MIVR model under most demand scenarios.

Data-driven Vehicle Rebalancing with Predictive Prescriptions in the Ride-hailing System

In this chapter, we introduce a novel data-driven optimization approach, predictive prescriptions, into the vehicle rebalancing problem to handle demand uncertainty in the ride-hailing system. Building upon a state-of-the-art vehicle rebalancing model, MIVR proposed in the Chapter 2, point-prediction-driven optimization models and data-driven optimization models are proposed to improve the model performance against demand uncertainty. Compared to robust MIVR models, data-driven MIVR models achieve competitive operational performances while being more computational efficient.

Regarding point-prediction-driven optimization models, a graph convolutional LSTM and two zone-based LSTM models are constructed in this paper to predict future demand for each sub-region. As for data-driven optimization models, SAA and predictive prescription with KNN and ORT are introduced in this paper. A real-world simulation with NYC data is used to evaluate model performances under four different demand scenarios. Between the data-driven optimization and point-prediction-driven optimization models, one should make a decision based on supply to demand ratio and the prediction accuracy. When the future demand can be predicted accurately, point-prediction-driven optimization models should be adopted. When the demand is volatile and hard to predict, data-driven optimization models perform better. The system performances can be further improved for data-driven optimization models when the supply to demand ratio is higher, indicating more idle vehicles are available to be redistributed. Among all data-driven optimization methods, predictive prescriptions perform better by leveraging auxiliary information.

Meanwhile, prediction errors over the future demand in the vehicle rebalancing problem can be beneficial to system performances when errors come from demand underestimation. The "conservativeness" brought by the demand underestimation improves the system performance due to highly uncertain demand in the future. The best-performing data-driven optimization model, predictive prescription with KNN-5,

is also compared with the robust MIVR proposed in Chapter 2, which utilizes robust optimization techniques to protect rebalancing decisions against demand uncertainty. The robust MIVR model reduces the customer unsatisfaction rate while conducting fewer vehicle rebalancing trips. On the other hand, predictive prescriptions reduce the average customer wait time but serve fewer customers.

From a practical perspective, rebalancing models need to be selected ahead of schedule. When considering a whole day's demand, demand uncertainty and prediction accuracy of predictive models change from time to time. Therefore, a good operation strategy is to separate the whole operation period into high and low uncertainty periods based on historical demand data. For low uncertainty periods, point-prediction-driven optimization models should be adopted. As for high uncertainty periods, data-driven optimization models, including robust and predictive prescription models, can be applied.

Disparity-Reducing Vehicle Rebalancing in the Ride-hailing System

Service disparity issues are naturally embedded in the vehicle rebalancing problem. This chapter presents a pioneering framework aimed at reducing disparity in both predicting ride-hailing demand and delivering equitable service to riders. The framework introduces a Socio-Aware Spatio-Temporal Graph Convolutional Network (SA-STGCN), which integrates a socio-enriched adjacency matrix and bias-reduction regularization methods. Additionally, it features a vehicle rebalancing engine that incorporates equity considerations into its objective function. This framework was evaluated using a simulator with real-world ride-hailing data, demonstrating that the SA-STGCN model not only outperforms standard demand prediction models in increasing accuracy but also in reducing error disparity. Significantly, mitigation in disparity at the demand prediction stage lead to more equitable service delivery in the vehicle rebalancing process. The vehicle rebalancing module, enhanced with equity weights, showed a notable reduction in the standard deviation of customer wait times by 6.5%, while not diminishing the system efficiency for ride-hailing platforms.

The proposed framework offers a viable approach for ride-hailing companies to

reduce service disparity into their operations, and it provides a basis for government regulations aimed at preventing service imbalances across different areas. However, realizing the win-win scenario highlighted in the study involves addressing practical challenges. A key solution lies in developing driver incentive mechanisms. These mechanisms should ensure that drivers are motivated to serve in underserved communities and that their earnings remain stable despite such commitments. As the role of ride-hailing services becomes more central in our everyday activities, it's crucial to make certain that these platforms maintain a strong commitment to social responsibility and proactively enhance the well-being and inclusiveness of the communities they operate in.

Robust Transit Frequency Setting Problem with Demand Uncertainty

In this chapter, two major issues are addressed when generating transit schedules: i) inherent demand uncertainties, and ii) gigantic OD matrices. To protect transit schedules against demand variations, a robust TFSP model is proposed. To the best of the authors' knowledge, this chapter is the first to apply RO technique for solving TFSPs. A nominal optimization model is formulated to solve the TFSPs under a single transit line setting, and an extended model considering crowding levels on transit vehicles is proposed.

To solve optimization problems efficiently given real-world transit instances, the Transit Downsizing (TD) approach is proposed based on the observation where transit demand matrices are sparse. We theoretically prove that the optimal objective function of the problem after TD is close to that of the original problem (i.e., the difference is bounded from above). A benchmark stochastic TFSP model is also formulated to demonstrate the robust performance of the TFSP model. Real-world transit lines operated by CTA are used to test the performance of the transit schedules generated with the proposed models compared to the current transit schedule. Both stochastic and robust transit schedules reduce wait times and in-vehicle travel times simultaneously for passengers over multiple demand scenarios. Compared to benchmark stochastic schedules, robust schedules further reduce passengers' wait times by up to

2.94% and in-vehicle travel times by up to 1.27% under the surge demand scenario.

Design of Transit-Centric Multimodal Urban Mobility System with Autonomous Mobility-on-Demand

In this chapter, we introduce a comprehensive optimization framework designed to jointly optimize the transit networks and frequencies, specify the size and distribution of AMoD fleets, and determine service pricing, all with the objective of minimizing commuters' total disutility, which consists of waiting and walking time. We propose a Mixed Integer Non-Linear Program (MINLP) for the integrated design of TCMUM-AMoD systems, employing a first-order approximation algorithm that efficiently solves the problem on a large scale. This framework has been tested through a real-world case study in Chicago, encompassing a variety of demand scenarios.

The outcomes validate the effectiveness of our model in generating system design solutions. Moreover, the findings underscore the efficiency of AMoD vehicles in meeting local demand patterns and the efficacy of transit vehicles in accommodating downtown and long-distance commuting needs. Crucially, the study highlights the importance of striking an optimal balance between AMoD vehicle availability and transit service levels when designing the integrated urban mobility systems. There exists an optimal transit-AMoD configuration under any demand patterns. Also, the results show that passengers' mode and route choice behaviors play an important role when designing the multimodal mobility system.

7.2 Implications

In this section, we discuss the implications of this dissertation, particularly how the promising numerical results of the proposed methodologies offer practical insights for the industry. Industry professionals have a substantial opportunity to learn from the findings of this dissertation and consider adopting some of the proposed methodologies to optimize their operations.

7.2.1 Ride-sharing Industry

Chapters 2, 3, and 4 of this dissertation introduce methodologies designed to address the vehicle rebalancing problem, which is pivotal for the efficiency of ride-sharing platforms like Uber and Lyft. Currently, these platforms incentivize drivers to reposition themselves through dynamic pricing mechanisms. Although vehicle rebalancing algorithms are not directly implemented in the current system, decisions generated from the proposed MIVR models could provide a foundational baseline to establish such dynamic pricing more effectively.

With the rapid development of autonomous driving technologies, the emergence of "robo-taxis" is anticipated globally. This shift will give platforms complete control over their fleets, making vehicle rebalancing the most critical operational challenge. Therefore, it is important for the ride-sharing industry to develop and implement models that can efficiently solve vehicle rebalancing problems.

Moreover, given the inherent uncertainties in ride-sharing demand, it is crucial to implement vehicle rebalancing models that are robust against such uncertainties. This dissertation proposes three types of models to manage demand uncertainty in vehicle rebalancing: 1) demand prediction + MIVR, 2) data-driven MIVR, and 3) robust MIVR. Table 7.1 details the implementation aspects of these three models.

Model	MIVR	Data-driven MIVR	Robust MIVR
Data requirement	No historical data required	Historical information	Estimation of uncertainty level from historical data
Demand prediction requirement	Yes	No	No
Computational efficiency	High	Medium	Low

Table 7.1: Implementation requirements for MIVR models.

For ride-sharing platforms, selecting the appropriate vehicle rebalancing model is crucial and depends on the predictability of demand. The baseline MIVR model is ideal when demand can be accurately predicted, suitable for scenarios with stable or known demand patterns, allowing efficient vehicle distribution and minimized waiting

times. In contrast, the data-driven MIVR model is recommended when demand is challenging to predict, striking a balance between computational efficiency and robust performance.

In terms of the robust MIVR model, it suffers from its higher computational complexity, although it can serve more customers and reduce waiting times. Moreover, the robust MIVR model requires the level of uncertainty to be predetermined manually, which can be cumbersome if the uncertainty level changes frequently. In general, each model offers distinct advantages and challenges, necessitating a strategic approach to model selection based on the specific conditions and demands of the operational environment.

7.2.2 Public Transit Industry

Chapters 5 and 6 of the dissertation delve into robust operations and integration with AMoD services for transit systems, particularly relevant in the context of permanently reduced transit ridership in the post-pandemic world. This shift provides an opportunity for transit agencies to reassess and improve their system designs and operations to enhance service delivery.

Chapter 5 introduces a robust transit frequency setting model that can help transit agencies create more reliable schedules by accounting for demand fluctuations. Such robust schedules are better equipped to adapt to uncertainties and changes in demand. Implementing this model involves establishing a process for estimating the level of demand uncertainty using historical data. Since transit agencies typically update their schedules quarterly based on historical data, integrating the robust model into the current decision-making process could be done seamlessly.

Chapter 6 explores the integration of transit and AMoD services, offering insights from the perspective of transit agencies. It discusses two approaches for replacing parts of the bus fleet with AMoD vehicles, as demonstrated in numerical experiments. From a capital cost perspective, this replacement could reduce overall passenger disutility and enhance user experience. For transit agencies looking to implement such an integrated system, key steps would include purchasing a sufficient number of AMoD

vehicles for local deployment, developing a user-friendly app for AMoD service requests, and implementing backend operations for real-time management of AMoD system and design of integrated system. While implementing such a comprehensive system poses challenges, transit agencies might also consider outsourcing AMoD services to specialized companies like Waymo or Cruise. In such cases, the primary responsibility of the transit agency would be the design of an integrated system, which can be effectively addressed using the model proposed in this dissertation. Also, fares structures of two systems need to be integrated. This approach simplifies the operational demands on the transit agency while leveraging the expertise of established AMoD providers.

7.3 Future Research Directions

7.3.1 Overcome Limitations in Existing Studies

The studies discussed in this dissertation have several limitations that offer opportunities for further studies in future research. In Chapter 2, the uncertainty set $\bar{\mathcal{U}}^k(\Gamma)$ has a limited impact on system performance. More effective and interpretable uncertainty sets could be designed to model uncertainty in the ride-hailing system. In addition, the MIVR model could be extended to solve the problem of vehicle rebalancing in the shared MoD system.

In Chapter 3, we found that the vehicle rebalancing problem can be beneficial to system performances when errors come from demand underestimation. However, predictive models usually aim for "unbiasedness", and weight overestimation and underestimation equally. A possible future research direction is to develop predictive models for ride-hailing systems which have an asymmetric loss function that favors underestimation over overestimation. Meanwhile, the extra benefits brought by conservativeness due to demand underestimation should have a limit. Future research could identify such an underestimation level where vehicle rebalancing benefits the most. Another future research direction can be the introduction of robust data-driven

optimization techniques into the MIVR model, which combines the benefits of both data-driven optimization and robust optimization.

In Chapter 4, we do not make a formal judgement on what is a *fair* vehicle rebalancing operation. Instead, we try to understand and reduce the disparity within the system, which serves as the foundation for understanding the fairness in the system. More analyses can be performed to better understand and achieve fairness in vehicle rebalancing. Meanwhile, it would be beneficial to include a focus on driver behavior and earnings, which this study has not addressed. A comprehensive framework could be developed to reduce disparity in all aspects of ride-hailing vehicle rebalancing operations: error disparity in demand prediction, pricing disparity for riders, and earnings disparity for drivers. Such a framework should ensure that drivers who are redirected to serve underserved communities are compensated equitably, comparable to those serving in city centers. This approach would create a more balanced and fair environment for all parties involved in the ride-hailing ecosystem.

In Chapter 5, the main limitation of this study is using heuristics to solve the robust TFSP model without proof of optimality. Meanwhile, the parameter controlling the size of the uncertainty set needs to be selected manually. Future studies could develop methodologies for decreasing problem sizes while maintaining a certain level of optimality loss. Data-driven approaches can be introduced to automatically select the value of uncertain parameter Γ. Also, our demand data only provides the time information when passengers actually board transit vehicles or enter subway stations, knowing more time information (e.g., the deadline for passengers to arrive at their destinations) could further introduce passengers' time preferences into the model. Another interesting research direction is pattern generation. Our model has the ability to select an optimal set of patterns to operate on a single transit line. However, how to generate a set of potential patterns for a single transit line can be a challenging task. Performances of different pattern generation algorithms can be evaluated through our proposed TFSP model. Meanwhile, other sources of uncertainty in transit systems can be considered when generating robust transit schedules, e.g., supply uncertainty (last-minute driver absence) and travel time uncertainty. Lastly,

the proposed TFSP model can be extended to solve a network-level frequency setting problem with multiple transit lines.

In Chapter 6, there are several limitations and can be potentially addressed in the future study. Firstly, it does not account for dynamic information within the commuter decision-making process. The discrete choice model applied assumes that commuter preferences for modes and routes are based on static information, an assumption that may not hold true in real-world scenarios. While real-time travel data for traditional transit systems pose a challenge, AMoD systems offer up-to-the-minute information on waiting times and estimated arrivals. Secondly, the model overlooks real-time traffic conditions, which can be significantly impacted by the introduction of AMoD vehicles, particularly in congested areas near subway stations. Lastly, the operational dynamics of AMoD systems, including vehicle rebalancing to redistribute vacant vehicles across different areas, are not considered. This aspect could notably enhance system efficiency.

7.3.2 Generalization of Dissertation

In this dissertation, the numerical experiments evaluating each proposed model are specifically conducted for New York City or Chicago. This context-specific numerical results raise the question of whether the insights and findings can be generalized to other cities or different contexts. Generalizability is a critical aspect of research, and exploring this could provide a valuable direction for future studies. In this section, we will discuss the potential for generalizing the MIVR model as a specific example.

Additionally, the approaches proposed in this dissertation, such as transit downsizing, possess potential for broader application beyond the specific cases examined. These methodologies could be adapted and applied to a variety of problems in urban mobility design and operations. Future research could explore these possibilities, examining how the methods developed in this dissertation could be modified to suit different urban contexts or applied to solve different types of problems within the realm of urban mobility.

Generalization of Matching-Integrating Vehicle Rebalancing

In Chapter 2, we examined the performance of the MIVR model in relation to the supply-to-demand ratio, in comparison to the standalone VR model that does not include a matching component. The findings indicate that the MIVR model redistributes fewer vehicles when the supply-to-demand ratio is high, and more vehicles when it is low. The advantages of the MIVR model over the VR model increase as the supply-to-demand ratio grows. These results are specific to the Manhattan network, which has a grid-based structure. However, the outcomes might differ or not hold true in different network configurations.

My hypothesis regarding the generalization of MIVR performances is that the MIVR model will generate an increased number of rebalancing trips as the demand-to-supply ratio rises. Figure 7-1 illustrates my expectations for how the number of rebalancing trips will change with increasing demand-to-supply ratios in two distinct road networks: Boston and Manhattan. The Manhattan network represents a standard grid-based structure. In contrast, the Boston road network exhibits a clique-based structure, characterized by well-connected local regions that are linked together by several express roads.

Figure 7-1: Relationship between demand-to-supply ratio and number of rebalancing trips.

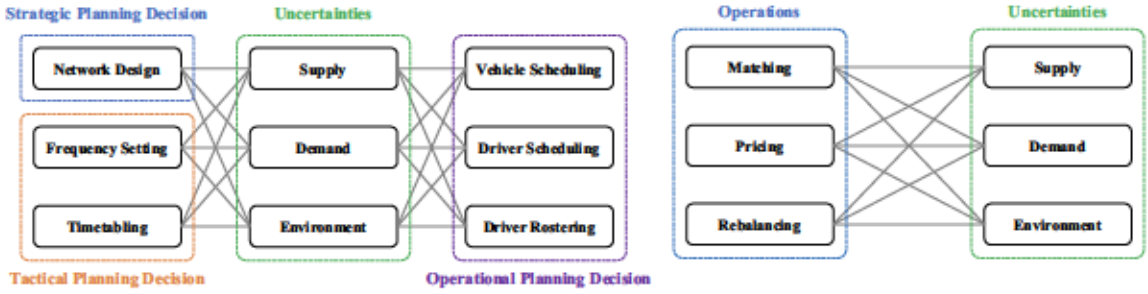
In grid-based networks, the number of rebalancing trips increases smoothly as the demand-to-supply ratio rises. Conversely, in clique-based networks, we expect to observe a "jump" in the number of rebalancing trips when the demand-to-supply

ratio increases. This is because when the ratio is low, each local clique has sufficient available vehicles, making inter-clique rebalancing unnecessary. However, once the demand surpasses a certain threshold, it becomes necessary to rebalance vehicles across different cliques.

To further validate this hypothesis, experiments could be conducted in various cities to observe how these patterns manifest in different urban layouts. Additionally, these relationships should be theoretically modeled within a vehicle rebalancing framework to predict and understand the dynamics under different urban structures and demand conditions.

7.3.3 Incorporate Different Sources of Uncertainty

As introduced in Chapter 1, there are three principal sources of uncertainty impacting urban mobility systems: i) supply uncertainty, ii) demand uncertainty, and iii) environmental uncertainty. Figure 7-2 illustrates how each operational problem can be aligned with a source of uncertainty to forge new research avenues. In this section, I will explore a specific future research direction that involves integrating travel time uncertainty into the matching decisions of ride-sharing systems.



(a) Transit system (b) Ride-sharing system

Figure 7-2: Uncertainty in public transit and ride-sharing systems.

Robust Matching with Travel Time Uncertainty

Existing methodology frameworks for driver-customer matching make assumption of deterministic travel times, where in reality travel times are highly uncertain. When

customers request rides from ride-sharing platforms, they typically receive time information, such as estimated time of arrival (ETA) and expected wait times. The time estimation could easily go wrong and significantly impact customer satisfaction with the platform. Customers could experience excessive wait times and leave the system or arrive at their destinations beyond the promised ETA.

Figure 7-3 provides an example showing why it is important to incorporate travel time uncertainty in the ride-sharing operation. Two available vehicles $\{v_1, v_2\}$, and three customer requests $\{r_1, r_2, r_3\}$ are included in this example. Red arrows represent the optimal pick-up routes for vehicles to serve all customer requests under each scenario. Colored road segments have volatile traffic conditions and different matching and routing decisions are made under different traffic conditions. Under normal traffic conditions shown in Figure 7-3a, vehicle v_1 utilizes colored road segments and picks up both customer requests r_1 and r_3. Under congested traffic conditions shown in Figure 7-3b, both vehicles avoid colored road segments, and vehicle v_2 picks up both customer requests r_1 and r_3. With existing matching and routing algorithms, vehicle routes in Figure 7-3a are adopted and customers could experience excessive waiting when traffic becomes congested. With robust matching and routing algorithms considering travel time uncertainty, vehicle routes in Figure 7-3b are adopted since it is less likely that customers will experience excessive wait times under given routes.

(a) Matching with normal traffic condition (b) Matching with congested traffic condition

Figure 7-3: Ride-sharing example under normal and congested traffic conditions.

To integrate travel time uncertainty into the driver-customer matching component of ride-sharing systems, robust and stochastic optimization techniques can be utilized. However, the primary challenge with these techniques is maintaining com-

putational efficiency. The driver-customer matching problem in dynamic ride-sharing requires decisions to be made within seconds, and incorporating sophisticated models to account for travel time uncertainty typically leads to increased computational complexity. Therefore, any proposed approaches must be specifically designed to balance the ability to handle uncertainty with the need for real-time decision-making.

7.3.4 Quantitative Analysis between Integration and Robustness

Chapter 6 is motivated by the hypothesis that system integration serves as a response to the uncertainty in the urban mobility system. As depicted in Figure 7-4, there are three general strategies to address system uncertainty: 1) improving prediction accuracy, 2) enhancing system responsiveness, and 3) mitigating the consequences of uncertainty. The robust models developed in Chapters 2 and 5 exemplify the approach of mitigating consequences of demand uncertainty in ride-sharing and transit systems, respectively.

The integration of AMoD with traditional transit systems enhances system responsiveness by introducing more adaptable AMoD services. This integration allows for greater flexibility in managing fluctuating demand and can potentially reduce the negative impacts of uncertainty in certain scenarios. However, the effectiveness of this integration may vary depending on the specific uncertainties faced by the system.

The integrated system tends to be more robust against changes in demand patterns, primarily because it offers a broader array of travel options to customers. Ride-sharing services, which provide convenient door-to-door transportation across the city, complement traditional transit systems that might struggle with demand fluctuations due to their fixed schedules and limited adaptability in real-time.

Conversely, system-level uncertainties such as sudden surges in demand can expose vulnerabilities in the integrated system. During peak demand periods, the flexibility of ride-sharing, while generally advantageous, becomes a liability due to the low capacity of individual vehicles, which are ill-equipped to handle large volumes of

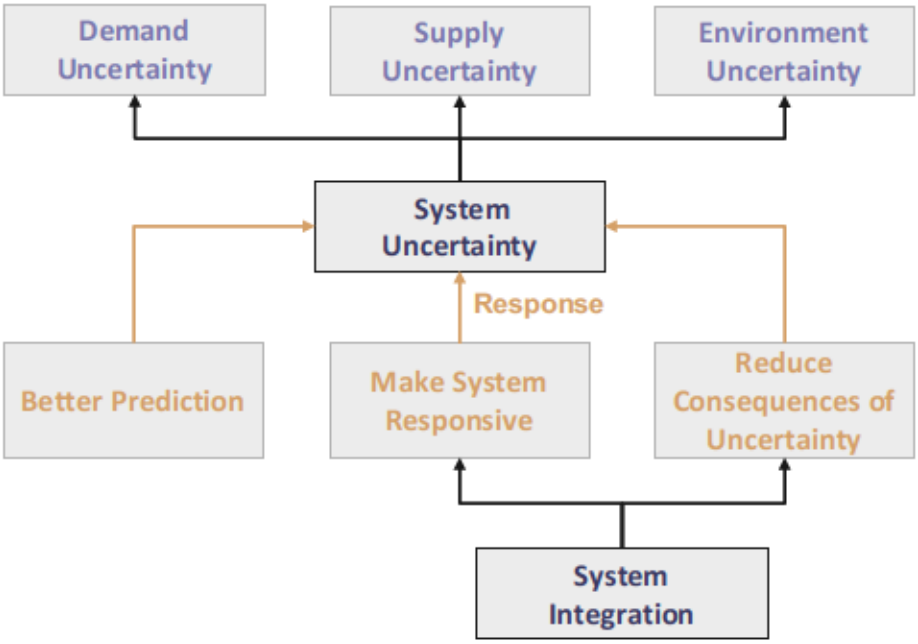


Figure 7-4: Integration as a response to uncertainty.

passengers. In contrast, transit systems, despite their inflexibility, can transport large groups of people efficiently over longer distances. In such scenarios, integration could potentially exacerbate system weaknesses rather than mitigate them.

The effectiveness of using integration as a response to uncertainty within the urban mobility context requires a more meticulous examination. It is crucial to identify under what conditions system integration enhances robustness and against which types of uncertainty. Furthermore, the degree of robustness that system integration contributes should be quantifiable.

7.3.5 Operations of Integrated Urban Mobility System

Lastly, the operations of an integrated urban mobility system represent a significant area for future research. In Chapter 6, we proposed a methodological framework to design such a system. One key insight from our findings is the importance of co-designing the systems for optimal performance. Merely layering one service onto another can lead to sub-optimal outcomes for the system as a whole. This insight holds true for the operations phase as well.

When operating an integrated public transit and ride-sharing system, implementing timed transfers can create seamless connections between the two services. For instance, transit schedules could provide precise time windows that ride-sharing services need to meet for first-mile trips to transit stations. Similarly, knowing the transit schedules allows for proactive rebalancing of ride-sharing vehicles to transit stations to meet demand for last-mile trips from the stations to passengers' final destinations. Such strategic operational integration enhances the efficiency and effectiveness of the integrated urban mobility system.

Appendix A

Chapter 2 Appendix

A.1 Derivation of The Robust Counterpart

Given the following generic constraint

$$L(\cdot) + v^T \zeta \leq c \quad \forall \zeta \in \mathcal{U}, \tag{A.1}$$

where $L(\cdot)$ indicates a function of decision variables in problem (P'), v is a vector in dimension $n\kappa$ and c is a scalar, it is equivalent to

$$L(\cdot) + \max_{\zeta \in \mathcal{U}} v^T \zeta \leq c. \tag{A.2}$$

By taking the convex conjugate of constraint (A.2) we derive the following equivalent constraint

$$L(\cdot) + \delta^*(v \mid \mathcal{U}) \leq c, \tag{A.3}$$

where $\delta(v \mid \mathcal{U})$ is an indicator function such that $\delta(v \mid \mathcal{U}) = 0$ if $v \in \mathcal{U}$, otherwise $\delta(v \mid \mathcal{U}) = \infty$. $\delta^*(v \mid \mathcal{U})$ is the convex conjugate of $\delta(v \mid \mathcal{U})$. Then we introduce Lemma 2 to help with deriving the robust counterpart [16].

Lemma 2. For a constraint $\bar{a}^T x + \delta^*(P^T x \mid Z) \leq b$, let $Z_1, ..., Z_k$ be closed convex

sets, such that $\bigcap_i ri(Z_i) \neq \emptyset$[1], and let $Z = \cap_{i=1}^{k} Z_i$. Then,

$$\delta^*(y \mid Z) = \min_{y^1,\ldots,y^k} \{\sum_{i=1}^{k} \delta^*(y^i \mid Z_i) \mid \sum_{i=1}^{k} y^i = y\},$$

and the constraint becomes

$$\begin{cases} \bar{a}^T x + \sum_{i=1}^{k} \delta^*(y^i \mid Z_i) \leq b \\ \sum_{i=1}^{k} y^i = P^T x \end{cases}$$

Let $\mathcal{U}_0 = \{\zeta : \|\zeta\|_\infty \leq \rho\}$ and $\mathcal{U}_k = \{\zeta : \left|e^T(\zeta \circ \Sigma^k)\right| \leq \Gamma\}, \forall k \in K$, where $\Sigma^k \in \mathbb{R}^{n\kappa}$ denotes a vector with (ik)-th entry equals to σ_i^k, $\forall i \in N$, and other entries equal to zero. The uncertainty set $\mathcal{U}$ can be written as: $\mathcal{U} = \cap_{k=0}^{\kappa} \mathcal{U}_k$. By applying Lemma 2 to constraint (A.3), we develop the following robust counterpart for constraint (A.1):

$$\begin{cases} L(\cdot) + \sum_{k=0}^{\kappa} \delta^*(\theta_k \mid \mathcal{U}_k) \leq c \\ \sum_{k=0}^{\kappa} \theta_k = v \end{cases} \tag{A.4}$$

Which is equivalent to

$$\begin{cases} L(\cdot) + \rho \|\theta_0\|_1 + \Gamma \sum_{k=1}^{\kappa} (\eta_1^k + \eta_2^k) \leq c \\ (\eta_1^{k'} - \eta_2^{k'}) \sigma_i^{k'} = \theta_{k'}^{i,k} \quad \forall i \in N,\ \forall k = k' \in K \\ \theta_{k'}^{i,k} = 0 \quad \forall i \in N,\ \forall k \neq k' \in K \\ \eta_1^k, \eta_2^k \geq 0 \quad \forall k \in K \\ \sum_{k=0}^{\kappa} \theta_k = v \end{cases} \tag{A.5}$$

Where $\theta_k \in \mathbb{R}^{n\kappa}$ and $\theta_{k'}^{i,k}$ represents (ik)-th entry of vector $\theta_{k'}$, $\forall k' \in K$.

[1] $ri(Z_i)$ indicates the relative interior of the set Z_i.

A.2 Benchmark Vehicle Rebalancing (VR) Model

In this section, we formulate a benchmark vehicle rebalancing (VR) model to test the performance of our MIVR model. With similar notations to the MIVR model, we introduce several additional parameters. Let P_v^k, Q_v^k be regional transition matrices regarding vacant vehicles in time period k, which are learned from the historical data. $P_{v,ij}^k$ stands for the probability for a vacant vehicle in sub-region i at time k to be in sub-region j at time $k+1$ and becomes occupied. Similarly, $Q_{v,ij}^k$ denotes the probability for a vacant vehicle in sub-region i at time k to be in sub-region j at time $k+1$ and remains vacant. Two regional transition matrices satisfy the following condition:

$$\sum_{j=1}^{n}(P_{v,ij}^k + Q_{v,ij}^k) = 1, \quad \forall i \in N,\ \forall k \in K.$$

Then the benchmark VR model is:

$$(VR) \quad \min_{x_{ij}^k} \quad \sum_{k=1}^{\kappa}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^k d_{ij}^k + \alpha \cdot \sum_{k=1}^{\kappa}\sum_{i=1}^{n} \mid S_i^k - r_i^k \mid \tag{A.6a}$$

$$\text{s.t.} \quad S_i^k = \sum_{j=1}^{n} x_{ji}^k - \sum_{j=1}^{n} x_{ij}^k + V_i^k \quad \forall i \in N,\ \forall k \in K \tag{A.6b}$$

$$V_i^{k+1} = \sum_{j=1}^{n} Q_{v,ji}^k S_j^k + \sum_{j=1}^{n} Q_{ji}^k O_j^k \quad \forall i \in N,\ \forall k \in K \setminus \{\kappa\} \tag{A.6c}$$

$$O_i^{k+1} = \sum_{j=1}^{n} P_{v,ji}^k S_j^k + \sum_{j=1}^{n} P_{ji}^k O_j^k \quad \forall i \in N,\ \forall k \in K \setminus \{\kappa\} \tag{A.6d}$$

$$\sum_{j=1}^{n} x_{ij}^k \le V_i^k \quad \forall i \in N,\ \forall k \in K \tag{A.6e}$$

$$a_{ij}^k \cdot x_{ij}^k = 0 \quad \forall i \in N,\ \forall k \in K \tag{A.6f}$$

$$x_{ij}^k \in \mathbb{R}^+ \quad \forall i, j \in N,\ \forall k \in K \tag{A.6g}$$

$$S_i^k, V_i^k, O_i^k \in \mathbb{R}^+ \quad \forall i \in N,\ \forall k \in K \tag{A.6h}$$

Where the objective function (B.1a) consists of vehicle rebalancing cost and a

service availability function with a weight parameter α to minimize the difference between available vehicles and estimated demand in each sub-region. Constraints (B.1b) to (B.1d) define the relationship between available vehicles S_i^k, vacant vehicles V_i^k and occupied vehicles O_i^k. The maximum number of available vehicles that can be rebalanced is restricted by constraints (B.1e). Constraints (B.1f) impose the feasibility restrictions for rebalancing decisions, and the non-negativity of integer decision variables are guaranteed by constraints (B.1g) and (B.1h). To increase the computational efficiency while maintaining a satisfying solution, we further relax integer decision variables x_{ij}^k, S_i^k, V_i^k and O_i^k to positive real numbers $\mathbb{R}^+$.

The VR model proposed in this section is sufficient to show the benefit of integrating matching into the VR problem. When having different VR models with the area partitioning assumption, a matching-integrated version can always be constructed.

A.3 Optimal Assignment of Drivers to Customers

In this section, the driver-customer assignment problem implemented in the matching engine of the simulator is described. Within each matching decision time interval δ, let $\mathcal{R} = \{r_1, ..., r_n\}$ denote a set of waiting customers and $\mathcal{V} = \{v_1, ..., v_m\}$ represent a set of vacant vehicles in the system. Between a customer r_i and a vehicle v_j, let $\tau(r_i, v_j)$ indicate the minimum travel time for the vehicle to pick up the customer. The maximum pickup time for customers is denoted by $\bar{w}$. First, we construct a bipartite graph $G = (V, E)$, where $V = \mathcal{R} \cup \mathcal{V}$ and $E = \{e(r_i, v_j) : \forall r_i \in \mathcal{R}, \forall v_j \in \mathcal{V}, \tau(r_i, v_j) \leq \bar{w}\}$, meaning that an edge exists between a vehicle and a customer if the customer can be picked up by the vehicle within the maximum pickup time. The cost of each edge $e(r_i, v_j)$ equals to the pickup time, i.e., $c_{e(r_i,v_j)} = \tau(r_i, v_j)$. The decision variables for the optimal assignment problem are $x_{e(r_i,v_j)} \in \{0, 1\}$ for each edge $e(r_i, v_j) \in E$ in the bipartite graph G, and $y_{r_i} \in \{0, 1\}$ for each customer $r_i \in \mathcal{R}$. $x_{e(r_i,v_j)} = 1$ indicates that the customer r_i will be picked up by the vehicle v_j in the optimal assignment. $y_{r_i} = 1$ implies that the customer r_i will not be assigned to any vehicles during the current decision time interval δ. Let $\mathcal{I}(r_i)$ represent the set of

edges connected to a customer vertex r_i in G. Similarly, let $\mathcal{I}(v_j)$ indicate the set of edges connected to a driver vertex v_j in G. The optimal driver-customer assignment problem is:

$$\min \quad \sum_{e(r_i,v_j)\in E} c_{e(r_i,v_j)} x_{e(r_i,v_j)} + \gamma \cdot \sum_{r_i \in \mathcal{R}} y_{r_i} \tag{A.7a}$$

$$\text{s.t.} \quad \sum_{e(r_i,v_j)\in \mathcal{I}(v_j)} x_{e(r_i,v_j)} \leq 1 \quad \forall v_j \in \mathcal{V} \tag{A.7b}$$

$$\sum_{e(r_i,v_j)\in \mathcal{I}(r_i)} x_{e(r_i,v_j)} + y_{r_i} = 1 \quad \forall r_i \in \mathcal{R} \tag{A.7c}$$

$$x_{e(r_i,v_j)} \in \{0,1\} \quad \forall e(r_i,v_j) \in E \tag{A.7d}$$

$$y_{r_i} \in \{0,1\} \quad \forall r_i \in \mathcal{R} \tag{A.7e}$$

The objective function (C.1a) minimizes the summation of the total pickup time and penalties for unsatisfied requests, where γ stands for the penalty VMT for each unsatisfied customer. Constraints (C.1b) ensure that each vehicle can only be assigned to at most one customer. Constraints (C.1c) guarantee that each customer is either served by a vehicle or remained waiting during the current matching period. Constraints (C.1d) and (C.1e) make sure that the decision variables are binary. The optimal driver-customer assignment problem can be solved efficiently by the off-the-shelf ILP solvers (e.g., Gurobi) in the simulation.

A.4 Estimation of Regional Transition Matrix

In this section, the process for estimating the regional transition probability matrices for occupied and vacant vehicles, P, Q, P_v and Q_v, with the real travel time and demand data are described. There are several assumptions we made to generate these matrices:

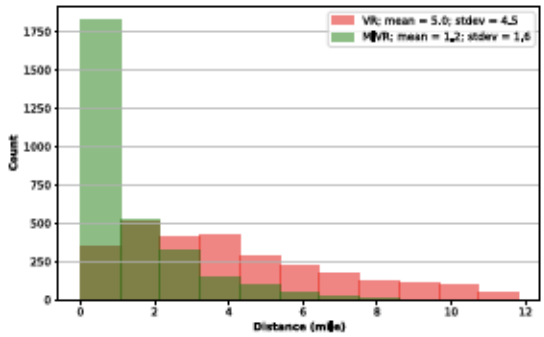
- Given a travel time and distance between the origin and the destination of a request, the vehicle travels with a constant speed.

- Given the origin and the destination of a request, the vehicle travels along the shortest path with regards to travel time.
- For vacant vehicles within sub-regions, 100% of vehicles remain in the same sub-region.

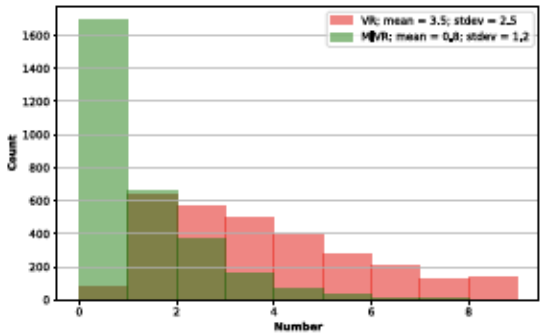
The detailed procedure is described as follows. First, the list of sub-regions crossed by the shortest path between each origin and destination pair was determined. The time spent within each sub-region for each origin-destination pair was weighted by the total demand to get the average time spent in each sub-region across all trips. For a given starting sub-region, the interzonal shortest paths, sub-region durations and origin-destination demand patterns were used to determine the likelihood of a given vehicle remaining in the starting sub-region, transitioning to a nearby sub-region or making a dropoff within a time interval. These probabilities were then used to populate P and Q. Because the taxi dataset only contains information about occupied vehicles, assumptions were made for the vacant vehicle zone transition probability matrices P_v and Q_v.

A.5 Benchmark VR Comparison Results for Different Demand Scenarios

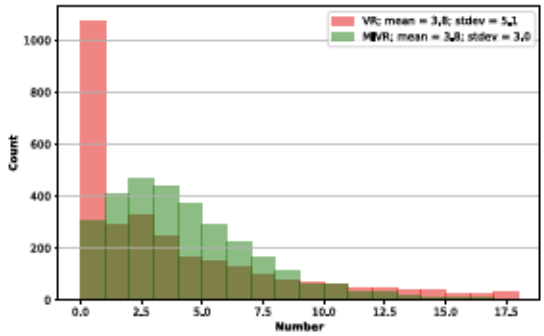
In this section, we provide the base case simulation results for four different demand scenarios: low demand with accurate estimation in Figure A-1, high demand with accurate estimation in Figure A-2, demand underestimation in Figure A-3 and demand overestimation in Figure A-4.



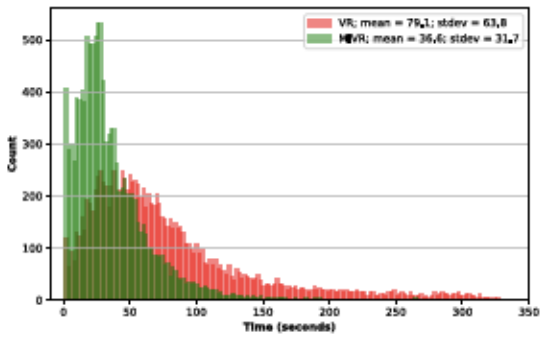
(a) Vehicle non-occupied travel distance distribution



(b) Vehicle rebalancing trip distribution

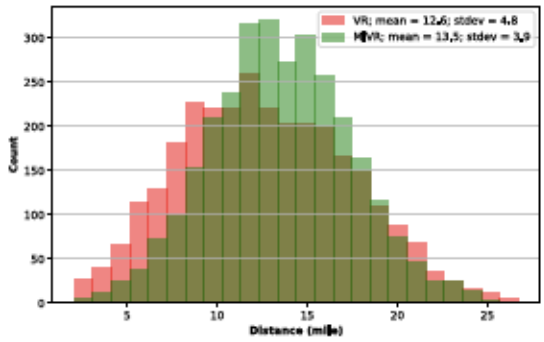


(c) Number of customers served (per vehicle) distribution

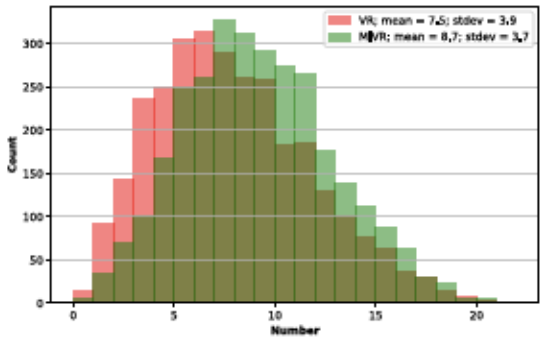


(d) Customer wait time distribution

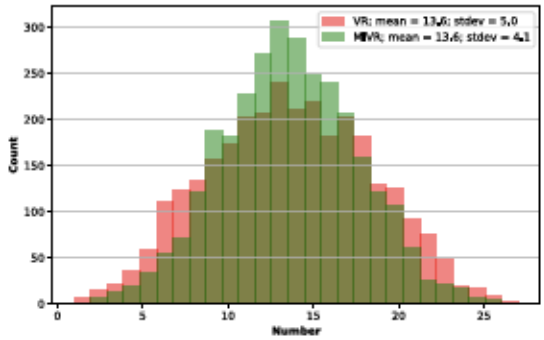
Figure A-1: Vehicle- and customer-related metrics in the simulation for the base case under the low demand with accurate estimation scenario (0 - 6).



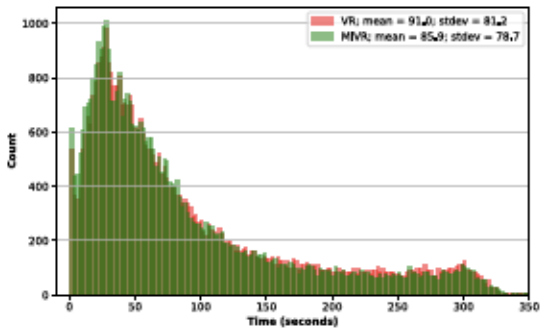
(a) Vehicle non-occupied travel distance distribution



(b) Vehicle rebalancing trip distribution

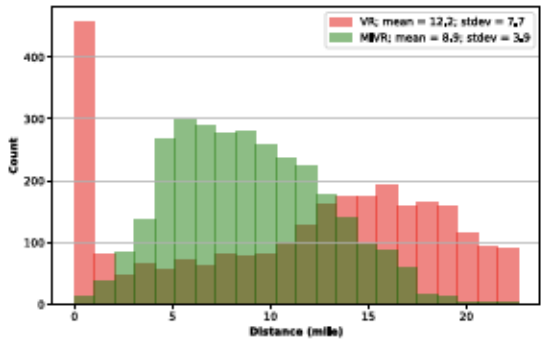


(c) Number of customers served (per vehicle) distribution

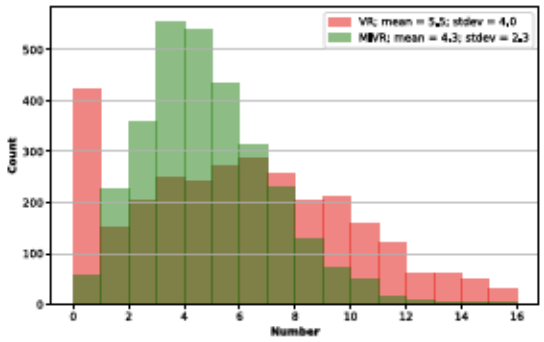


(d) Customer wait time distribution

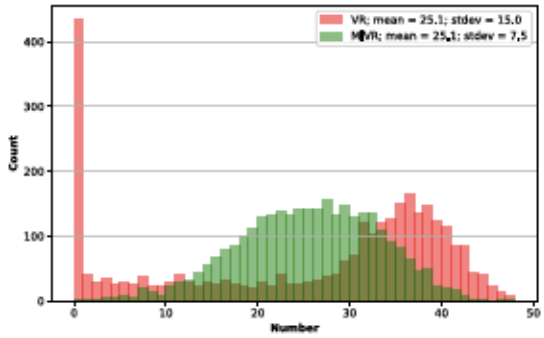
Figure A-2: Vehicle- and customer-related metrics in the simulation for the base case under the high demand with accurate estimation scenario (6 - 10).



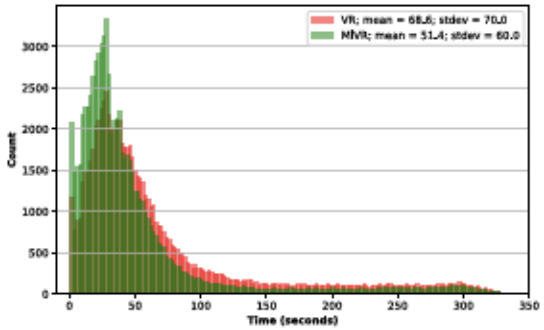
(a) Vehicle non-occupied travel distance distribution



(b) Vehicle rebalancing trip distribution

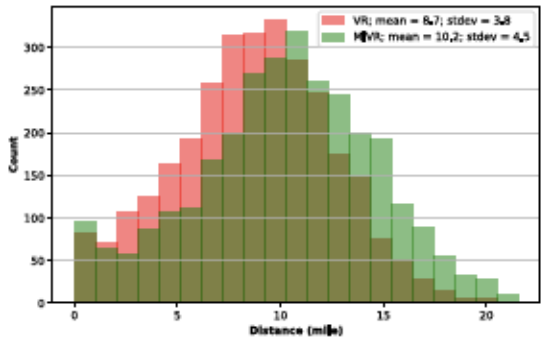


(c) Number of customers served (per vehicle) distribution

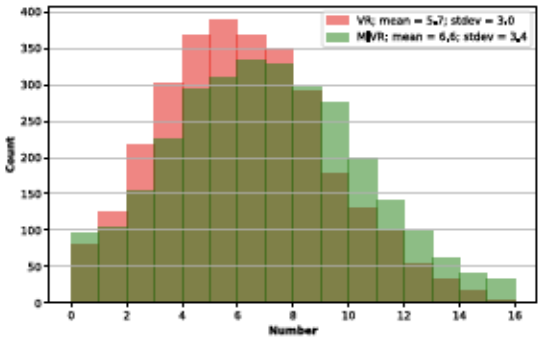


(d) Customer wait time distribution

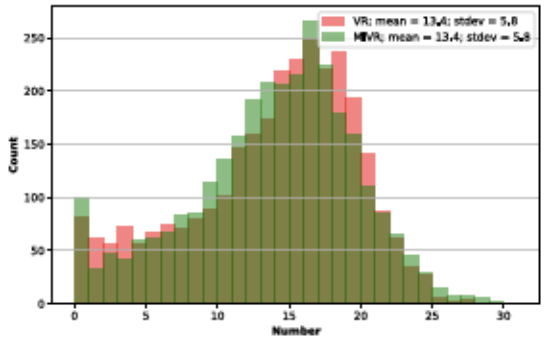
Figure A-3: Vehicle- and customer-related metrics in the simulation for the base case under demand underestimation scenario (11 - 17).



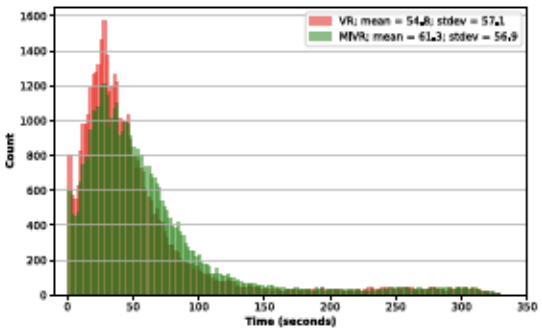
(a) Vehicle non-occupied travel distance distribution



(b) Vehicle rebalancing trip distribution



(c) Number of customers served (per vehicle) distribution



(d) Customer wait time distribution

Figure A-4: Vehicle- and customer-related metrics in the simulation for the base case under demand overestimation scenario (20 - 24).

Appendix B

Chapter 3 Appendix

B.1 Driver-Customer Matching Problem

In this section, the driver-customer matching problem utilized in the simulation for evaluating the performances of vehicle rebalancing models is described. Given locations for available vehicles $\mathcal{V} = \{v_1, ..., v_m\}$ and locations for customers who have requested a demand $\mathcal{R} = \{r_1, ..., r_n\}$, a driver-customer matching problem is solved to assign customer requests to drivers. Between a customer r_i and a vehicle v_j, let $d(r_i, v_j)$ and $\tau(r_i, v_j)$ represent the distance and travel time for picking up the customer, respectively. A customer will leave the system if the customer is not assigned to any drivers within the maximum wait time $\tilde{w}$.

To solve the driver-customer matching problem, we first construct a bipartite graph $G = (V, E)$, where $V = \mathcal{R} \cup \mathcal{V}$ and $E = \{e(r_i, v_j) : \forall r_i \in \mathcal{R}, \forall v_j \in \mathcal{V}, \tau(r_i, v_j) \leq \bar{w}\}$, indicating that an edge exists between a vehicle and a customer if the customer can be picked up by the vehicle within the maximum pickup time. The cost of each edge $e(r_i, v_j)$ equals to the pickup distance, i.e., $c_{e(r_i,v_j)} = d(r_i, v_j)$. The decision variables for the driver-customer matching problem are $x_{e(r_i,v_j)} \in \{0, 1\}$ for each edge $e(r_i, v_j) \in E$ in the bipartite graph G, and $y_{r_i} \in \{0, 1\}$ for each customer $r_i \in \mathcal{R}$. $x_{e(r_i,v_j)} = 1$ represents that the customer r_i will be picked up by the vehicle v_j in the optimal matching. $y_{r_i} = 1$ denotes that the customer r_i can not be satisfied. Let $\mathcal{I}(r_i)$ represent the set of edges connected to a customer vertex r_i in G. Similarly,

let $\mathcal{I}(v_j)$ indicate the set of edges connected to a driver vertex v_j in G. The optimal driver-customer matching problem is formulated as:

$$\min \quad \sum_{e(r_i,v_j)\in E} c_{e(r_i,v_j)} x_{e(r_i,v_j)} + \gamma \cdot \sum_{r_i \in \mathcal{R}} y_{r_i} \tag{B.1a}$$

$$\text{s.t.} \quad \sum_{e(r_i,v_j)\in \mathcal{I}(v_j)} x_{e(r_i,v_j)} \leq 1 \quad \forall v_j \in \mathcal{V} \tag{B.1b}$$

$$\sum_{e(r_i,v_j)\in \mathcal{I}(r_i)} x_{e(r_i,v_j)} + y_{r_i} = 1 \quad \forall r_i \in \mathcal{R} \tag{B.1c}$$

$$x_{e(r_i,v_j)} \in \{0,1\} \quad \forall e(r_i,v_j) \in E \tag{B.1d}$$

$$y_{r_i} \in \{0,1\} \quad \forall r_i \in \mathcal{R} \tag{B.1e}$$

The objective function (B.1a) minimizes a generalized cost for the driver-customer matching which consists of total pickup distance and penalties for unsatisfied customers. γ is the penalty parameter for each unsatisfied customer. Constraints (B.1b) guarantee that each vehicle can only be matched with at most one customer. Each customer is either assigned to a vehicle or remained to wait in the system, which is ensured by constraints (B.1c). Constraints (B.1d) and (B.1e) make sure that the decision variables are binary.

B.2 Uncertainty Set in the Robust MIVR Model

In this section, we briefly describe the uncertainty set utilized in the robust MIVR model, more details can be found in [67]. Given the future demand r_i^k of sub-region i at time k, the uncertainty set in the robust MIVR model consists of a box uncertainty set controlled by the parameter ρ and a polyhedral uncertainty set defined by the parameter Γ.

Let μ_i^k and σ_i^k represent the mean and standard deviation of the demand in sub-region i at time k from the historical data, respectively. The box uncertainty set $\tilde{\mathcal{U}}_i^k(\rho)$ is defined as

$$\tilde{\mathcal{U}}_i^k(\rho) = \left\{ r_i^k : \left| \frac{r_i^k - \mu_i^k}{\sigma_i^k} \right| \leq \rho \right\} \quad \forall i \in N,\ \forall k \in K,$$

where ρ indicates the parameter controlling the difference between historical average demand and future demand for each sub-region i at time k.

The polyhedral uncertainty set $\bar{\mathcal{U}}^k(\Gamma)$ is defined as

$$\bar{\mathcal{U}}^k(\Gamma) = \left\{ (r_1^k, ..., r_n^k) : \left| \sum_{i=1}^{n} (r_i^k - \mu_i^k) \right| \leq \Gamma \right\} \quad \forall k \in K,$$

where Γ denotes the parameter ensuring that the overall changes across all sub-regions at time k should not exceed Γ. Then the complete uncertainty set $\mathcal{U}$ used in the robust MIVR model is

$$\mathcal{U} = \left[\bigcap_{i=1}^{n} \bigcap_{k=1}^{\kappa} \tilde{\mathcal{U}}_i^k(\rho) \right] \cap \left[\bigcap_{k=1}^{\kappa} \mathcal{U}^k(\Gamma) \right].$$

Larger values of ρ and Γ lead to larger uncertainty set in the robust MIVR model, which leads to more conservative rebalancing decisions. Besides, decentralized vehicle rebalancing systems were proposed as contingency plans when AVs lost connections with central dispatch systems. Chen et al. [44] proposed a decentralized cooperative cruising method for offline operations of AMoD fleets. Their proposed method shows significant performance improvements compared to strategies with random-selected destinations for rebalancing AVs under different fleet sizes.

Appendix C

Chapter 5 Appendix

C.1 Derivation of the Robust Counterpart

In this section, we will derive the robust counterpart of the robust problem (5.13). In problem (5.13), there are two constraints with uncertain parameter ζ, Constraints (5.13b) and (5.13d). We reformulate two constraints as

$$C_1 + M \sum_{(o,d,t)\in\mathcal{F}} \left(\sigma_t^{o,d}\zeta_t^{o,d}\right) \leq \alpha, \quad \forall \boldsymbol{\zeta} \in \mathcal{U}(\Gamma), \tag{C.1a}$$

$$-\sigma_t^{o,d}\zeta_t^{o,d} \leq \mu_t^{o,d} - C_{t,2}^{o,d}, \quad \forall (o,d,t) \in \mathcal{F}, \forall \boldsymbol{\zeta} \in \mathcal{U}(\Gamma), \tag{C.1b}$$

where

$$\begin{aligned} C_1 = &\sum_{(o,d,t)\in\mathcal{F}} \sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \left(w_{t,\tau}^{o,d,p,v} + \gamma\phi^{o,d,p}\right) \lambda_{t,\tau}^{o,d,p,v} \\ &- M \sum_{(o,d,t)\in\mathcal{F}} \sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \lambda_{t,\tau}^{o,d,p,v} + M \sum_{(o,d,t)\in\mathcal{F}} \mu_t^{o,d}, \\ C_{t,2}^{o,d} = &\sum_{v\in\mathcal{V}} \sum_{p\in\mathcal{P}^{o,d}} \sum_{\tau=\tau_t^{o,d,p}}^{T} \lambda_{t,\tau}^{o,d,p,v}, \quad \forall (o,d,t) \in \mathcal{F}. \end{aligned}$$

The uncertainty set

$$\mathcal{U}(\Gamma) = \{\boldsymbol{\zeta} : \|\boldsymbol{\zeta}\|_\infty \leq 1, \|\boldsymbol{\zeta}\|_1 \leq \Gamma\}$$

can be reformulated as

$$\zeta_t^{o,d} \leq 1, \quad \forall (o,d,t) \in \mathcal{F},$$
$$\sum_{(o,d,t)\in\mathcal{F}} \zeta_t^{o,d} \leq \Gamma.$$

Let $\zeta_t^{o,d} = \rho_t^{o,d,+} - \rho_t^{o,d,-}$ given two auxiliary non-negative variables $\rho_t^{o,d,+} \geq 0, \rho_t^{o,d,-} \geq 0$, we have $\zeta_t^{o,d} = \rho_t^{o,d,+} + \rho_t^{o,d,-}$ and the uncertainty set can be formulated as

$$\rho_t^{o,d,+} + \rho_t^{o,d,-} \leq 1, \quad \forall (o,d,t) \in \mathcal{F},$$
$$\sum_{(o,d,t)\in\mathcal{F}} \left(\rho_t^{o,d,+} + \rho_t^{o,d,-}\right) \leq \Gamma.$$

For the Constraints (C.1a), we can rewrite it as

$$C_1 + \max_{\zeta\in\mathcal{U}(\Gamma)} \left\{ M \sum_{(o,d,t)\in\mathcal{F}} \sigma_t^{o,d} \zeta_t^{o,d} \right\} \leq \alpha. \tag{C.2}$$

The second term in Equation (C.2) can be written as an optimization problem while replacing $\zeta_t^{o,d}$ with the previous definition:

$$\max \quad \sum_{(o,d,t)\in\mathcal{F}} M\sigma_t^{o,d}(\rho_t^{o,d,+} - \rho_t^{o,d,-}) \tag{C.3a}$$
$$\text{s.t.} \quad \rho_t^{o,d,+} + \rho_t^{o,d,-} \leq 1, \quad \forall (o,d,t) \in \mathcal{F}, \tag{C.3b}$$
$$\sum_{(o,d,t)\in\mathcal{F}} \left(\rho_t^{o,d,+} + \rho_t^{o,d,-}\right) \leq \Gamma, \tag{C.3c}$$
$$\rho_t^{o,d,+} \geq 0, \quad \forall (o,d,t) \in \mathcal{F}, \tag{C.3d}$$
$$\rho_t^{o,d,-} \geq 0, \quad \forall (o,d,t) \in \mathcal{F}. \tag{C.3e}$$

Taking the dual of problem (C.4), we have

$$\min \quad \sum_{(o,d,t)\in\mathcal{F}} \nu_1^{o,d,t} + \Gamma \cdot \nu_2 \tag{C.4a}$$
$$\text{s.t.} \quad \nu_1^{o,d,t} + \nu_2 \geq M\sigma_t^{o,d}, \quad \forall (o,d,t) \in \mathcal{F} \tag{C.4b}$$
$$\nu_1^{o,d,t} + \nu_2 \geq -M\sigma_t^{o,d}, \quad \forall (o,d,t) \in \mathcal{F} \tag{C.4c}$$

$$\nu_1^{o,d,t} \geq 0, \quad \forall (o,d,t) \in \mathcal{F} \tag{C.4d}$$

$$\nu_2 \geq 0, \tag{C.4e}$$

where $\nu_1^{o,d,t}$ and ν_2 are dual variables. Therefore, the Constraints (C.2) can be reformulated by plugging in the minimization problem (C.4):

$$C_1 + \sum_{(o,d,t)\in\mathcal{F}} \nu_1^{o,d,t} + \Gamma \cdot \nu_2 \leq \alpha, \tag{C.5a}$$

$$\nu_1^{o,d,t} + \nu_2 \geq M\sigma_t^{o,d}, \quad \forall (o,d,t) \in \mathcal{F} \tag{C.5b}$$

$$\nu_1^{o,d,t} + \nu_2 \geq -M\sigma_t^{o,d}, \quad \forall (o,d,t) \in \mathcal{F} \tag{C.5c}$$

$$\nu_1^{o,d,t} \geq 0, \quad \forall (o,d,t) \in \mathcal{F} \tag{C.5d}$$

$$\nu_2 \geq 0, \tag{C.5e}$$

and we have the robust counterpart for Constraints (5.13b).

Similarly, for Constraints (C.1b), we can rewrite it as

$$\max_{\zeta\in\mathcal{U}(\Gamma)} \left\{ -\sigma_t^{o,d}\zeta_t^{o,d} \right\} \leq \mu_t^{o,d} - C_{t,2}^{o,d}, \quad (o,d,t) \in \mathcal{F}. \tag{C.6}$$

For each (o,d,t) pair, the first term in Equation (C.6) can be written as an optimization problem using the same replacement for uncertain parameter $\zeta_t^{o,d}$:

$$\max \quad -\sigma_t^{o,d}(\rho_t^{o,d,+} - \rho_t^{o,d,-}) \tag{C.7a}$$

$$\text{s.t.} \quad \rho_{t'}^{o',d',+} + \rho_{t'}^{o',d',-} \leq 1, \quad \forall (o',d',t') \in \mathcal{F}, \tag{C.7b}$$

$$\sum_{(o',d',t')\in\mathcal{F}} \left(\rho_{t'}^{o',d',+} + \rho_{t'}^{o',d',-} \right) \leq \Gamma, \tag{C.7c}$$

$$\rho_{t'}^{o',d',+} \geq 0, \quad \forall (o',d',t') \in \mathcal{F}, \tag{C.7d}$$

$$\rho_{t'}^{o',d',-} \geq 0, \quad \forall (o',d',t') \in \mathcal{F}. \tag{C.7e}$$

Taking the dual of problem (C.8), we have

$$\min \quad \sum_{(o',d',t')\in\mathcal{F}} \nu_{o',d',t',3}^{o,d,t} + \nu_4^{o,d,t} \tag{C.8a}$$

$$\text{s.t.} \quad \nu_{o,d,t,3}^{o,d,t} + \nu_4^{o,d,t} \geq -\sigma_t^{o,d}, \tag{C.8b}$$

$$\nu^{o,d,t}_{o,d,t,3} + \nu^{o,d,t}_{4} \geq \sigma^{o,d}_{t}, \tag{C.8c}$$

$$\nu^{o,d,t}_{o',d',t',3} + \nu^{o,d,t}_{4} \geq 0, \quad (o',d',t') \in \mathcal{F} \neq (o,d,t), \tag{C.8d}$$

$$\nu^{o,d,t}_{o',d',t',3} \geq 0, \quad \forall (o',d',t') \in \mathcal{F}, \tag{C.8e}$$

$$\nu^{o,d,t}_{4} \geq 0, \tag{C.8f}$$

where $\nu^{o,d,t}_{o',d',t',3}$ and $\nu^{o,d,t}_{4}$ are dual variables. The Constraints (C.6) can then be rewritten by inserting the minimization problem (C.8) for each (o,d,t) pair:

$$\sum_{(o',d',t') \in \mathcal{F}} \nu^{o,d,t}_{o',d',t',3} + \nu^{o,d,t}_{4} \leq \mu^{o,d}_{t} - C^{o,d}_{t,2}, \quad \forall (o,d,t) \in \mathcal{F} \tag{C.9a}$$

$$\nu^{o,d,t}_{o,d,t,3} + \nu^{o,d,t}_{4} \geq -\sigma^{o,d}_{t}, \quad \forall (o,d,t) \in \mathcal{F}, \tag{C.9b}$$

$$\nu^{o,d,t}_{o,d,t,3} + \nu^{o,d,t}_{4} \geq \sigma^{o,d}_{t}, \quad \forall (o,d,t) \in \mathcal{F}, \tag{C.9c}$$

$$\nu^{o,d,t}_{o',d',t',3} + \nu^{o,d,t}_{4} \geq 0, \quad (o',d',t') \neq (o,d,t) \in \mathcal{F}, \tag{C.9d}$$

$$\nu^{o,d,t}_{o',d',t',3} \geq 0, \quad \forall (o,d,t), (o',d',t') \in \mathcal{F}, \tag{C.9e}$$

$$\nu^{o,d,t}_{4} \geq 0, \quad \forall (o,d,t) \in \mathcal{F}, \tag{C.9f}$$

and we have the robust counterpart for Constraints (5.13d).

C.2 Robust Transit Schedules

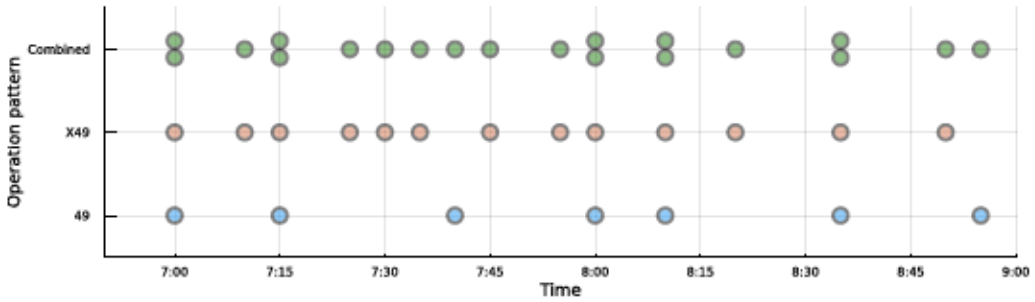


Figure C-1: The robust transit schedule with $\Gamma = 0$.

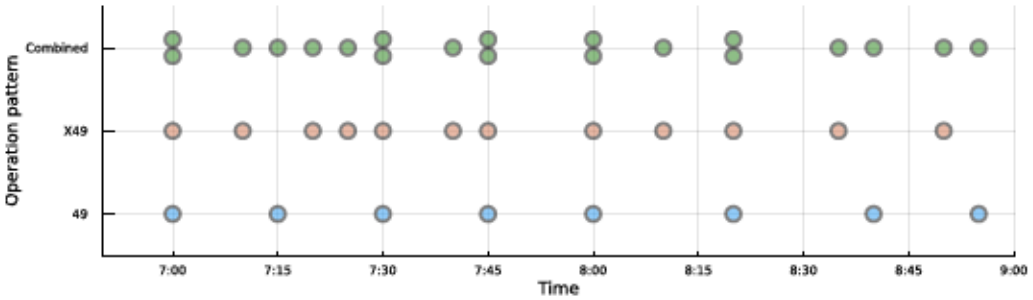


Figure C-2: The robust transit schedule with $\Gamma = 1$.

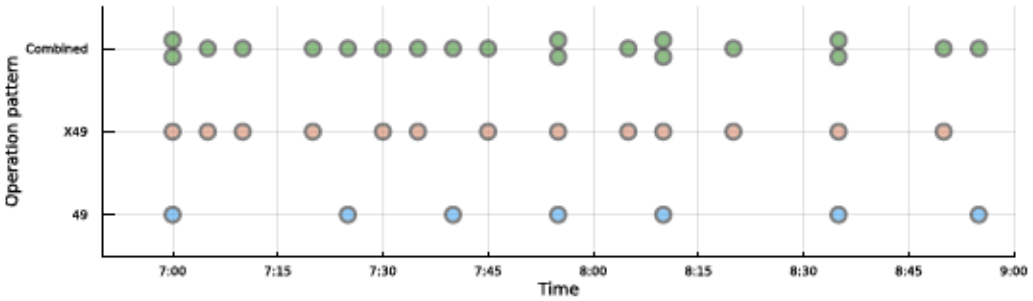


Figure C-3: The robust transit schedule with $\Gamma = 2$.

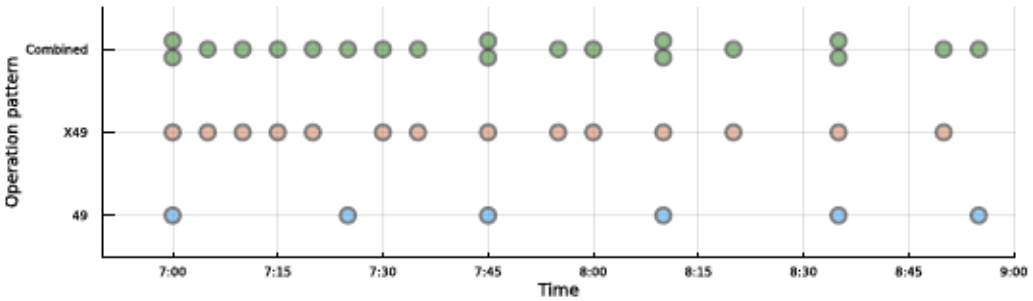


Figure C-4: The robust transit schedule with $\Gamma = 3$.

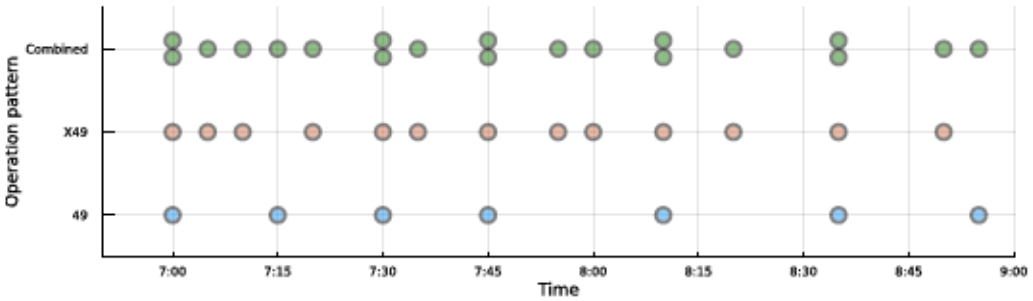


Figure C-5: The robust transit schedule with $\Gamma = 4$.

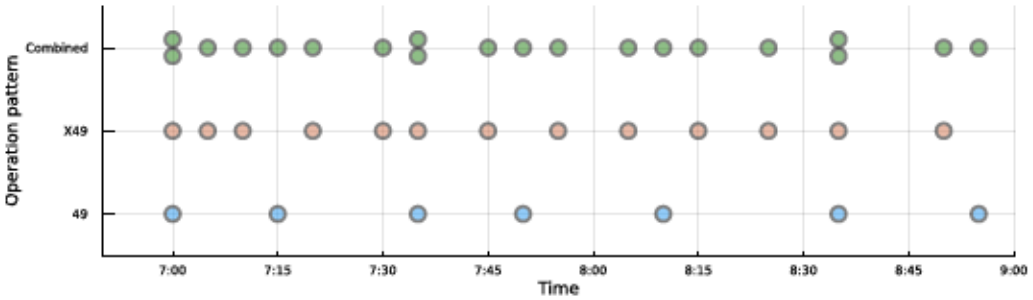


Figure C-6: The robust transit schedule with $\Gamma = 5$.

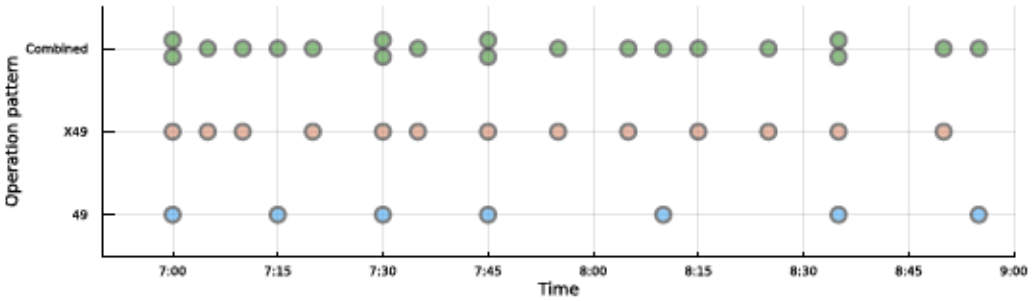


Figure C-7: The robust transit schedule with $\Gamma = 6$.

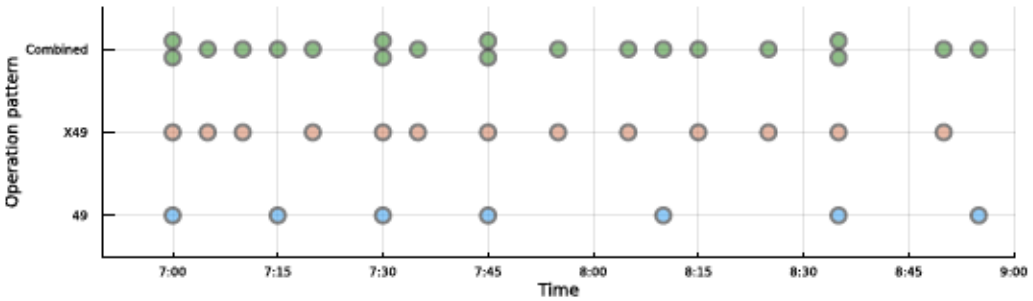


Figure C-8: The robust transit schedule with $\Gamma = 7$.

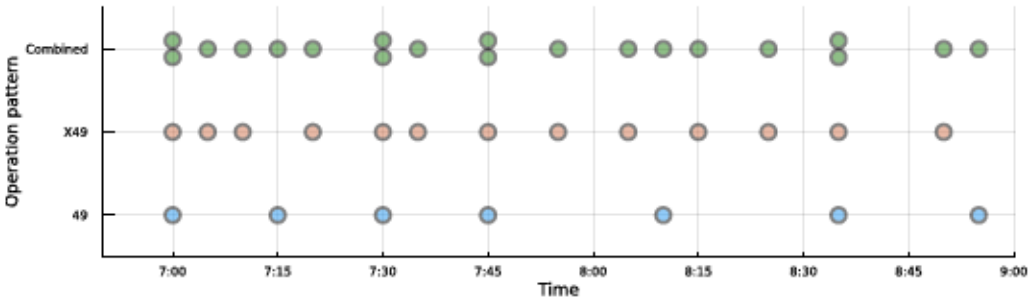


Figure C-9: The robust transit schedule with $\Gamma = 8$.

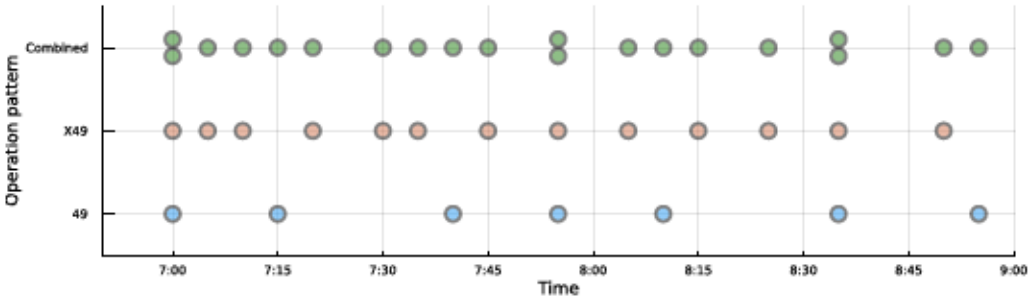


Figure C-10: The robust transit schedule with $\Gamma = 9$.

Bibliography

[1] Niels Agatz, Alan Erera, Martin Savelsbergh, and Xing Wang. Optimization for dynamic ride-sharing: A review. *European Journal of Operational Research*, 223(2):295–303, 2012.

[2] Lina Al-Kanj, Juliana Nascimento, and Warren B Powell. Approximate dynamic programming for planning a ride-hailing system using autonomous fleets of electric vehicles. *European Journal of Operational Research*, 284(3):1088–1106, 2020.

[3] Javier Alonso-Mora, Samitha Samaranayake, Alex Wallar, Emilio Frazzoli, and Daniela Rus. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences*, 114(3):462, Jan 2017.

[4] American Public Transportation Association. APTA ridership trends, 2022. https://transitapp.com/apta.

[5] James Atwood and Don Towsley. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[6] Jiaru Bai, Kut C. So, Christopher S. Tang, Xiqun (Michael) Chen, and Hai Wang. Coordinating supply and demand on an on-demand service platform with impatient customers. *Manufacturing & Service Operations Management*, 21(3):556–570, Jun 2018.

[7] Roberto Baldacci, Vittorio Maniezzo, and Aristide Mingozzi. An exact method for the car pooling problem based on lagrangean column generation. *Operations Research*, 52(3):422–439, Jun 2004.

[8] Siddhartha Banerjee, Daniel Freund, and Thodoris Lykouris. Pricing and optimization in shared vehicle systems: An approximation framework. *Operations Research*, 70(3):1783–1805, 2022.

[9] Siddhartha Banerjee, Chamsi Hssaine, Noémie Périvier, and Samitha Samaranayake. Real-time approximate routing for smart transit systems, 2021.

[10] Jose Maria Barrero, Nicholas Bloom, and Steven J Davis. Why working from home will stick. Working Paper 28731, National Bureau of Economic Research, April 2021.

[11] Moshe Ben-Akiva and Michel Bierlaire. *Discrete Choice Models with Applications to Departure Time and Route Choice*, pages 7–37. Springer US, Boston, MA, 2003.

[12] A. Ben-Tal, Laurent El Ghaoui, and A. S. Nemirovskiĭ. *Robust optimization*. Princeton series in applied mathematics. Princeton University Press, 2009.

[13] Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

[14] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.

[15] Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[16] Dimitris Bertsimas and Dick den Hertog. *Robust and adaptive optimization*. Dynamic Ideas LLC, Belmont, Massachusetts, 2020.

[17] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, Jul 2017.

[18] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, Feb 2018.

[19] Dimitris Bertsimas, Patrick Jaillet, and Sébastien Martin. Online vehicle routing: The edge of optimization in large-scale applications. *Operations Research*, 67(1):143–162, 2019.

[20] Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

[21] Dimitris Bertsimas, Yee Sian Ng, and Julia Yan. Joint frequency-setting and pricing optimization on multimodal transit networks at scale. *Transportation Science*, 54(3):839–853, May 2020.

[22] Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[23] Dimitris Bertsimas, Melvyn Sim, and Meilin Zhang. Adaptive distributionally robust optimization. *Management Science*, 65(2):604–618, 2019.

[24] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. (arXiv:1901.04562), Jan 2019. arXiv:1901.04562 [cs, stat].

[25] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[26] Tierra S. Bills and Joan L. Walker. Looking beyond the mean for equity analysis: Examining distributional impacts of transportation improvements. *Transport Policy*, 54:61–69, February 2017.

[27] Reuben Binns. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR, January 2018. ISSN: 2640-3498.

[28] John R. Birge and François Louveaux. *Introduction to Stochastic Programming*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2011.

[29] Eszter Bokányi and Anikó Hannák. Understanding inequalities in ride-hailing services through simulations. *Scientific Reports*, 10, 12 2020.

[30] Anton Braverman, Jim G Dai, Xin Liu, and Lei Ying. Empty-car routing in ridesharing systems. *Operations Research*, 67(5):1437–1452, 2019.

[31] Brett Helling. How much does uber cost? fare pricing, rates, and cost estimates explained, 2024.

[32] Eliot Brown. The Ride-Hail Utopia That Got Stuck in Traffic, February 2020.

[33] Gérard P. Cachon, Kaitlin M. Daniels, and Ruben Lobel. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19(3):368–384, Jun 2017.

[34] Yi Cao, Shan Wang, and Jinyang Li. The optimization model of ride-sharing route for ride hailing considering both system optimization and user fairness. *Sustainability (Switzerland)*, 13:1–17, 1 2021.

[35] Nicholas S. Caros, Xiaotong Guo, Anson Stewart, John Attanucci, Nicholas Smith, Dimitris Nioras, Anna Gartsman, and Alissa Zimmer. Ridership and operations visualization engine: An integrated transit performance and passenger journey visualization engine. *Transportation Research Record*, 2022.

[36] Nicholas S. Caros, Xiaotong Guo, and Jinhua Zhao. The emerging spectrum of flexible work locations: implications for travel demand and carbon emissions, 2023.

[37] Nicholas S. Caros, Xiaotong Guo, Yunhan Zheng, and Jinhua Zhao. The impacts of remote work on travel: insights from nearly three years of monthly surveys, 2023.

[38] Ennio Cascetta, Armando Cartenì, Ilaria Henke, and Francesca Pagliara. Economic growth, transport accessibility and regional equity impacts of high-speed railways in Italy: ten years ex post evaluation and future perspectives. *Transportation Research Part A: Policy and Practice*, 139:412–428, September 2020.

[39] Juan Camilo Castillo, Dan Knoepfle, and Glen Weyl. Surge pricing solves the wild goose chase. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, page 241–242, New York, NY, USA, 2017. Association for Computing Machinery.

[40] Juan Camilo Castillo, Daniel T. Knoepfle, and E. Glen Weyl. Matching in Ride Hailing: Wild Goose Chases and How to Solve Them. *SSRN Electronic Journal*, 2022.

[41] Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey, October 2020. arXiv:2010.04053 [cs, stat].

[42] Avishai Ceder. *Public Transit Planning and Operation*. 2007.

[43] Avishai Ceder and Nigel H.M. Wilson. Bus network design. *Transportation Research Part B: Methodological*, 20(4):331–344, 1986.

[44] Linji Chen, Amir Hosein Valadkhani, and Mohsen Ramezani. Decentralised cooperative cruising of autonomous ride-sourcing fleets. *Transportation Research Part C: Emerging Technologies*, 131(May):103336, 2021.

[45] Rongsheng Chen and Michael W Levin. Dynamic user equilibrium of mobility-on-demand system with linear programming rebalancing strategy. *Transportation Research Record*, 2673(1):447–459, 2019.

[46] Maxime C Cohen and Renyu Zhang. Competition and coopetition for two-sided platforms. *SSRN Working Paper No. 3028138*, page 48, 2017.

[47] Dragoš Cvetković, Peter Rowlinson, and Slobodan Simić. An introduction to the theory of graph spectra. *(No Title)*, 2009.

[48] Florian Dandl, Michael Hyland, Klaus Bogenberger, and Hani S Mahmassani. Evaluating the impact of spatio-temporal demand forecast aggregation on the operational performance of shared autonomous mobility fleets. *Transportation*, 46(6):1975–1996, 2019.

[49] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering, February 2017. arXiv:1606.09375 [cs, stat].

[50] Mi Diao, Hui Kong, and Jinhua Zhao. Impacts of transportation network companies on urban mobility. *Nature Sustainability*, 4(6):494–500, 2021.

[51] Edsger W Dijkstra et al. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[52] Iain Dunning, Joey Huchette, and Miles Lubin. Jump: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017.

[53] Marvin Erdmann, Florian Dandl, and Klaus Bogenberger. Dynamic car-passenger matching based on tabu search using global optimization with time windows. In *2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO)*, pages 1–5. IEEE, 2019.

[54] Reza Zanjirani Farahani, Elnaz Miandoabchi, W. Y. Szeto, and Hannaneh Rashidi. A review of urban transportation network design problems. *European Journal of Operational Research*, 229(2):281–302, 2013.

[55] Peter G. Furth and Nigel H.M. Wilson. Setting Frequencies on Bus Routes: Theory and Practice. *Transportation Research Record*, pages 1–7, 1981.

[56] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. Spatiotemporal Multi-Graph Convolution Network for Ride-Hailing Demand Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3656–3663, 2019.

[57] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, page 1263–1272. PMLR, Jul 2017.

[58] K Gkiotsalitis, M Schmidt, and E Van Der Hurk. Subline frequency setting for autonomous minibusses under demand uncertainty. 2021.

[59] Gregory A. Godfrey and Warren B. Powell. An adaptive dynamic programming algorithm for dynamic fleet management, i: Single period travel times. *Transportation Science*, 36(1):21–39, Feb 2002.

[60] Gregory A. Godfrey and Warren B. Powell. An adaptive dynamic programming algorithm for dynamic fleet management, ii: Multiperiod travel times. *Transportation Science*, 36(1):40–54, Feb 2002.

[61] Bram L. Gorissen, Ihsan Yanıkoğlu, and Dick den Hertog. A practical guide to robust optimization. *Omega*, 53:124–137, Jun 2015. arXiv: 1501.02634.

[62] Grand View Research. Global Ride Hailing Services Market Size & Share Report, 2030, January 2023.

[63] Robert Grone, Russell Merris, and V.S. Sunder. The laplacian spectrum of a graph. *SIAM Journal on matrix analysis and applications*, 11(2):218–238, 1990.

[64] Maxime Guériau and Ivana Dusparic. SAMoD: Shared autonomous mobility-on-demand using decentralized reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1558–1563. IEEE, 2018.

[65] Ge Guo, Wei Yuan, Jinyuan Liu, Yisheng Lv, and Wei Liu. Traffic forecasting via dilated temporal convolution with peak-sensitive loss. *IEEE Intelligent Transportation Systems Magazine*, 15:48–57, 1 2023.

[66] Ge Guo and Tianqi Zhang. A residual spatio-temporal architecture for travel demand forecasting. *Transportation Research Part C: Emerging Technologies*, 115, 6 2020.

[67] Xiaotong Guo, Nicholas S. Caros, and Jinhua Zhao. Robust matching-integrated vehicle rebalancing in ride-hailing system with uncertain demand. *Transportation Research Part B: Methodological*, 150:161–189, 2021.

[68] Xiaotong Guo, Andreas Haupt, Hai Wang, Rida Qadri, and Jinhua Zhao. Understanding multi-homing and switching by platform drivers. *Transportation Research Part C: Emerging Technologies*, 154:104233, 2023.

[69] Xiaotong Guo, Baichuan Mo, Haris N. Koutsopoulos, Shenhao Wang, and Jinhua Zhao. Transit frequency setting problem with demand uncertainty, 2022.

[70] Xiaotong Guo, Ao Qu, Hongmou Zhang, Peyman Noursalehi, and Jinhua Zhao. Dissolving the segmentation of a shared mobility market: A framework and four market structure designs. *Transportation Research Part C: Emerging Technologies*, 157:104397, 2023.

[71] Xiaotong Guo, Qingyi Wang, and Jinhua Zhao. Data-driven vehicle rebalancing with predictive prescriptions in the ride-hailing system. *IEEE Open Journal of Intelligent Transportation Systems*, 3:251–266, 2022.

[72] Xiaotong Guo, Hanyong Xu, Dingyi Zhuang, Yunhan Zheng, and Jinhua Zhao. Fairness-enhancing vehicle rebalancing in the ride-hailing system, 2023.

[73] Xiaotong Guo and Jinhua Zhao. Design of transit-centric multimodal urban mobility system with autonomous mobility-on-demand, 2024.

[74] Gurobi Optimization, LLC. Gurobi optimizer reference manual, 2023.

[75] Jonathan V. Hall and Alan B. Krueger. An Analysis of the Labor Market for Uber's Driver-Partners in the United States. *ILR Review*, 71(3):705–732, 2018.

[76] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[77] Richard A Harshman et al. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.

[78] Long He, Zhenyu Hu, and Meilin Zhang. Robust repositioning for vehicle sharing. *Manufacturing & Service Operations Management*, 2019.

[79] Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

[80] Sin C Ho, Wai Yuen Szeto, Yong-Hong Kuo, Janny MY Leung, Matthew Petering, and Terence WH Tou. A survey of dial-a-ride problems: Literature review and recent developments. *Transportation Research Part B: Methodological*, 111:395–421, 2018.

[81] Zhengfeng Huang, Gang Ren, and Haixu Liu. Optimizing bus frequencies under uncertain demand: Case study of the transit network in a developing city. *Mathematical Problems in Engineering*, 2013, 2013.

[82] Michael Hyland, Charlotte Frei, Andreas Frei, and Hani S. Mahmassani. Riders on the storm: Exploring weather and seasonality effects on commute mode choice in chicago. *Travel Behaviour and Society*, 13:44–60, 10 2018. VOT.

[83] O.J. Ibarra-Rojas, F. Delgado, R. Giesen, and J.C. Muñoz. Planning, operation, and control of bus transport systems: A literature review. *Transportation Research Part B: Methodological*, 77:38–75, Jul 2015.

[84] Ramon Iglesias, Federico Rossi, Kevin Wang, David Hallac, Jure Leskovec, and Marco Pavone. Data-driven model predictive control of autonomous mobility-on-demand systems. *arXiv:1709.07032 [cs, stat]*, Sep 2017. arXiv: 1709.07032.

[85] Young-Seon Jeong, Young-Ji Byon, Manoel Mendonca Castro-Neto, and Said M. Easa. Supervised weighting-online learning algorithm for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1700–1707, Dec 2013.

[86] Jingjing Jiang. More Americans are using ride-hailing apps, January 2019.

[87] Weiwei Jiang and Jiayun Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921, November 2022.

[88] Yan Jiao, Xiaocheng Tang, Zhiwei Qin, Shuaiji Li, Fan Zhang, Hongtu Zhu, and Jieping Ye. Real-world ride-hailing vehicle repositioning using deep reinforcement learning. *arXiv:2103.04555 [cs]*, Jul 2021. arXiv: 2103.04555.

[89] Ishan Jindal, Zhiwei Tony Qin, Xuewen Chen, Matthew Nokleby, and Jieping Ye. Optimizing taxi carpool policies via reinforcement learning and spatio-temporal mining. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1417–1426. IEEE, 2018.

[90] Jintao Ke, Siyuan Feng, Zheng Zhu, Hai Yang, and Jieping Ye. Joint predictions of multi-modal ride-hailing demands: A deep multi-task multi-graph learning-based approach. *Transportation Research Part C: Emerging Technologies*, 127:103063, 2021.

[91] Jintao Ke, Hai Yang, Xinwei Li, Hai Wang, and Jieping Ye. Pricing and equilibrium in on-demand ride-pooling markets. *Transportation Research Part B: Methodological*, 139:411–431, Sep 2020.

[92] Jintao Ke, Hongyu Zheng, Hai Yang, and Xiqun (Michael) Chen. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Research Part C: Emerging Technologies*, 85(June):591–608, 2017.

[93] Zemian Ke and Sean Qian. Leveraging ride-hailing services for social good: Fleet optimal routing and system optimal pricing. *Transportation Research Part C: Emerging Technologies*, 155, 10 2023.

[94] Dara Kerr. Lyft grows gangbusters in 2017, bringing competition to Uber, January 2018.

[95] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pages 1–14, 2017.

[96] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks, February 2017. arXiv:1609.02907 [cs, stat].

[97] Anton J. Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, Jan 2002.

[98] Pramesh Kumar and Alireza Khani. Planning of integrated mobility-on-demand and urban transit networks. *Transportation Research Part A: Policy and Practice*, 166(March):499–521, 2022.

[99] Richard C Larson and Amadeo R Odoni. *Urban operations research*. Prentice-Hall, NJ, 1981.

[100] Dongyuan Zhan Leon Yang Chu, Zhixi Wan. Harnessing the double-edged sword via routing: Information provision on ride-hailing platforms. *Available at SSRN 3266250*, 2018.

[101] Wilhelm Lerner. The future of urban mobility: towards networked, multimodal cities of 2050, 2018.

[102] Elyse O'Callaghan Lewis, Don MacKenzie, and Jessica Kaminsky. Exploring equity: How equity norms have been applied implicitly and explicitly in transportation research and practice. *Transportation Research Interdisciplinary Perspectives*, 9:100332, March 2021.

[103] Can Li, Lei Bai, Wei Liu, Lina Yao, and S Travis Waller. A multi-task memory network with knowledge adaptation for multimodal demand forecasting. *Transportation Research Part C: Emerging Technologies*, 131:103352, 2021.

[104] Xiaolong Li, Gang Pan, Zhaohui Wu, Guande Qi, Shijian Li, Daqing Zhang, Wangsheng Zhang, and Zonghui Wang. Prediction of urban human mobility using large-scale taxi traces and its applications. *Frontiers of Computer Science*, 6(1):111–121, 2012.

[105] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. Feb 2018. MAG ID: 2963358464.

[106] Yanhong Li, Wangtu Xu, and Shiwei He. Expected value model for optimizing the multiple bus headways. *Applied Mathematics and Computation*, 219(11):5849–5861, 2013.

[107] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2015.

[108] Marco Lippi, Matteo Bertini, and Paolo Frasconi. Short-Term Traffic Flow Forecasting: An Experimental Comparison of Time-Series Analysis and Supervised Learning. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):871–882, June 2013. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[109] Todd Alexander Litman. Evaluating Transportation Equity. March 2023.

[110] Jiachao Liu, Wei Ma, and Sean Qian. Optimal curbside pricing for managing ride-hailing pick-ups and drop-offs. *Transportation Research Part C: Emerging Technologies*, 146:103960, 2023.

[111] Yifang Liu, Will Skinner, and Chongyuan Xiang. Globally-optimized realtime supply-demand matching in on-demand ridesharing. In *The World Wide Web Conference*, pages 3034–3040, 2019.

[112] Andrew W. Lo and Mark T. Mueller. Warning: Physics envy may be hazardous to your wealth!, 2010.

[113] Youshui Lu, Yong Qi, Saiyu Qi, Yue Li, Hongyu Song, and Yuhao Liu. Say no to price discrimination: Decentralized and automated incentives for price auditing in ride-hailing services. *IEEE Transactions on Mobile Computing*, 21:663–680, 2 2022.

[114] Qi Luo, Shukai Li, and Robert C. Hampshire. Optimal design of intermodal mobility networks under uncertainty: Connecting micromobility with mobility-on-demand transit. *EURO Journal on Transportation and Logistics*, 10(June 2020):100045, 2021.

[115] Qi Luo, Samitha Samaranayake, and Siddhartha Banerjee. Multimodal mobility systems: joint optimization of transit network design and pricing. In *Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems*, ICCPS '21, page 121–131, New York, NY, USA, 2021. Association for Computing Machinery.

[116] Guodong Lyu, Wang Chi Cheung, Chung-Piaw Teo, and Hai Wang. Multi-objective online ride-matching. *Available at SSRN 3356823*, 2019.

[117] Changxi Ma, Wei Hao, Ruichun He, Xiaoyan Jia, Fuquan Pan, Jing Fan, and Ruiqi Xiong. Distribution path robust optimization of electric vehicle with multiple distribution centers. *PloS One*, 13(3), 2018.

[118] Christopher MacKechnie. How much does a bus cost to purchase and operate?, 2019.

[119] Daniel McFadden. The measurement of urban travel demand. *Journal of Public Economics*, 3(4):303–328, 1974.

[120] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning, January 2022. arXiv:1908.09635 [cs].

[121] Fei Miao, Shuo Han, Abdeltawab M. Hendawi, Mohamed E Khalefa, John A. Stankovic, and George J. Pappas. Data-driven distributionally robust vehicle balancing using dynamic region partitions. In *Proceedings of the 8th International Conference on Cyber-Physical Systems - ICCPS '17*, page 261–271. ACM Press, 2017.

[122] Fei Miao, Shuo Han, Shan Lin, Qian Wang, John A Stankovic, Abdeltawab Hendawi, Desheng Zhang, Tian He, and George J Pappas. Data-driven robust taxi dispatch under demand uncertainties. *IEEE Transactions on Control Systems Technology*, 27(1):175–191, 2017.

[123] Tripp Mickle, Yiwen Lu, and Mike Isaac. 'this experience may feel futuristic': Three rides in waymo robot taxis, 2023.

[124] Baichuan Mo, Zhejing Cao, Hongmou Zhang, Yu Shen, and Jinhua Zhao. Competition between shared autonomous vehicles and public transit: A case study in Singapore. *Transportation Research Part C: Emerging Technologies*, 127(April):103058, 2021.

[125] Baichuan Mo, Haris N. Koutsopoulos, Max Zuo-Jun Shen, and Jinhua Zhao. Robust Path Recommendations During Public Transit Disruptions Under Demand Uncertainty. 2022.

[126] Baichuan Mo, Zhenliang Ma, Haris N Koutsopoulos, and Jinhua Zhao. Capacity-constrained network performance model for urban rail systems. *Transportation Research Record*, 2674(5):59–69, 2020.

[127] Abood Mourad, Jakob Puchinger, and Chengbin Chu. A survey of models and algorithms for optimizing shared mobility. *Transportation Research Part B: Methodological*, 123:323–346, 2019.

[128] Vedant Nanda, Pan Xu, Karthik Abinav Sankararaman, John P. Dickerson, and Aravind Srinivasan. Balancing the tradeoff between profit and fairness in rideshare platforms during high-demand hours. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 131, New York, NY, USA, 2020. Association for Computing Machinery.

[129] G. F. Newell. Dispatching Policies for a Transportation Route. *Transportation Science*, 5(1):91–105, 1971.

[130] NYC Taxi and Limousine Commission. TLC Trip Record Data, 2019. "[Online; accessed 15-June-2020]".

[131] Marco Pavone, Stephen L Smith, Emilio Frazzoli, and Daniela Rus. Robotic load balancing for mobility-on-demand systems. *The International Journal of Robotics Research*, 31(7):839–854, Jun 2012.

[132] Dana Pessach and Erez Shmueli. A Review on Fairness in Machine Learning. *ACM Computing Surveys*, 55(3):51:1–51:44, February 2022.

[133] Helen K.R.F. Pinto, Michael F. Hyland, Hani S. Mahmassani, and I. Ömer Verbas. Joint design of multimodal transit networks and shared autonomous mobility fleets. *Transportation Research Part C: Emerging Technologies*, 113:2–20, 2020.

[134] Shiyou Qian, Jian Cao, Frédéric Le Mouël, Issam Sahel, and Minglu Li. SCRAM: A Sharing Considered Route Assignment Mechanism for Fair Taxi Route Recommendations. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 955–964, New York, NY, USA, August 2015. Association for Computing Machinery.

[135] Naveen Raman, Sanket Shah, and John Dickerson. Data-driven methods for balancing fairness and efficiency in ride-pooling. 10 2021.

[136] Mohsen Ramezani and Mehdi Nourinejad. Dynamic modeling and control of taxi services in large-scale urban networks: A macroscopic approach. *Transportation Research Part C: Emerging Technologies*, 94:203–219, 2018.

[137] Hannah Ritchie and Max Roser. Urbanization. *Our World in Data*, 2018. https://ourworldindata.org/urbanization.

[138] Mauro Salazar, Nicolas Lanzetti, Federico Rossi, Maximilian Schiffer, and Marco Pavone. Intermodal autonomous mobility-on-demand. *IEEE Transactions on Intelligent Transportation Systems*, 21(9):3946–3960, 2020.

[139] Bruce Schaller. Unsustainable? The Growth of App-Based Ride Services and Traffic, Travel and the Future of New York City, 2017.

[140] Susan Shaheen, Adam Cohen, Balaji Yelchuru, and Sara Sarkhili. Mobility on demand operational concept report. (FHWA-JPO-18-611), Sep 2017.

[141] Yu Shen, Hongmou Zhang, and Jinhua Zhao. Integrating shared autonomous vehicle in public transportation system: A supply-side simulation of the first-mile service in singapore. *Transportation Research Part A: Policy and Practice*, 113:125–136, Jul 2018.

[142] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 802–810, Cambridge, MA, USA, 2015. MIT Press.

[143] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248, London United Kingdom, July 2018. ACM.

[144] Kevin Spieser, Samitha Samaranayake, Wolfgang Gruel, and Emilio Frazzoli. Shared-vehicle mobility-on-demand systems: a fleet operator's guide to rebalancing empty vehicles. In *Transportation Research Board 95th Annual Meeting*, number 16-5987. Transportation Research Board, 2016.

[145] Konrad Steiner and Stefan Irnich. Strategic planning for integrated mobility-on-demand and urban public bus networks. *Transportation Science*, 54(6):1616–1639, 2020.

[146] Jiahui Sun, Haiming Jin, Zhaoxing Yang, Lu Su, and Xinbing Wang. Optimizing long-term efficiency and fairness in ride-hailing via joint order dispatching and driver repositioning. pages 3950–3960. Association for Computing Machinery, 8 2022.

[147] Arslan Ali Syed, Irina Gaponova, and Klaus Bogenberger. Neural network-based metaheuristic parameterization with application to the vehicle matching problem in ride-hailing services. *Transportation Research Record*, 2673(10):311–320, 2019.

[148] Gabriel E. Sánchez-Martínez. Inference of public transportation trip destinations by using fare transaction and vehicle location data: Dynamic programming approach. *Transportation Research Record*, 2652(1):1–7, 2017.

[149] Tom Sühr, Asia J. Biega, Meike Zehlike, Krishna P. Gummadi, and Abhijnan Chakraborty. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. pages 3082–3092. Association for Computing Machinery, 7 2019.

[150] Amirmahdi Tafreshian, Neda Masoud, and Yafeng Yin. Frontiers in service science: Ride matching for peer-to-peer ride sharing: A review and future directions. *Service Science*, 12(2-3):44–60, 2020.

[151] Terry A. Taylor. On-demand service platforms. *Manufacturing & Service Operations Management*, 20(4):704–720, Jul 2018.

[152] The White House. President biden issues executive order on safe, secure, and trustworthy artificial intelligence, 2023.

[153] Transport for London. Traffic modelling guidelines (version 4.0), 2021.

[154] Transportation Research Board. *Transit Capacity and Quality of Service Manual, Third Edition*. 2013.

[155] Matthew Tsao, Dejan Milojevic, Claudio Ruch, Mauro Salazar, Emilio Frazzoli, and Marco Pavone. Model predictive control of ride-sharing autonomous mobility-on-demand systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6665–6671, 2019.

[156] Uber Technologies, Inc. Uber Movement, New York City travel speeds, 2019. "[Online; accessed 15-June-2020]".

[157] United States Environmental Protection Agency. Sources of greenhouse gas emissions, 2022. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions.

[158] U.S. Department of Transportation. Public transportation's role in responding to climate change, 2010.

[159] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[160] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. (arXiv:1710.10903), Feb 2018. arXiv:1710.10903 [cs, stat].

[161] I. Verbas and Hani Mahmassani. Optimal allocation of service frequencies over transit network routes and time periods. *Transportation Research Record*, (2334):50–59, 2013.

[162] Ömer Verbas, Charlotte Frei, Hani S. Mahmassani, and Raymond Chan. Stretching resources: sensitivity of optimal bus frequency allocation to stop-level demand elasticities. *Public Transport*, 7(1):1–20, 2015.

[163] Eleni I. Vlahogianni, Matthew G. Karlaftis, and John C. Golias. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43:3–19, June 2014.

[164] Alex Wallar, Menno Van Der Zee, Javier Alonso-Mora, and Daniela Rus. Vehicle rebalancing for mobility-on-demand systems with ride-sharing. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4539–4546. IEEE, 2018.

[165] Hai Wang and Hai Yang. Ridesourcing systems: A framework and review. *Transportation Research Part B: Methodological*, 129:122–155, Nov 2019.

[166] Qingyi Wang, Shenhao Wang, Dingyi Zhuang, Haris Koutsopoulos, and Jinhua Zhao. Uncertainty quantification of spatiotemporal travel demand with probabilistic graph neural networks. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–12, 2024.

[167] Xiaohan Wang, Zhan Zhao, Hongmou Zhang, Xiaotong Guo, and Jinhua Zhao. Quantifying the uneven efficiency benefits of ridesharing market integration, 2023.

[168] Xudong Wang, Yuankai Wu, Dingyi Zhuang, and Lijun Sun. Low-rank hankel tensor completion for traffic speed estimation. *IEEE Transactions on Intelligent Transportation Systems*, 24(5):4862–4871, 2023.

[169] Yineng Wang, Xi Lin, Fang He, and Meng Li. Designing transit-oriented multi-modal transportation systems considering travelers' choices. *Transportation Research Part B: Methodological*, 162(October 2021):292–327, 2022.

[170] Yu Wang, Yu Zhang, and Jiafu Tang. A distributionally robust optimization approach for surgery block allocation. *European Journal of Operational Research*, 273(2):740–753, 2019.

[171] Jian Wen, Yu Xin Chen, Neema Nassir, and Jinhua Zhao. Transit-oriented autonomous vehicle operation with integrated demand-supply interaction. *Transportation Research Part C: Emerging Technologies*, 97:216–234, Dec 2018.

[172] Jian Wen, Jinhua Zhao, and Patrick Jaillet. Rebalancing shared mobility-on-demand systems: A reinforcement learning approach. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 220–225, 2017.

[173] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling, May 2019.

[174] Peng Xiong, Panida Jirutitijaroen, and Chanan Singh. A distributionally robust optimization model for unit commitment considering uncertain wind power generation. *IEEE Transactions on Power Systems*, 32(1):39–49, 2016.

[175] Haitao Xu, Jing Ying, Hao Wu, and Fei Lin. Public bicycle traffic flow prediction based on a hybrid model. *Applied Mathematics & Information Sciences*, 7(2):667–674, 2013.

[176] Zhe Xu, Zhixin Li, Qingwen Guan, Dingshui Zhang, Qiang Li, Junxiao Nan, Chunyang Liu, Wei Bian, and Jieping Ye. Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 905–913, 2018.

[177] An Yan and Bill Howe. Fairness-aware demand prediction for new mobility. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1079–1087, Apr. 2020.

[178] An Yan and Bill Howe. Equitensors: Learning fair integrations of heterogeneous urban data. pages 2338–2347. Association for Computing Machinery, 2021.

[179] Chiwei Yan, Helin Zhu, Nikita Korolko, and Dawn Woodard. Dynamic pricing and matching in ride-hailing platforms. *Naval Research Logistics (NRL)*, 67(8):705–724, 2020.

[180] Yadan Yan, Zhiyuan Liu, Qiang Meng, and Yu Jiang. Robust optimization model of bus transit network design with stochastic travel time. *Journal of Transportation Engineering*, 139(6):625–634, 2013.

[181] Hai Yang, Xiaoran Qin, Jintao Ke, and Jieping Ye. Optimizing matching time interval and matching radius in on-demand ride-sourcing markets. *Transportation Research Part B: Methodological*, 131:84–105, 2020.

[182] Hai Yang, Chaoyi Shao, Hai Wang, and Jieping Ye. Integrated reward scheme and surge pricing in a ridesourcing market. *Transportation Research Part B: Methodological*, 134:126–142, 2020.

[183] Huaxiu Yao, Xianfeng Tang, Hua Wei, Guanjie Zheng, and Zhenhui Li. Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(0101):5668–5675, Jul 2019.

[184] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Zhenhui Li, Jieping Ye, and Didi Chuxing. Deep multi-view spatial-temporal network for taxi demand prediction. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 2588–2595, 2018.

[185] Jiexia Ye, Juanjuan Zhao, Kejiang Ye, and Chengzhong Xu. How to Build a Graph-Based Deep Learning Architecture in Traffic Domain: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[186] Mischa Young and Steven Farber. The who, why, and when of Uber and other ride-hailing trips: An examination of a large sample household travel survey. *Transportation Research Part A: Policy and Practice*, 119(November 2018):383–392, 2019.

[187] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3634–3640, July 2018. arXiv:1709.04875 [cs, stat].

[188] Gioele Zardini, Nicolas Lanzetti, Marco Pavone, and Emilio Frazzoli. Analysis and control of autonomous mobility-on-demand systems: A review. *arXiv:2106.14827 [cs, eess]*, Jun 2021. arXiv: 2106.14827.

[189] Liteng Zha, Yafeng Yin, and Hai Yang. Economic analysis of ride-sourcing markets. *Transportation Research Part C: Emerging Technologies*, 71:249–266, 2016.

[190] Hongmou Zhang, Xiaotong Guo, and Jinhua Zhao. Economies and diseconomies of scale in segmented mobility sharing markets, 2022.

[191] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(11), Feb 2017.

[192] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. Dnn-based prediction model for spatio-temporal data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 1–4, Burlingame California, Oct 2016. ACM.

[193] Ning Zhang, Yunlong Zhang, and Haiting Lu. Seasonal autoregressive integrated moving average and support vector machine models: prediction of short-term traffic flow on freeways. *Transportation Research Record*, 2215(1):85–92, 2011.

[194] Rick Zhang and Marco Pavone. Control of robotic mobility-on-demand systems: a queueing-theoretical perspective. *arXiv:1404.4391 [cs]*, Apr 2014. arXiv: 1404.4391.

[195] Rick Zhang, Federico Rossi, and Marco Pavone. Model predictive control of autonomous mobility-on-demand systems. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1382–1389. IEEE, 2016.

[196] Tianqi Zhang and Ge Guo. Graph attention lstm: A spatiotemporal approach for traffic flow forecasting. *IEEE Intelligent Transportation Systems Magazine*, 14:190–196, 2022.

[197] Xiaojian Zhang, Qian Ke, and Xilei Zhao. Enhancing Fairness in AI-based Travel Demand Forecasting Models, March 2023. arXiv:2303.01692 [cs].

[198] Jinhua Zhao, Adam Rahbee, and Nigel H. M. Wilson. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, 22(5):376–387, 2007.

[199] Yunhan Zheng, Qingyi Wang, Dingyi Zhuang, Shenhao Wang, and Jinhua Zhao. Fairness-enhancing deep learning for ride-hailing demand prediction, March 2023. arXiv:2303.05698 [cs].

[200] Yunhan Zheng, Shenhao Wang, and Jinhua Zhao. Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models. *Transportation Research Part C: Emerging Technologies*, 132:103410, November 2021.