Toward a Resilient Public Transportation System: Effective Monitoring and Control under Service Disruptions

by

Baichuan Mo
B.E., Tsinghua University (2018)
M.S., Massachusetts Institute of Technology (2020)

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
September 2022

Author . . . .
Department of Civil and Environmental Engineering
August 12, 2022

Certified by. . . . .
Jinhua Zhao
Edward H. and Joyce Linde Associate Professor of Transportation and City Planning, MIT
Thesis Supervisor

Certified by. . . . .
Haris N. Koutsopoulos
Professor of Civil and Environmental Engineering, Northeastern University
Thesis Supervisor

Accepted by . . . .
Colette L. Heald
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Toward a Resilient Public Transportation System: Effective Monitoring and Control under Service Disruptions

by

Baichuan Mo

Submitted to the Department of Civil and Environmental Engineering
on August 12, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Transportation

Abstract

Urban public transit is an important component of transportation systems and plays a critical role in providing mobility in many metropolitan areas. However, with aging systems, continuous expansion, and near-capacity operations, transit systems are susceptible to unplanned delays and service disruptions caused by equipment, weather, passengers, or other internal and external factors, resulting in great inconvenience for passengers and economic loss for operators. Ensuring good service provision during service disruptions is important for public transit management.

Resilience is an important concept related to incidents. It usually refers to the ability of an entity to return to its initial conditions after it is disturbed. Since monitoring, control, and planning are the three major tasks for public transit system management, we define the resilience of a public transit system as the ability to monitor, control, and plan for incidents (service disruptions) in ways to mitigate congestion, improve travel efficiency, and reduce safety risks.

This dissertation focuses on the first two tasks to improve the resilience of public transit operation in light of disruptions that regularly take places. Specifically, we aim to 1) understand the impact of unplanned incidents on PT systems (i.e., monitoring); and 2) design mitigating strategies to relieve incident impacts (i.e., control). The specific topics we cover in the thesis can be categorized by a two-by-two matrix. The first dimension considers short-term (e.g., less than a couple of minutes) v.s. long-term (e.g., more than 1 hour) incidents while the second dimension considers monitoring and control tasks. Five different studies under the umbrella of this two-by-two matrix are presented.

The first study evaluates a transit system performance under random short-term service suspensions using a bulk-service queue model. We prove that under random suspensions, headways can be represented as the difference between two compound Poisson exponential variables. Assuming no vehicle overtaking, we approximate the headway as a zero-inflated truncated normal distribution to obtain a closed-form moment generating function (MGF). Based on the MGF, we derive the system stability conditions and the mean and variance of queue length and waiting time at each station with analytical formulations. The second study provides an empirical analysis of the impact of service disruptions. We use a real-world train collision incident at the Chicago Transit Authority (CTA) system to analyze the impact of unplanned long-term incidents on the system's demand, supply, and passenger behavior. We also propose a redundancy index to quickly identify alternative capacity in CTA under service disruptions. The third study proposes a probabilistic method to infer passengers' behavior (e.g., waiting, switching to another line, transferring to a bus) under disruptions. The main contribution is a probabilistic model to recognize whether an observed smart card record (e.g., transfer to a bus stop) is a normal behavior or due to the incident. This model allows us to extract the actual behavioral responses and outperforms the typical rule-based methods. The fourth study proposes a station-based path recommendation model to reduce the total system travel time during disruptions. We use a robust optimization-based formulation to address the demand uncertainty. The closed-form robust counterpart is derived. To tackle the lack of an analytical formulation of travel times due to left behind, we propose a simulation-based first-order approximation to transform the original problem into a linear program and solve it iteratively with the method of successive average. The fifth study proposes an individual-based path recommendation model with the objective of minimizing total system travel time and respecting passengers' path choice preferences. Passengers' behavior uncertainty in path choices given recommendations and travel time equity are also considered in the formulation. We model the behavior uncertainty based on passenger's prior preferences and the posterior path choice probability distribution with two new concepts: ϵ-feasibility and Γ-concentration, which control the mean and variance of path flows in the optimization problem. We show that these two concepts can be transformed into linear constraints using Chebyshev's inequality. The individual path recommendation problem with behavior uncertainty is efficiently solved using Benders decomposition. Finally, we use a post-adjustment heuristic to address equity requirements.

Future research directions and potential applications of the work are discussed in the last chapter.

Thesis Supervisor: Jinhua Zhao
Title: Edward H. and Joyce Linde Associate Professor of Transportation and City Planning, MIT

Thesis Supervisor: Haris N. Koutsopoulos
Title: Professor of Civil and Environmental Engineering, Northeastern University

Acknowledgments

Throughout my four years at MIT, I am indebted to many people for their companionship, support, encouragement, and help.

First, I would like to express my deepest gratitude to my advisors Prof. Jinhua Zhao and Prof. Haris N. Koutsopoulos. Jinhua and Haris are two different spirits that consistently inspire my life. Jinhua is a creator, showing me rich and exciting research ideas. He is the compass in my research ship, guiding me towards a vast and attractive ocean. I pretty much enjoyed the extensive discussions and sparkling moments with him. Haris is a craftsman, digging into the depth of techniques with a thorough understanding. He is the sail in my research ship, providing the essential energy support in a challenging journey. I pretty much appreciated his dedicated revision to every corner of my papers, rigorous attitude towards every problem encountered, and enlightening suggestions in even the complicated methodology details. From them, I learned what a great mentor should be. I feel very honored to have both of them as my advisors during my studies at MIT. Jinhua and Haris are more than my advisors. In many moments of my life, I feel they are also close friends. I am super glad to share with them my newest achievements, my recent experiences, and any other interesting things in my life. They are always supportive of every decision I have made. They care about my daily life as much, if not more than my academic development. Because of them, my PhD journey goes so smoothly, enjoyably, and successfully.

I am also grateful for my other two committee members, Prof. Max Zuo-Jun Shen and Prof. Cathy Wu. Max is an expert in operations research and management science. Many of his suggestions still greatly impact my future research directions and tastes. Cathy is a young genius scholar. Many of her comments are super constructive to my presentation and dissertation. Thanks for their tremendous support throughout my several committee meetings. I also thank other faculty members in MIT Transit Lab, Jim Aloisi, John Attanucci, Fred Salvucci, and Prof. Nigel Wilson.

I am thankful to my friends at JTL-Transit Urban Mobility Lab. I would like to especially thank Qing Yi Wang, Yunhan Zheng, and Xiaotong Guo, with whom I

discussed a lot of research ideas and had a lot of fun. I also thank Zhenliang Ma, Shen Yu, Shenhao Wang, Hongmou Zhang, Zhan Zhao, Hui Kong, Jintai Li, Dingyi Zhuang, Yuzhu Huang, Nick Caros, Dajiang Suo, Rachel Luo, Ruben Morgan, Jake Gao, Patrick Meredith-Karam, John Moody, Joanna Moody, Anson Stewart, Mei-Chun Ruth Tse Yiu, Michael Martello, Ehab Ebeid, Andreas Haupt, Peyman Noursalehi, and Nate Bailey for cooperation, help, encouragement, and friendship.

I would also like to thank my friends at MIT CEE, MIT CSSA, MIT CEO, and Northeastern University. Thanks to Yu Qiu, Yifei Xie, Siyu Chen, Jie Deng, Yunpo Li, Tong Bo, Yue Meng, Ruijiao Sun, Zhaozhong Zhuang, Tian Zhao, Jie Yun, Ping Xu, Tiancheng Yu, Zhichu Ren, Zhuyu Peng, Xinyi Gu, Xiaonuo Yang, Xinhua Wu, Kerem Tuncel, Jiali Zhou, and Joseph Rodriguez.

I am grateful to the Hong Kong Mass Transit Rail program, the Chicago Transit Authority program, the Singapore–MIT Alliance for Research and Technology Future Mobility program, and MIT UPS Fellowship for providing financial and data support during my staying at MIT. I also thank MIT Center for Transportation and Logistics (CTL). CTL is my second lab, I will never forget having spent many days and nights there working for my research.

I am extremely grateful to my mother, Qingping Tang. Thank you for respecting my choices and providing me with opportunities to pursue my dreams. Your support is one of my strongest spiritual dependents during many of my dark times. I would like to offer special thanks to my father, Guoshan Mo, who, although no longer with me when I was fourteen. I am sorry you are not able to witness my growth. I am confident that what I have achieved today is much more than your expectation. You should be proud of your son in heaven.

Finally, a special thank to ZMJ, who just entered my life this summer, greatly lighting up my mood. You might not even know how special you are to me. I am looking forward to our upcoming story.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and motivation

In this section, we briefly discuss the background of service disruptions in public transportation systems and the concept of resilience. We also introduce our definition of a resilient public transit system.

1.1.1 Public transit systems and disruptions

Public transit plays an important role in the urban mobility system. Millions of passengers use urban transit systems for daily commuting and accessing various activities. However, with aging systems, continuous expansion, and near-capacity operations, transit systems are susceptible to unplanned delays and service disruptions caused by equipment failure, weather, passengers, and other internal and external factors. Take the Chicago Transit Authority (CTA) system as an example (see Table 1.1). In 2019, 27,198 unplanned incidents are reported in the rail system alone, which implies 75 incidents per day on average. The average number of major incidents per day (duration $\geq$ 20 minutes) is 1.04.

Table 1.1: CTA incident statistics[1] (2019)

Total # of incidents	Avg # incidents per day	Avg # major[2] incidents per day
27,198	75	1.04

[1]: Data calculated by the author.

[2]: Major incidents are those with a duration greater than 20 minutes

These statistics show that service disruptions and incidents are not unusual in public transit systems. They cause great inconvenience to both customers and operators. Figure 1-1a shows a photo of the train crash incident at the Chicago O'Hare station on March 24, while Figure 1-1b shows a typical crowding situation at platforms during service disruptions. Incidents result in passenger delays, cancellation of trips, economic losses, and safety concerns (e.g., due to crowding at platforms) [6]. It is important to understand the impacts of incidents and propose mitigating strategies to relieve them.

(a) Train crash

(b) Crowding at platforms

Figure 1-1: Example of service disruptions and the consequences

1.1.2 Resilience

Resilience is an important concept related to incidents. Deriving from the Latin verb "resiliere" (literally means "to bounce back"), resilience usually refers to the ability of an entity to return to its initial conditions after it is disturbed [7]. As a concept, resilience has gained increasing attention in various fields, including ecology, engineering, social sciences, and climate science. The proliferation of the concept in different domains makes it polysemic, resulting in a diversifying set of definitions, each of which may be sensible in its own context. Therefore, to properly discuss resilience within an academic setting, it must be prefaced by the context within which it is to be invoked. Keenan [1] illustrates the concept of resilience across a range of academic disciplines (Figure 1-2). These definitions can be classified on a spectrum of varying degrees of equilibrium states and normative characteristics.

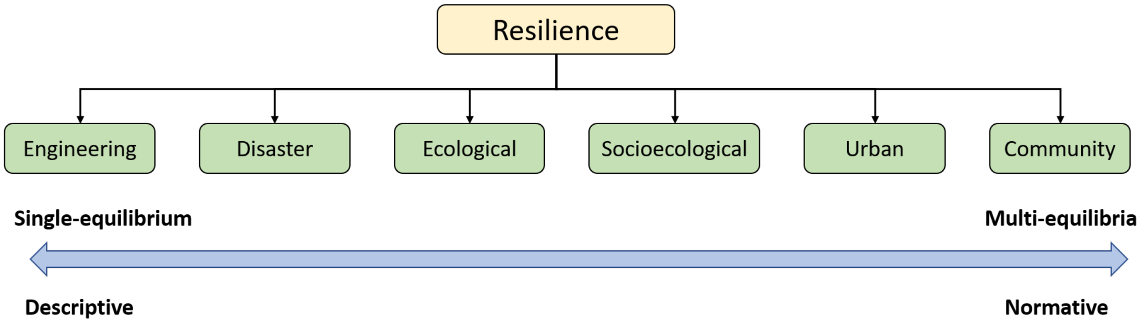


Figure 1-2: Topology of resilience definitions across fields of study (adapted from Keenan [1] and Martello [2])

In the context of engineered (or closed) systems, definitions tend to assume a single-equilibrium system with different components arranged to achieve a predefined system state (i.e., with intentionality [8]). Within such systems, definitions of resilience are based on predefined value judgment according to the desired system state. Therefore such systems are of a descriptive nature.

In contrast, systems that are more open and indeterminate, such as socio-ecological systems, may achieve multiple equilibria. Therefore, the definition of resilience needs to be interpreted or specified with desirable actions or system states, implying resilience is of a normative nature [9]. That is, within such systems, the description of system processes and responses to perturbation are ultimately dependent on the predefined cultural, organizational, or ideological norms [2].

Public transit systems are engineered systems. Resilience for an engineered system is defined as:

"the endogenous capacity of a system to cope with a predefined exogenous perturbation, responding or reorganizing in ways that maintain its perceived essential function, identity, and structure, while also maintaining the capacity for adaptation and transformation" [10].

The engineering resilience in transport areas has been studied considerably in recent years. Examples can be found in air [11], road [12], supply chain [13], waterborne [14] and railway networks [15].

In an engineered single-equilibrium system, we may evaluate the system with a predefined system function. The following building characteristics represent distinct system states: vulnerability, survivability, response, and recovery (Figure 1-3).

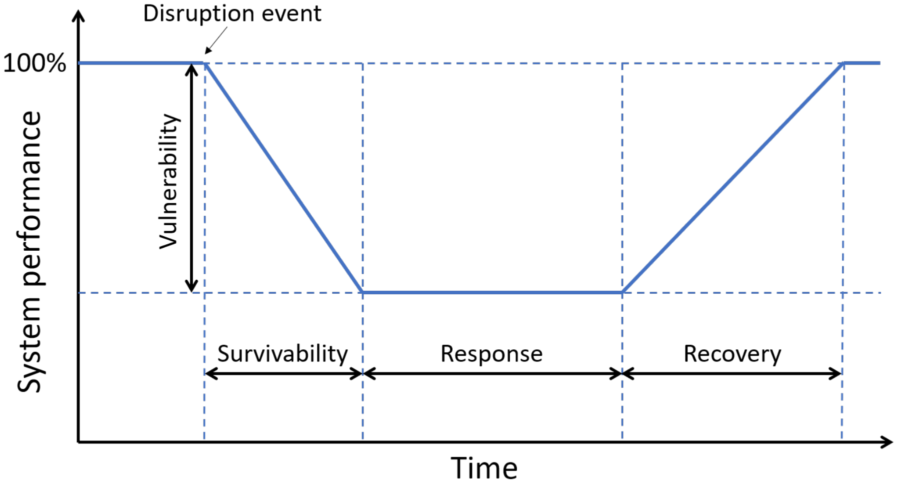


Figure 1-3: Different system states for the definition of the resilience of an engineered system (adapted from Bešinović [3])

Vulnerability is defined as how much performance remains after a disruption event [15]. A similar definition can be found in Zhou et al. [16]. Other terms related to vulnerability are resistance, flexibility, and redundancy. Robustness can be considered as a counterpart of vulnerability. **Survivability** is the functionality of the system that remains when it transitions from the normal system performance (i.e. 100%) to a disrupted one. In practice, when a disruption happens, the system can degrade differently (e.g. fail completely at once or have its performance deteriorate slowly until finally reaches the disrupted steady-state). **Response** is the set of actions taken directly and immediately after a disruption in order to provide the best level of service possible during a disruption. This phase represents a disrupted steady-state of the system. Depending on the nature of the disruption, the duration of the phase varies. **Recovery** is the ability of the system to return from the disrupted state to its original condition. The nature of the response may affect the recovery process (e.g., better response actions can shorten the recovery process). In certain types of disruptions, some of these states may not exist. Also, in some cases, survivability is considered as part of the response, while in others, the response may be part of the recovery phase.

1.1.3 Public transit system management

To define the resilience of a public transit system, it is essential to first understand the main functions of a public transit system and how it is managed. Figure 1-4 shows the three major tasks in public transit system management, including (historical) performance monitoring, (real-time) control, and (future) operation planning. Each task contains many associated sub-problems, which cover different dimensions of service management.

Monitoring means estimating the system performance (e.g., travel delays, train load) and passenger behavior (e.g., path choices) to understand the situation of the system, which is usually conducted for history scenarios. Control aims to adjust the operations (e.g., rerouting) in real-time to provide better services, especially under incident conditions. Planning tries to design and prepare the system to serve future demands. The sub-tasks include designs of networks, fare policies, schedules, etc. These tasks can allow operators to understand, inform, and improve transit services.

1.1.4 Resilience for public transit systems

Given the major tasks of a public transit system, we define its resilience as *the ability to cope with incidents (or service disruptions) through monitoring, control, and planning to maintain connectivity, mitigate congestion, and reduce travel delays and safety*

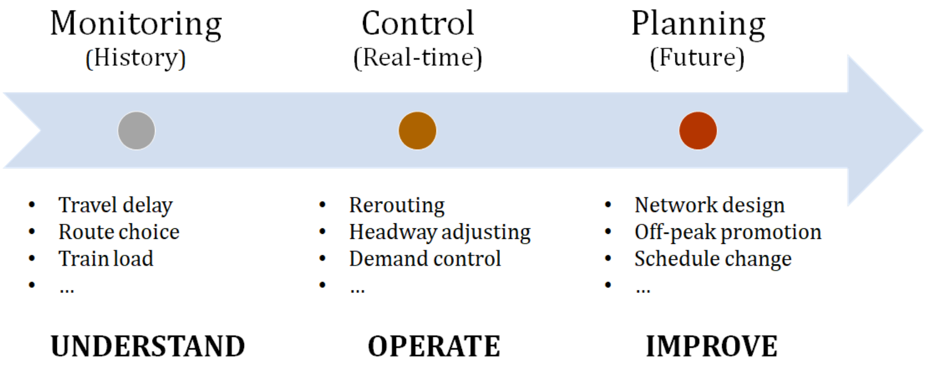


Figure 1-4: Three major tasks for public transit system management

risks.

Our definition of resilience includes three aspects: retrospective, reactive, and proactive. The first one implies the review of previous disruptions, understanding the impact of incidents and the performance of response strategies. The second aspect protects against possible disruptions with real-time control strategies. The last aspect indicates the preparation and strategic planning for future disruptions.

The recent review of resilience definitions in Zhou et al. [17] indicates that there is no unique choice on how to define resilience. However, certain similarities can be observed across these resilience definitions. We summarize the comparison in Table 1.2. The definition we use in the thesis belongs to the first category.

Incidents can happen at different aspects of a public transit system [5]. From the infrastructure aspect, disruptions may be caused by technical failures (e.g., bridge collapses, power outages, vehicle malfunctions, etc.), extreme operating conditions (floods, snowstorms, etc.), or deliberate attacks. From the service aspect, disruptions can occur from events such as human errors or crew shortages due to sickness or labor disputes. From the operations aspect, disruptions may arise from policy shifts or budget cuts. From the demand aspect, incidents (e.g., demand surge) may occur due to special events (e.g., concerts, Olympic games) or various forms of societal upheaval and crises. To improve the system's resilience, strategies can be designed from all these different aspects.

Given the different layers of a public transit system, its system performance func-

Table 1.2: Definitions of resilience in transport systems (adapted from Bešinović [3])

Definitions	**Reference**
Ability to recover quickly from a disruption	Bababeik et al. [18], Chan and Schofer [19], Lu [20], Adjetey-Bahun et al. [21], **This dissertation**.
Remaining system's performance during a disruption	Khaled et al. [15], Diab and Shalaby [22], Ferranti et al. [23], Dawson et al. [24], Dorbritz [25]
Described with four properties: robustness, redundancy, resourcefulness and rapidity	Bruneau et al. [26], Beiler et al. [27], Bocchini et al. [28]
A function of the system's vulnerability against potential disruption	Mansouri et al. [14], Saadat et al. [29], Zhang et al. [30]

tion also has different implications. Figure 1-5 shows two examples of system functions from the supply and demand sides. From the supply side, we can define the system function for a transit system as the service frequency. Disruptions may directly decrease the service frequency due to suspensions of services. The recovery of service frequency is also an indicator of the end of the incident from the operators' point of view. From the demand side, we may define the system function as the service quality index defined based on the passenger's waiting time (e.g., $\frac{\text{No-incident waiting time}}{\text{Incident waiting time}}$). Note that a higher waiting time indicates a lower system performance. The evolution processes of supply and demand system functions are asynchronous. For example, though the system may recover to the normal service frequency, the dissipation of congestion may take additional time due to accumulated populations during the incident period.

1.2 Research questions

This dissertation focuses on the first two tasks in Figure 1-4 (monitoring and control) to develop a resilient public transit system. Specifically, we aim to

- **understand the impact of unplanned incidents on PT systems (i.e.,**

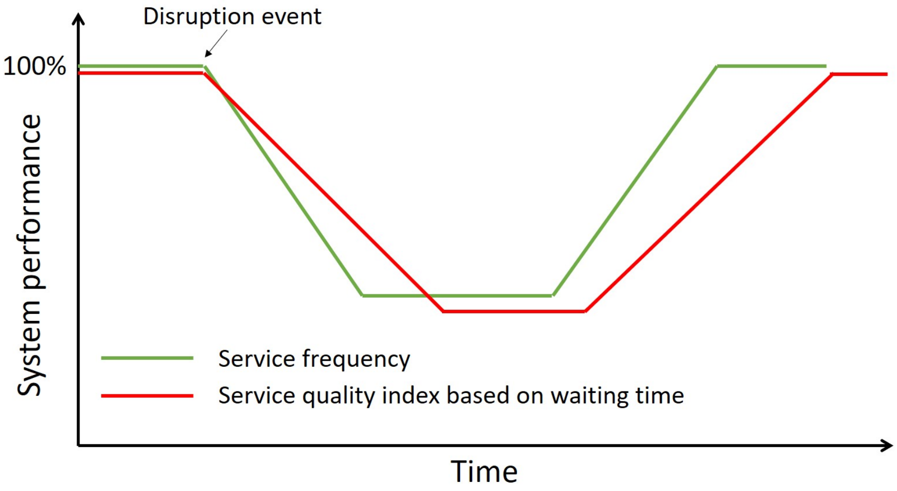


Figure 1-5: Different system performance function definitions for public transit systems

monitoring) and

- **design mitigating strategies to relieve incident impacts (i.e., control)**.

There are a lot of research questions that can be studied for monitoring and control tasks during service disruptions.

For the monitoring task, we can analyze incidents' impact on public transit systems from different angles (demand, supply, level of service), incident types (planned vs. unplanned, long-term vs. short-term vs. special events), and research approaches (data-based empirical studies, theoretical studies such as queuing models, dynamical systems, graph theories, and simulation-based studies). We can also analyze incidents' impact on passenger behavior. The analysis can be categorized based on the nature of the behavior change (instantaneous, temporary, permanent), analysis level (aggregate-level, individual-level), and used data sources (survey-based, empirical data-based).

For the control tasks, we can design models and strategies from the supply and demand sides. On the demand side, we can design station-based inflow control strategies during disruptions to reduce system crowding. We can also design information for passengers in the form of path recommendations to guide their route choices and reduce congestion. On the supply side, we can adjust operations for existing services (e.g., re-scheduling, rolling stock adjustment, re-routing), design shuttle bus services (e.g., routes and schedules), and develop multimodal integration models such as integrating with Transportation Network Companies (TNC), bike/scooter sharing systems, local bus companies, etc.

1.3 Research objectives and conceptual framework

Though there are many research questions under the public transit resilience umbrella, the dissertation only addresses a subset of them. Figure 1-6 presents the conceptual framework of the dissertation. It can be characterized by a two-by-two matrix. The first dimension is incident type. We consider short-term service suspensions (or perturbations) and long-term service disruptions (both are unplanned). The second dimension is monitoring and control tasks.

Short-term service suspensions usually cause some delays (e.g., less than 5 minutes). The whole system is still working and no repairs are needed. These suspensions may be caused by congestion or improper operating behavior. Most of the passengers would choose to wait in this scenario. In this dissertation, the monitoring task for short-term service suspensions is to evaluate the change in passengers' queue length and waiting time at a station. A bulk-service queue model is used to derive some theoretical results. And the control task is to derive optimal control strategies to reduce passengers' waiting time based on the results in the monitoring part.

Long-term service disruptions usually cause the shutdown of the service with a duration greater than 30 minutes. The impact of long-term service disruptions is substantial. PT operators need to repair the system and passengers usually need to change their travel modes. In the thesis, the monitoring task for long-term service suspensions is to measure the impact of service disruptions on the system's demand and supply empirically, and infer passenger's travel mode choices during disruptions. And the control task is to design route recommendation strategies for passengers and propose control strategies for operators (such as shuttle buses) so as to mitigate the system impact and reduce passengers' travel times.

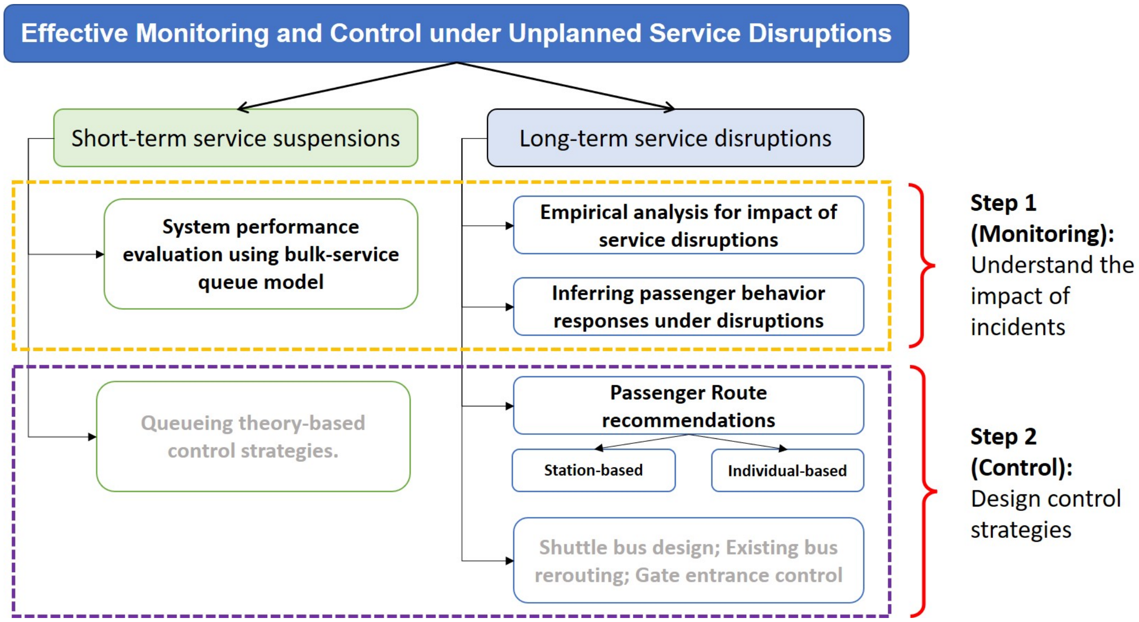

Effective Monitoring and Control under Unplanned Service Disruptions
Short-term service suspensions
Long-term service disruptions
System performance evaluation using bulk-service queue model
Empirical analysis for impact of service disruptions
Inferring passenger behavior responses under disruptions
Step 1 (Monitoring): Understand the impact of incidents
Queueing theory-based control strategies.
Passenger Route recommendations
Station-based
Individual-based
Shuttle bus design; Existing bus rerouting; Gate entrance control
Step 2 (Control): Design control strategies

Figure 1-6: Conceptual framework of the dissertation (Gray texts are topics not included in the dissertation)

Within this two-by-two matrix, the dissertation presents five different topics:

- System performance evaluation under short-term service suspensions using a bulk-service queue model (Chapter 2),
- Empirical analysis for the impact of service disruptions (Chapter 3),
- Inferring passenger behavioral responses under disruptions (Chapter 4),
- Station-based passenger path recommendations (Chapter 5), and
- Individual-based passenger path recommendations (Chapter 6).

1.4 Data and context

1.4.1 Automatically collected data

There are two important automatically-collected data in public transit systems: automated fare collection (AFC) data and automated vehicle location (AVL) data [31]. AFC data include passengers' usage transactions from smart card data. Based on the system, AFC data may contain transactions on both rail and buses (e.g., the CTA system in Chicago) or only rail (e.g., Hong Kong). AFC systems are either open or closed. Open systems require that passengers only tap in when they enter the system (e.g. the MBTA system in Boston). Closed systems require both, tapping in and tapping out (e.g. the transit system in Seoul, Korea). Many systems are hybrid, utilizing an open architecture on the bus side and closed on the subway side (e.g. London). For a closed system, AFC data can provide accurate origins and destinations of passengers' trips with tap-in and tap-out times. Given the rich information provided, AFC data have been used for understanding travel patterns [32], predicting individual trips [33], improving transit planning [34], etc. A complete review of the use of AFC data for transit system management can be found in Pelletier et al. [35].

AVL data contains information on the time-dependent location of vehicles. Train locations are collected from the rail tracking system. Bus locations are collected from vehicles' GPS units. AVL data have been used for measuring travel time variability and reliability [36], predicting vehicle arrival time [37], updating real-time scheduling [38]. A complete review of using AVL data for transit planning and management can be found in Levy and Lawrence [39].

1.4.2 Case studies and application

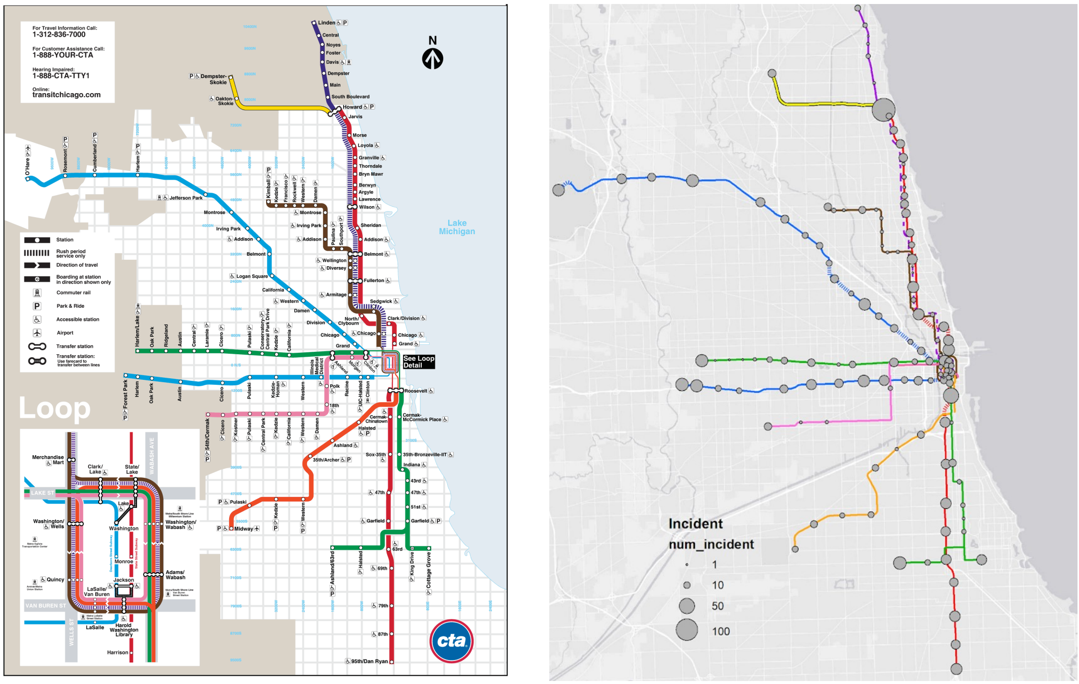
Data from the Chicago Transit Authority (CTA) are used in the various case studies. CTA is the second-largest transit system in the United States, providing services in Chicago, Illinois, and some of its surrounding suburbs. It operates 24 hours each day and is used by 0.84 million bus and 0.81 million train passengers per weekday on average in 2019 [40]. CTA has approximately 1,800 buses that operate over 140 routes traveling along 2,230 miles. Buses serve more than 12,000 posted bus stops. CTA's 1,450 train cars operate over eight routes and 222 miles of track, serving 145 stations in Chicago and seven suburbs.

The map of the CTA rail system is shown in Figure 1-7a. The rail system consists of eight lines (named after their color) and the "Loop". The Loop, located in the Chicago downtown area, is a 1.79 miles long circuit of elevated rail that forms the hub of the Chicago rail system. Its eight stations account for around 10% of the weekday boardings of the CTA trains.

CTA's AFC system is open. Passengers use their farecards only when entering a rail station or boarding a bus, so no information about a trip's destination is directly provided. The train tracking system provides train arrival and departure times at each station (i.e., AVL data).

1.4.3 CTA incidents

We obtain CTA's incident information from the control center data. The control center records every incident that occurred in the system with information on time, location, duration, and causes. Figure 1-7b shows the number of incidents distribution over different stations (only consider incidents with a duration of more than 10 minutes). In general, transfer stations (e.g., Howard and Roosevelt) and terminal stations (e.g., Forest park, O'Hare) have more incidents than others. This may be due to the fact that these stations have more complex infrastructure systems.



(a) CTA rail system

(b) Incidents location distribution

Figure 1-7: CTA incident overview

Figure 1-8 shows the number of incidents distributed over different lines. We find that the Red Line has the most incidents. Moreover, the proportion of major incidents (duration more than 20 minutes) is also the highest in the Red Line. The reason may be that the Red Line is the busiest line on the rail system (an average of 209,085 passengers boarding each weekday in 2019). It runs 24 hours a day, 365 days a year. The intense operations under high demands bring more infrastructure issues and management challenges (thus a high incident rate).

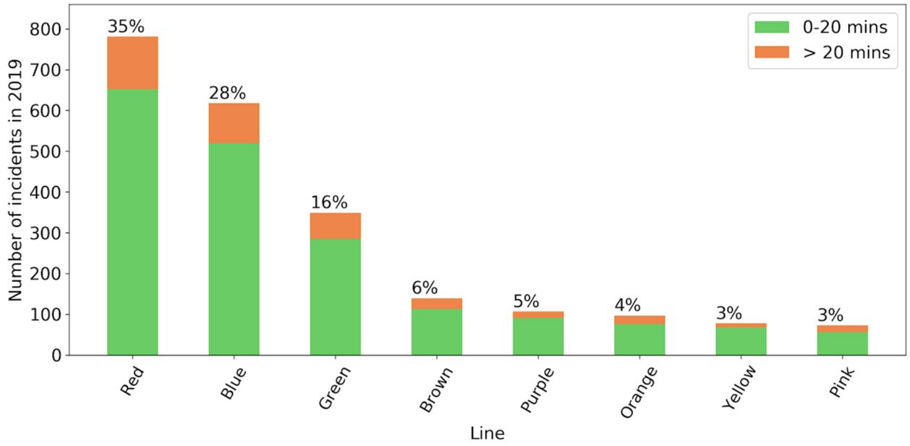


Figure 1-8: Incidents distribution over different lines (numbers in each bar show the proportion of incidents for this line)

Figure 1-9 shows the distribution of the number of incidents over months in 2019. January has the most incidents, which may be due to weather conditions.

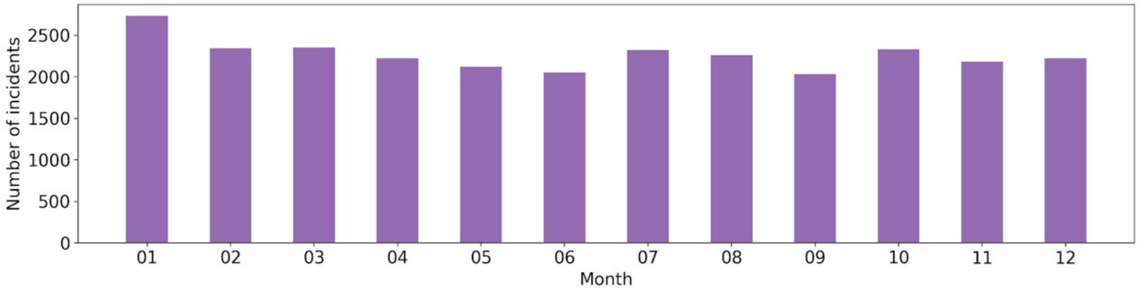


Figure 1-9: CTA number of incidents over months in 2019

Figure 1-10 shows the incident duration distributions. Around 80% of incidents in the CTA system are less than 5 minutes. The average incident duration is 4.97

minutes. This emphasizes the importance of service management under short-term service suspensions.

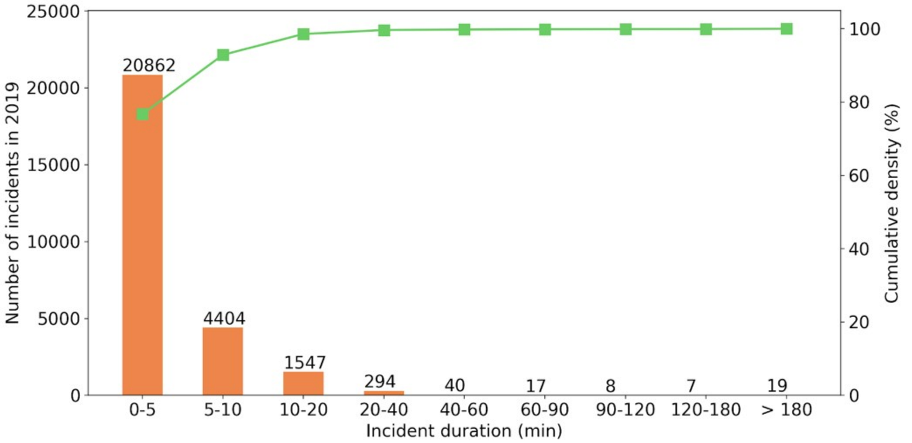


Figure 1-10: Incidents duration distribution

The incident causes distribution is shown in Figure 1-11. The fraction of incidents due to "crime" is the highest. Notably, though "power and track" only accounts for 4% of all incidents, it has a high probability (around 50%) of causing disruptions of more than 20 minutes.

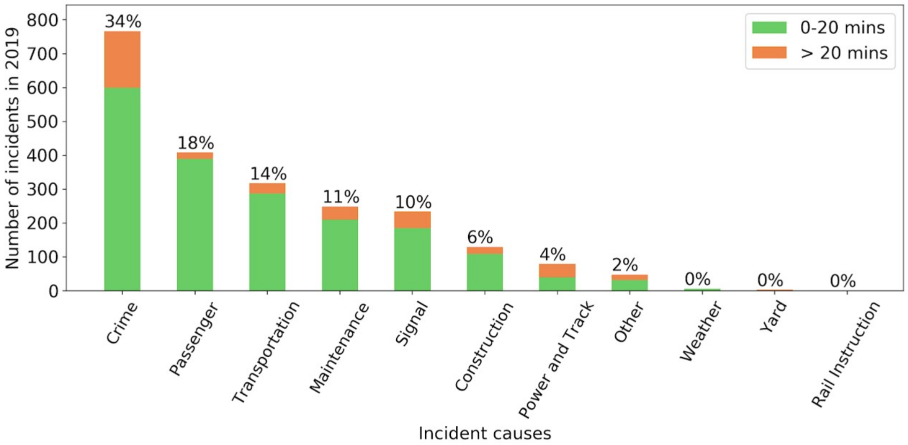


Figure 1-11: Incidents causes distribution (numbers in each bar show the proportion of number of incidents for this cause)

1.5 Methodology and dissertation outline

1.5.1 Chapter 2: System performance evaluation under short-term service suspensions using bulk-service queue model

According to our preliminary analysis of the CTA system, around 80% of incidents in the system are less than 5 minutes, which makes it important to quantify the impact of short-term incidents. The objective of this study is to evaluate how a public transportation system is affected by random short-term service suspension. Specifically, we aim to derive closed-form formulations for the stability condition and mean and variance of the passengers' queue length and waiting time at a platform.

There is no previous study using the bulk-service queue model to evaluate public transit system performance under random service disruptions. Powell [41] has derived closed-form formulations for a $M/G^{[S]}/1$ model (i.e., arbitrary bulk-service distribution) considering one terminal stations. In this study, we aim to derive the formulations of $G^{[S]}$ (i.e., the service distribution) for a single-route public transit system under random service suspension.

We assume that vehicles may have random suspensions when traveling. The most important part of this research is to derive the headway distribution under incidents and the probability generating function (PGF) of the number of arriving passengers within a headway.

Our analysis shows that headways can be represented as the difference between two compound Poisson-exponential variables. Assuming no vehicle overtaking, we approximate the headway as a zero-inflated truncated normal distribution to obtain a closed-form moment generating function. Based on the headway distribution, the PGF of the number of arriving passengers within a headway is derived. This is a theoretical study that aims to extend the bulk-service queueing theory in the literature. The results can be used to calculate the public transit systems' performance and stability under different incident rates and duration lengths in an efficient way.

1.5.2 Chapter 3: Empirical analysis for the impact of service disruptions

For a long-term service disruption, it is important to understand how the operators and passengers are affected during the incident empirically. This topic aims to propose a general urban rail incident analysis framework from the supply and demand sides using AFC and AVL data.

Specifically, on the supply side, we propose an incident-based network redundancy index to analyze the network's ability to provide alternative services under a specific rail disruption. The impacts on the operations are analyzed through the headway changes. On the demand side, the analysis takes place at two levels: aggregate flows and individual responses. We calculate the demand changes of different rail lines, rail stations, bus routes, and bus stops to better understand the passenger flow redistribution under incidents. Individual behavior in terms of passengers' mode choices is analyzed using a binary logit model based on socio-demographics derived from AFC data.

The analysis is conducted for the PT system of the Chicago Transit Authority. Two rail disruption cases are analyzed, one with high network redundancy around the impacted stations and the other with low. Results show that the service frequency of the incident line was largely reduced (by around 30% 70%) during the incident time. Nearby rail lines with substitutional functions are also slightly affected. Passengers showed different behavioral responses in the two incident scenarios. In the low redundancy case, most of the passengers chose to use nearby buses to move, either to their destinations or to the nearby rail lines. In the high redundancy case, most of the passengers transferred directly to nearby lines.

1.5.3 Chapter 4: Inferring passenger behavior responses under disruptions

In a long-term service disruption case study, passengers may change their travel modes during the incident. Understanding their choices is crucial for operational responses.

The objective of this chapter is to use AFC data to infer passengers' travel mode choices during major incidents.

Most of the previous studies used surveys to investigate passengers' travel modes during incidents. Only a limited number of studies proposed inference models of passenger behavior (e.g., waiting, leaving the system) using AFC data. However, these models are all based on a closed system with both tap-in and tap-out data. The methods are rule-based without considering passengers' historical travel information.

This study aims to use a probabilistic framework to infer passengers' travel mode choices during incidents. The basic idea is that, based on individual historical trips and the observed behavior during an incident (from AFC data), we calculate the probability that this observed behavior is caused by the incident (instead of normal behavior).

We enumerate 19 possible behaviors that passengers may exhibit based on the stages of their trips when an incident happened (such as waiting, transferring to a bus, etc.). A probabilistic model is proposed to estimate the mean and variance of the number of passengers in each of the 19 behavioral groups. Results with synthetic data show that the proposed approach can well estimate passengers' behavior.

1.5.4 Chapters 5 and 6: Station-based and individual-based path recommendations under disruptions

During a service disruption, passengers may make suboptimal choices during the incident. For example, according to the empirical analysis, most of the passengers used bus routes that are parallel to the incident rail line. Since the capacity of bus routes is much less than the rail routes, congestion and high waiting times are caused. However, other alternative routes that transfer to rail lines and connect to downtown are not well utilized. Therefore, route recommendations are needed for passengers during incidents to better distribute passengers and utilize the capacity in the system.

Route recommendation problems are quite rich. As shown in Figure 1-12, whether providing information or not and how to provide it requires inputs from both passengers and operators. Different assumptions can lead to different formulations of the problem.

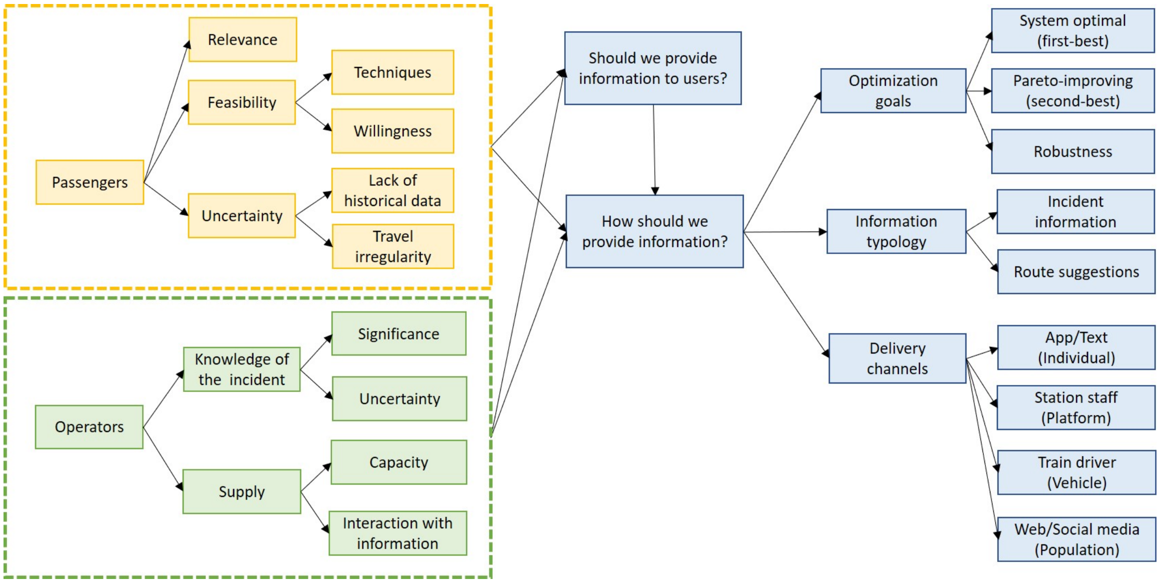


Figure 1-12: Richness of path recommendation problems

In this thesis, we consider two different route recommendation schemes: station-based and individual-based.

Station-based route recommendation: Station-based route recommendations are conducted by providing route information to passengers waiting at a station. Passengers with different departure times and destinations are provided with different strategies. The objective is to minimize passengers' total travel times. We model the path recommendation problem as an optimal flow problem with uncertain demand. To tackle the lack of analytical formulation of travel times due to capacity constraints, we propose a simulation-based first-order approximation to transform the original problem into a linear program. Uncertainties in demand are modeled using robust optimization to protect the path recommendation strategies against inaccurate estimates.

A real-world rail disruption scenario in the CTA system is used as a case study. Results show that even without considering uncertainty, the nominal model can reduce the system travel time by 9.1% (compared to the status quo), and outperforms the benchmark capacity-based path recommendation by around 3%. The average travel time of passengers using the incident line (i.e., passengers receiving recommendations) is reduced even more (-20.6% compared to the status quo). After incorporating the demand uncertainty, the robust model can further reduce system travel times. The best robust model can decrease the average travel time of incident-line passengers by 2.91% compared to the nominal model.

Individual-based route recommendation: Another route recommendation scheme is individual-based, where each individual is recommended a specific route so as to reduce the system travel time considering their preference heterogeneity (see Figure 1-13). In this study, we narrow the scope of the problem by the following assumptions: 1) We know the origins, destinations, and original path preferences of all passengers when the incident takes place. Their original choices may neither be optimal for themselves nor the system. 2) We know a subset of incident-relevant passengers who can receive individual messages by phone. They follow the recommendation with known probabilities. 3) We know the exact incident location, blocked links, and the duration distribution of the incident.

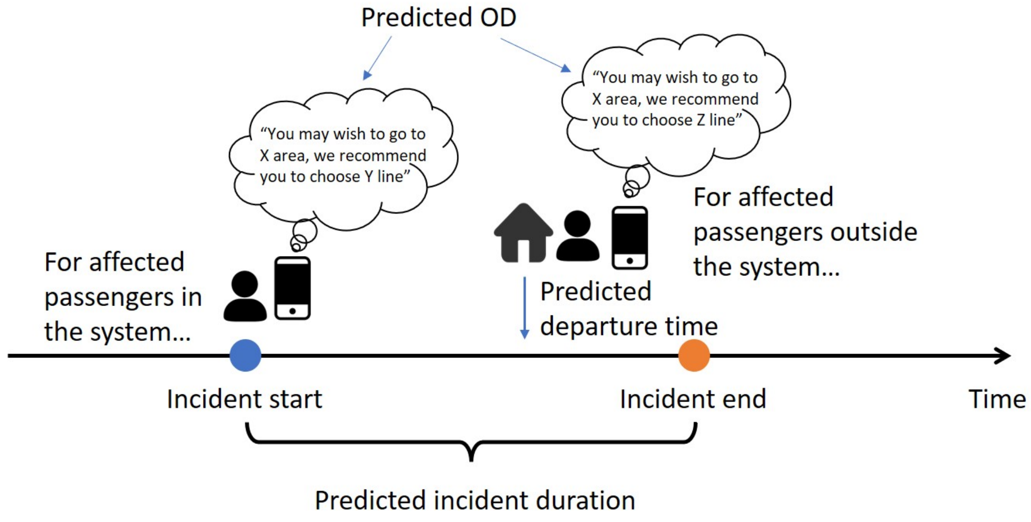

Predicted OD
"You may wish to go to X area, we recommend you to choose Y line"
"You may wish to go to X area, we recommend you to choose Z line"
For affected passengers in the system...
For affected passengers outside the system...
Predicted departure time
Incident start
Incident end
Time
Predicted incident duration

Figure 1-13: Illustration of individual-based path recommendation

We propose a mixed-integer programming (MIP) formulation to model the individual-based path (IPR) recommendation problem. Passengers' behavior uncertainty in path choices given recommendations and their travel time equity are also considered. We model the behavior uncertainty based on passenger's prior preferences and posterior path choice probability distribution with two new concepts: ϵ-feasibility and Γ-concentration, which control the mean and variance of path flows in the optimization problem. The IPR problem with behavior uncertainty is solved efficiently with Bender's decomposition. A post-adjustment heuristic is used to address the equity requirement. The proposed approach is implemented in the CTA system with a real-world urban rail disruption as the case study. Results show that the proposed IPR model significantly reduces the average travel times compared to the status quo and outperforms the capacity-based benchmark path recommendation strategy. We also show that incorporating behavior uncertainty with respect to responses to information achieves lower system travel times than assuming that all passengers would follow the recommendations. The post-adjustment heuristic effectively reduces the difference in passengers' travel times and increases equity, where in this study, equity is defined as all passengers with the same origin, destination, and departure times should have similar travel times if they follow the recommendations. The equity requirement slightly increases the system travel time, showing the trade-off between efficiency and equity. We also show that it is possible to make recommendations so that most of the passengers (e.g., more than 70%) use their preferred paths while only increasing the system travel time by 0.51%.

1.6 Related publications

The dissertation research has resulted in five papers.

The content in Chapter 2 is based on the paper "Resilience of public transit systems under short random service suspensions: A bulk-service queue model" by Baichuan Mo, Li Jin, Haris N. Koutsopoulos, Zuo-Jun Max Shen, Jinhua Zhao [42]. This paper is under review.

The content in Chapter 3 is based on the paper "Impact of unplanned service disruptions on urban public transit systems" by Baichuan Mo, Max Y von Franque, Haris N. Koutsopoulos, John P. Attanucci, and Jinhua Zhao [43]. This paper was presented at Transportation Research Board 100th Annual Meeting.

The content in Chapter 4 is based on the paper “Inferring passenger responses to urban rail disruptions using smart card data: A probabilistic framework” by Baichuan Mo, Haris N. Koutsopoulos, Jinhua Zhao [44]. This paper has been published in Transportation Research Part E: Logistics and Transportation Review.

The content in Chapter 5 is based on the paper “Robust Path Recommendations During Public Transit Disruptions Under Demand Uncertainty” by Baichuan Mo, Haris N. Koutsopoulos, Zuo-Jun Max Shen, Jinhua Zhao [45]. This paper is under review.

The content in Chapter 6 is based on the paper “Individual path recommendation under public transit service disruptions considering behavior uncertainty and equity” by Baichuan Mo, Haris N. Koutsopoulos, Jinhua Zhao [46]. This paper is under review.

Chapter 2

System performance evaluation under short-term service suspensions using a bulk-service queue model

2.1 Introduction

Public transit (PT) systems play a crucial role in cities worldwide, transporting people to jobs, homes, outings, and other activities. However, PT systems are usually susceptible to unplanned delays and service disruptions, which may be caused by equipment failures, weather, passengers, or other internal and external factors.

Short-term service suspension happens frequently in PT systems. According to Mo et al. [43], there are on average 75 incidents happening in the Chicago urban rail system per day and more than 75% of them are less than 5 minutes. Causes for these short-term suspensions can be signal system failures, passenger behavior, and infrastructure problems. For this reason, it is important to recognize how a PT system is affected by these short-term service suspensions.

An important concept related to the system's reaction to incidents is "resilience". Resilience, in the context of managed infrastructure systems, is defined as the endogenous capacity of a system to cope with exogenous perturbations [47]. The measurement of resilience varies in different studies. In this study, we consider two aspects regarding resilience: a) system's ability to stay stable under random service suspensions, and b) system performance changes under random service suspensions. The first aspect requires **stability analysis** of the system. For the second aspect, we calculate the mean and variance of passengers' **queue length** and **waiting time** under random service suspensions.

Queuing behavior at a PT station is usually modeled as a bulk-service queue model [48, 49, 50]. Bulk service means that customers are served in groups rather than individually. At a PT station, with the arrival of vehicles (e.g., buses or trains), a group of passengers will board (i.e., being served in groups). If the vehicle capacity is less than the number of customers waiting, some customers are left behind [51]. Most of the previous studies on a bulk service model for PT systems focus on stations [52, 41, 50]. Islam et al. [49] used a Markov model to extend the station-level analysis to the route level. However, these studies all considered PT systems under normal operating conditions. The studies of PT systems under service suspensions using queuing analysis are limited. Regarding the treatment of service disruptions in bulk-service queue models, Madan [53] first considered a single channel bulk service queue subject to interruptions. They assumed there are two states (work and repair) in the system and derived the probability generating function (PGF) of queue length using steady-state equations. Many researchers extended Madan [53]'s framework by considering more channels [54], more heterogeneous states [55, 56], different service interruption assumptions [57], and different repair policies [58, 59]. However, all these studies assumed that the service is offered with a fixed batch size (i.e., fixed capacity), which is not valid for PT systems where the available vehicle capacity for boarding is a random variable depending on the current vehicle load. Besides, all these studies used steady-state equations to derive multiple PGFs of queue length under different system states (e.g., work and repair). Results are usually mathematically tedious and the queue length and waiting time can only be analyzed with a very small service batch size (e.g., Madan [53] only analyzed the problem with service batch size equal to 1 and 2, for batch size more than 3, the closed-form formulas are hard to derive).

Finally, previous studies usually consider the breakdown of servers. But there is no trivial way to map the "breakdown of servers" to a PT system with valid real-world assumptions because, in a PT system, the assumption that each station is an independent server is not valid.

To fill the research gaps, we propose a bulk queue service-based framework to describe the passenger and vehicle dynamics for a PT system and analyze the system resilience under short random service suspensions. The objective of this study is to derive the stability condition of a PT system and the mean and variance of passengers' queue length and waiting time for each station under random suspensions. This analysis provides important insights into PT systems' resilience and performance changes under service disruptions, which is helpful for future control and planning strategies.

This work can be seen as an extension of Powell [41] and Islam et al. [49] from normal conditions to incident conditions. Powell [41] proposed a bulk service queue model for transportation terminals (i.e., station-level) with analytical queue length and waiting time formulations under normal conditions using transform methods (as opposed to steady-state equations methods) and Islam et al. [49] extended the analysis from station-level to route-level. In this study, we explicitly model the random service suspension in a single-route PT system (in reality, it represents a bus route or one-directional rail line, which is a basic element of more complex PT networks). Different from typical service interruption studies where servers may break down, we assume a **vehicle** in the PT system may suffer from random suspensions. A detailed discussion of this assumption is provided in Section 2.3.2, where we show how it corresponds to many real-world situations and can be seen as the first step toward a general incident representation in PT systems. Under this assumption, we extend Powell [41] and Islam et al. [49]'s work to obtain the mean and variance of passengers' queue length and waiting time at each station in the single route PT system by analyzing the headway distribution under random service suspensions. The major contribution of this chapter is fourfold:

- This is the first study to explore analytically the bulk-service queuing problem involving short random service suspensions applied to PT systems. We model the service suspension in PT systems by analyzing vehicles' speed profiles, which is a novel and practical way to consider "server breakdown" in PT systems.

- We prove that the headway under random service suspensions can be represented as the difference between two compound Poisson exponential variables. We assume there is no vehicle overtaking and approximate the headway distribution as a zero-inflated truncated normal distribution to obtain a closed-form moment generating function. Based on this we derive the PGF and corresponding moments of the number of arrival passengers within a headway (these are critical components for the bulk-service queue model). This is a new analytical contribution to the bulk-service queuing theory.

- Based on Islam et al. [49]'s work, we introduce a Markov chain model to capture the inter-station passenger flow dynamics, which extends the typical bulk-service queuing analysis from the station level to the route level.

- We propose an interpolation-based roots-solving method to find all complex roots for this study's model specification. Roots-solving is an essential step to obtain the queue length and waiting time for the bulk-service queuing model.

The remaining chapter is organized as follows. Section 2.2 reviews the literature on the bulk-service queue problem, random service disruptions, and queuing models for public transit systems. Section 2.3 presents the model settings for a single-route system with random service suspensions. Section 2.4 shows the analysis and derivations of the major results. Section 2.5 provides numerical examples to illustrate the theoretical results and validates the proposed approach using simulation. Section 2.6 concludes the chapter and discusses future research directions.

2.2 Literature review

2.2.1 Bulk queue models

In the bulk service queuing literature, customers are served in a batch of fixed or variable lengths. The service rate may depend on the number of customers waiting for service. The motivation of this model rises from addressing problems in manufacturing systems, elevators, transport systems, etc.

Bailey [60] originated the study of bulk queues by considering a system with simple Poisson arrivals at a server that serves, at particular points in time, all waiting customers up to a fixed capacity c. If no customers are waiting, a zero number of customers are served, implying that the server is never idle. The queue, denoted by M/G^c/1, is described using an embedded Markov chain defined at points of service completions. Immediately following Bailey [60], Downton [61] obtained the waiting time distribution of bulk service queues by considering random arrivals and random service time distribution. Jaiswal [62] confirmed the results in Downton [61]. He derived the waiting time distribution using the embedded Markov-chain approach.

The general bulk service rule was first introduced by Neuts [63], where a server, upon finishing a batch, may remain idle if there are fewer than m customers waiting for service. Thus all departing batches from the queue have at least m customers, although no more than the service capacity.

Along with and after those milestone studies, papers have appeared which can be differentiated on the basis of the queuing types (arrival process, service process, number of servers), objectives (queues, waiting times, busy periods, etc.), the time domain of the solution (i.e., steady-state or transient), and the method of solution (transforms or direct numerical methods). Chaudhry and Templeton [64] and Sasikala and Indhira [65] provide a more complete review of the developments in bulk service queue models.

2.2.2 Random service disruptions

The subject of queuing systems wherein the service channel is subject to breakdowns is a popular subject that has received a lot of attention in the past fifty years. For a recent survey of the related literature, readers can refer to Krishnamoorthy et al. [66].

However, most of the research on this topic deals with models where the server serves the customers one at a time. The related literature on bulk service is limited. Madan [53] studied a single-channel queueing system with Poisson arrivals and exponential service in batches of fixed size. The system is subject to random interruptions with an operating state and a repairing state. Both the operating times and the repair times of the service channel are assumed to be exponential. Madan [55] generalized the model in Madan [53] to the case where the repairs are performed in two phases. Singh and Ram [54] extended the model in Madan [53] by considering a system with three identical channels, with operating and repair times for all three service channels distributed exponentially. Jayaraman et al. [57] considered a single-server queueing system with general bulk service. Arrivals are Poisson but alternate between two modes according to whether the server is operational or in the failed state. The duration of the operating and repair periods are exponential and phase-type distributions, respectively. Tadj and Choudhury [58] analyzed a bulk service queueing system with an unreliable server, Poisson input, and general service and repair times. Tadj et al. [59] considered a bulk service queuing system where service is provided to groups of customers of fixed size. Service consists of two consecutive phases and may take a vacation following the second phase of service. While providing service, the server may break down and a delay period precedes the repair period.

2.2.3 Queuing models in public transit systems

Queuing theory in PT systems is usually conducted at the station level, aiming at obtaining the mean queue length and waiting time. In the case of regular services where headways are equal, assuming that a) passengers arrive at stops according to a Poisson process and b) passengers can be served by the first arriving vehicle, the mean waiting time of passengers ($\mathbb{E}[W]$) is given by:

$$\mathbb{E}[W] = H/2, \tag{2.1}$$

where H is the service headway and W is the passenger waiting time. This is the most widely used queuing assumption in transit studies [67, 68, 69]. However, in the case where service is not reliable, the assumption of regular service can be problematic. Numerous models have been proposed to account for the stochastic nature of headways [70, 71]. A well-known model proposed by Osuna and Newell [71] with Poisson arrival passengers and stochastic headways is

$$\mathbb{E}[W] = \frac{1}{2} \cdot \left[\mathbb{E}[H] + \frac{\mathrm{Var}[H]}{\mathbb{E}[H]}\right], \tag{2.2}$$

where $\mathbb{E}[H]$ and $\mathrm{Var}[H]$ are the expectation and variance of headways, respectively. In the case of regular services, the variance is zero and the model reverts to Equation 2.1.

However, the results in Eq. 2.1 and 2.2 do not consider the vehicle capacity (i.e., they assume all passengers can board the first vehicle). In a congested PT system, passengers may be left behind due to limited vehicle capacity, leading to an increase in waiting times [4]. Bulk service queue models have been applied in PT systems to capture the effects of capacity constraints. Powell [48, 72, 41] used a bulk service queue model to calculate the passenger queue length and waiting times at public transportation terminals. The closed-form mean and variance for these two quantities are derived using a transform method. Rapoport et al. [73] studied bulk service queues with constant or variable capacity and exogenously determined arrival times (e.g., passenger arrivals based on smart card data). Wang et al. [50] proposed a bulk service and batch arrival queuing model with reneging behavior to estimate passengers' waiting for public transport services.

All the aforementioned studies consider the queuing analysis at the station level. The extension of queuing analysis from a station level to a route level is not a trivial problem. First, the boarding and alighting behavior at upstream stations affect the available capacity distribution at downstream stations. Second, headways may be correlated across stations, leading to different headway distributions for different stations [74, 75]. To address this problem, Islam et al. [49, 76] proposed a Markov model to combine the Powell [48] and Hickman [75]'s approaches and used a bulk service model to analyze system performance at the route level. However, a limitation of their research is that the calculation of headway correlation does not consider the vehicle capacity (though the capacity constraint is considered in the queuing behavior), resulting in the inconsistency of model assumptions. Also, they assume that headways follow the Erlang distribution, which leads to model tractability but is not consistent with empirical observations [77].

Our study can be seen as an extension of Powell [41]'s and Islam et al. [49]'s work to incorporate random service suspensions in a PT system with more consistent and reasonable assumptions. And we also characterize the headway distribution under service suspensions.

2.2.4 Service interruptions in public transit systems

Studies on service interruption in public transit systems can be categorized into two groups: impact analysis and operations control. Impact analysis studies have used a variety of methods to analyze the impact of service disruptions on performance and level of service. Of these methods, the most common is based on graph theory, surveys, simulation, and empirical data. Graph theory-based methods usually derive resilience or vulnerability indicators based on the network topology [78, 30, 79, 80]. These methods are effective for understanding high-level network properties related to incidents. Survey-based methods investigate passenger behavior and opinions during incidents [81, 82, 83, 84, 85]. Passengers' individual-level behavior is analyzed and understood using econometric models. Simulation-based methods simulate passenger flows on the transit network under incident scenarios [86, 87, 88]. The empirical data-based methods use smart card and vehicle location data to analyze real-world incident impacts [89, 90, 43]. These studies can output many metrics of interest such as vehicle load, travel delays caused by incidents, distribution of the impact, etc. Studies focusing on operations control under service disruptions address aspects including shuttle bus design [91, 92], vehicle holding [93], integrating local services [94], and timetable adjustment [95].

The resilience analysis presented in this study belongs to the “impact analysis” category, which aims to obtain stability conditions for PT systems and the mean and variance of passengers’ queue length and waiting time of each station under short random suspensions. None of the previous studies has used the bulk service model for this type of analysis.

2.3 Model

2.3.1 Single-route public transit system and vehicle movements

Consider a single-route PT system with N stations as shown in Figure 2-1. Vehicles are dispatched from a transportation hub (also referred to as station 0) and travel from station 1 to station N. At a specific station n, we assume that passenger arrivals follow a Poisson process with a fixed rate $\lambda^{(n)}$ during the time period of interest. When a vehicle arrives at station n, each passenger in the vehicle has a probability of $\alpha^{(n)}$ to alight. Thus, the number of alighting passengers at station n follows a binomial distribution. Poisson arrivals and binomial alighting are two common assumptions in much of the PT-related literature [75]. In this study, we do not consider reneging behavior of passengers (i.e., passengers may leave the system if they have waited for too long) since the focus of the study is on “short” service suspensions and we assume passengers choose to wait. Empirical studies [89, 96] show that passengers start to leave the system only when delays are large (e.g., 30 minutes or more). Incorporating balking and reneging is outside the scope of this study and can be a future extension of this work.

Let $l = 1, 2, ...$ be a superscript denoting the vehicle run number (or vehicle ID). Smaller l means vehicles dispatched at an earlier time. Figure 2-2 summarizes the

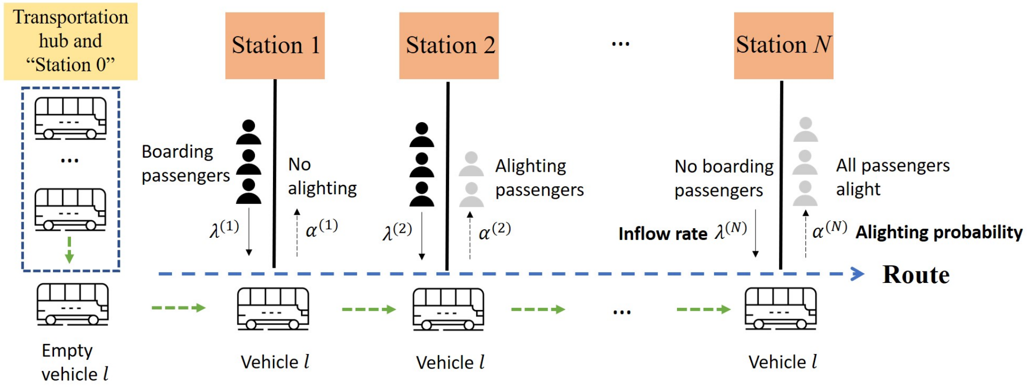


Figure 2-1: Schematic presentation of a single-route public transit system

vehicle and passenger interactions at station n over time. Let $t_A^{(n,l)}$ be the time that vehicle l arrives at station n, and $t_D^{(n,l)}$ the time that vehicle l departs station n. $H^{(n,l)}$ is the headway between the preceding vehicle $l-1$ and vehicle l, as they depart from stop n (i.e., $H^{(n,l)} = t_D^{(n,l)} - t_D^{(n,l-1)}$). When a vehicle arrives at station n, some of the on-board passengers alight first, then the queuing passengers start to board. Let $Q^{(n,l)}$ be the number of queuing passengers when vehicle l arrives at station n, $R^{(n,l)}$ the number of left behind passengers when vehicle l departs station n, and $Y^{(n,l)}$ the number of passengers arriving between $t_D^{(n,l)}$ and $t_A^{(n,l+1)}$. By definition,

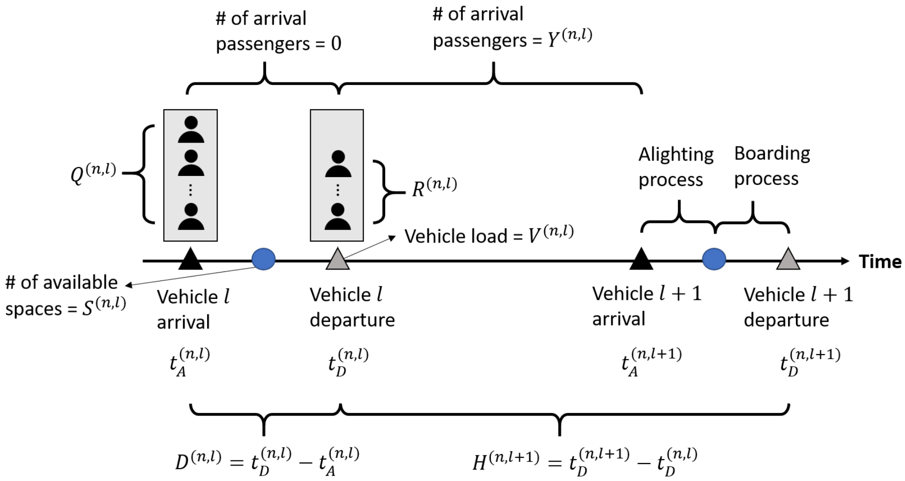
$$Q^{(n,l+1)} = R^{(n,l)} + Y^{(n,l)}. \tag{2.3}$$

In this study, we assume that the dwell time (i.e., $t_D^{(n,l)} - t_A^{(n,l)}$) is negligible compared to the vehicle travel time ($t_A^{(n+1,l)} - t_D^{(n,l)}$) such that the number of passengers arriving during the dwell time is zero (same assumption as in Powell [48]). Then, given the headway $H^{(n,l+1)}$, $Y^{(n,l)}|H^{(n,l+1)}$ follows a Poisson distribution with parameter $\lambda^{(n)}H^{(n,l+1)}$:

$$Y^{(n,l)} \mid H^{(n,l+1)} \sim \textbf{Poi}(\lambda^{(n)}H^{(n,l+1)}). \tag{2.4}$$

In other words, $Y^{(n,l)}$ can be seen as the number of arriving passengers within a headway (i.e., $H^{(n,l+1)}$).

From the vehicle's perspective, let $S^{(n,l)}$ be the number of available space after



# of arrival passengers = 0
# of arrival passengers = $Y^{(n,l)}$
$Q^{(n,l)}$
$R^{(n,l)}$
Vehicle load = $V^{(n,l)}$
Alighting process
Boarding process
Time
# of available spaces = $S^{(n,l)}$
Vehicle l arrival
Vehicle l departure
Vehicle $l+1$ arrival
Vehicle $l+1$ departure
$t_A^{(n,l)}$
$t_D^{(n,l)}$
$t_A^{(n,l+1)}$
$t_D^{(n,l+1)}$
$D^{(n,l)} = t_D^{(n,l)} - t_A^{(n,l)}$
$H^{(n,l+1)} = t_D^{(n,l+1)} - t_D^{(n,l)}$

Figure 2-2: Diagram of vehicles and passengers interaction at station n in the time dimension

passengers alighting from vehicle l at station n, $G^{(n,l)}$ the number of remaining passengers on vehicle l after passengers alighting at station n. By definition,

$$G^{(n,l)} = C - S^{(n,l)}, \tag{2.5}$$

where C is the capacity of vehicles. Denote $V^{(n,l)}$ as the vehicle load (i.e., number of on-board passengers) when vehicle l departs station n (i.e., the vehicle load when it arrives at station $n+1$). Then, the number of alighting passengers from vehicle l at station n given $V^{(n-1,l)}$ follows a binomial distribution:

$$V^{(n-1,l)} - G^{(n,l)} \mid V^{(n-1,l)} \sim \mathsf{Bin}(V^{(n-1,l)}, \alpha^n). \tag{2.6}$$

2.3.2 Random service suspensions and vehicle speed profile

Let us assume that there are random service suspensions when a vehicle travels in the system. Given these disturbances, the speed curve of vehicle l from station n to $n+1$ can be described by the red line in Figure 2-3. Every random incident causes a speed reduction or stop of the vehicle. In reality, these incidents can be caused by many reasons. For example, in a bus system, they may be caused by traffic congestion or accidents, drivers' or passengers' behavior, vehicle engine issues, etc. In a rail system, the reasons may be signal failures, infrastructure problems, and drivers' or passengers' behavior. The speed curve is a general representation of different incidents, interruptions, suspensions, or disruptions that impede the vehicle's movement.

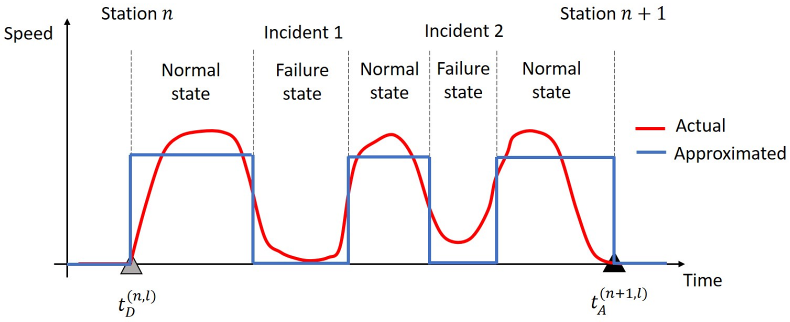


Figure 2-3: Schematic speed curve of vehicle l traveling from station n to $n+1$

The actual vehicle speed profile under interruptions can be complicated. To facilitate mathematical modeling, we assume that the speed of a vehicle under random interruptions can be approximated by an impulse function (blue line in Figure 2-3). The impulse function separates the vehicle trajectory into traveling and stopping phases, denoted as normal state and failure state, respectively. In the normal state, a vehicle travels at a constant speed. Once an incident happens, the vehicle stops immediately and enters the failure state. We assume that, in a sufficiently small time interval, Δ, the probability of incident occurrence is $\gamma\Delta$. Furthermore, the duration of an incident follows an exponential distribution with rate θ (i.e., mean of $\frac{1}{\theta}$). Then, the state of a vehicle is a two-state Markov process (Figure 2-4) with the state space of {Normal, Failure}.

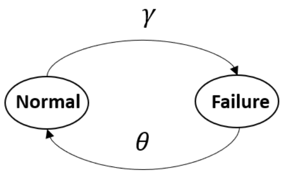


Figure 2-4: Transition diagram of vehicle states

For the two-state Markov process, the duration of failure and normal states follows the exponential distribution with rates θ and γ, respectively.

Approximating the actual speed curve as an impulse function can be seen as the first step toward a general incident representation in PT systems. Actually, any type of incident can be represented as a mixture of different types of normal and failure states. The normal and failure states can be defined with heterogeneous occurrence probabilities and duration for different categories of incidents, which results in a more sophisticated speed curve representation.

2.3.3 Headway under random service suspensions

Under the assumption of an impulse-function speed profile, all vehicles have the same fixed travel speed under the normal state. Therefore, if there is no incident in the system, all stations have the same deterministic headway (denoted as $\bar{H}$). The relationship among $\bar{H}$, route cycle time $\bar{E}$ (i.e., the time that a vehicle travels from the transportation hub to the last station and returns to the hub), and fleet size (denoted as $\bar{F}$) for the route is

$$\bar{H} = \frac{\bar{E}}{\bar{F}} \tag{2.7}$$

With random service suspensions, the route cycle time would increase. There are two possible responses for the transit agency: 1) To maintain the same planned headway $\bar{H}$, the agency needs to increase the fleet size for the route. 2) With the same fleet size (i.e., limited resources), the agency would have to increase the planned headway $\bar{H}$. In this study, we consider the second scenario because it reflects incidents' impact on headway and service performance, which is more relevant to this chapter's topic.

Therefore, we assume that at the route planning stage, transit agencies have an estimate of the average delay in the cycle time, $\bar{D}$. Let $I^{(n,l)}$ be the total duration of all incidents happening during the vehicle l's travel time from the transportation hub to station n (a random variable). Then $\mathbb{E}[I^{(N,l)}]$ is the expected incident duration for a vehicle traveling from the transportation hub to the last station N. Assuming the road conditions for two directions of the route are the same, then the total estimated delay for the cycle trip is

$$\bar{D} = 2 \cdot \mathbb{E}[I^{(N,l)}] \tag{2.8}$$

It is worth noting that some transit agencies may plan the headway by assuming a larger delay (e.g., not the mean, but the 85% percentile). Hence, we may also formulate $\bar{D}$ as a general function of $\mathbb{E}[I^{(N,l)}]$. In this study, we adopt Eq. 2.8 for simplicity. Then, the incident-adjusted planned headway (denoted as $\bar{H}^{\text{Adj}}$) is

$$\bar{H}^{\text{Adj}} = \frac{\bar{E} + \bar{D}}{\bar{F}} = \bar{H} + \frac{2 \cdot \mathbb{E}[I^{(N,l)}]}{\bar{F}} \tag{2.9}$$

where $\frac{2 \cdot \mathbb{E}[I^{(N,l)}]}{\bar{F}}$ is the **planned** headway adjustment term due to incidents. Note that we assume $I^{(N,l)}$ are identically distributed for all l. So the incident-adjusted planned headway is not affected by vehicle ID. In this study, we assume that the single-route PT system will dispatch vehicles based on the incident-adjusted planned headway $\bar{H}^{\text{Adj}}$ and all dispatches are on time.

Let $T^{(n)}$ be the travel time for vehicles from the transportation hub to station n when there is no incident (a fixed constant in this study due to the fixed speed assumption). Without loss of generality, let us assume vehicle $(l-1)$ departs from the transportation hub at time 0. Considering random service suspensions, vehicle $(l-1)$'s departure time from station n is:

$$t_D^{(n,l-1)} = T^{(n)} + I^{(n,l-1)} \tag{2.10}$$

Note that the dwell time is ignored as we assumed before.

Given that the incident-adjusted planned headway is $\bar{H} + \frac{2 \cdot \mathbb{E}[I^{(N,l)}]}{\bar{F}}$, the departure time of vehicle l from station n is

$$t_D^{(n,l)} = \bar{H} + \frac{2 \cdot \mathbb{E}[I^{(N,l)}]}{\bar{F}} + T^{(n)} + I^{(n,l)} \tag{2.11}$$

Therefore, with random incidents, the actual headway of vehicle l at station n is

$$H^{(n,l)} = t_D^{(n,l)} - t_D^{(n,l-1)} = \bar{H} + \frac{2 \cdot \mathbb{E}[I^{(N,l)}]}{\bar{F}} + I^{(n,l)} - I^{(n,l-1)}. \tag{2.12}$$

In this study, we assume $I^{(n,l)}$ and $I^{(n,l-1)}$ are independent. This assumption facilitates closed-form derivations. In reality, if the incidents are caused by road congestion or infrastructure issues, it is possible that the incident durations for two consecutive vehicles passing through the same route segment are correlated. However, addressing the correlation is not a trivial problem in the bulk service queue model [48] and is beyond the scope of this study.

2.4 Analysis

The objective of this study is to derive the stability conditions of a PT system and the mean and variance of passengers' queue length and waiting time at each station under random service suspensions. Figure 2-5 shows how the distributions of different random variables (particularly, $S^{(n,l)}, V^{(n,l)}, Q^{(n,l)}$) are calculated. The major calculation consists of three parts:

- Given the distribution of $V^{(n-1,l)}$, calculate the distribution of $S^{(n,l)}$. The details are shown in Section 2.4.1

- Given the distribution of $S^{(n,l)}$, calculate the distribution of $Q^{(n,l)}$ and the mean and variance of queue length and waiting time at station n. This is discussed in Section 2.4.3.

- Given the distribution of $S^{(n,l)}$ and $Q^{(n,l)}$, calculate the distribution of $V^{(n,l)}$, which is discussed in Section 2.4.2

With the three components, we can derive the distribution of $S^{(n,l)}, Q^{(n,l)}, V^{(n,l)}$ for all $n = 1, ..., N$ given the distribution of $V^{(0,l)}$ (i.e., vehicle load when vehicle l arrives at the first station, it is always zero by definition). Note that, in this section, we focus on the steady-state distribution of these variables (i.e., $l \to \infty$).

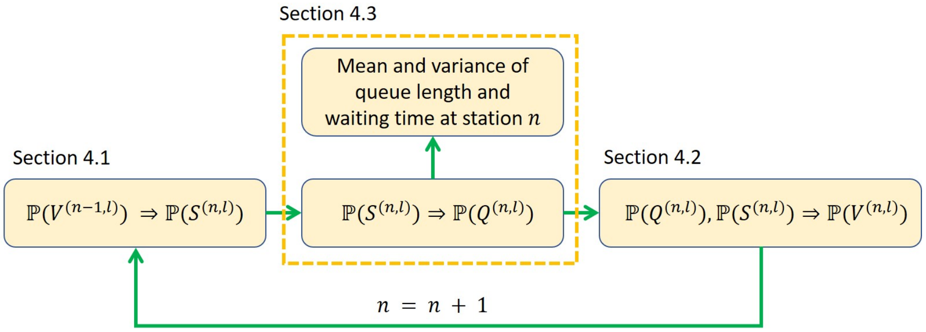

Section 4.3
Mean and variance of queue length and waiting time at station n
Section 4.1
$\mathbb{P}(V^{(n-1,l)}) \Rightarrow \mathbb{P}(S^{(n,l)})$
$\mathbb{P}(S^{(n,l)}) \Rightarrow \mathbb{P}(Q^{(n,l)})$
Section 4.2
$\mathbb{P}(Q^{(n,l)}), \mathbb{P}(S^{(n,l)}) \Rightarrow \mathbb{P}(V^{(n,l)})$
$n = n + 1$

Figure 2-5: Analysis framework

After obtaining the corresponding distributions, we discuss the stability conditions in Section 2.4.4 and summarize the approach in Section 2.4.5.

2.4.1 Available vehicle space steady-state distribution

In this section, we aim to derive the steady-state distribution of $S^{(n,l)}$ given the steady-state distribution of $V^{(n-1,l)}$. Define $v_k^{(n,l)} := \mathbb{P}(V^{(n,l)} = k)$, $s_k^{(n,l)} := \mathbb{P}(S^{(n,l)} = k)$, and $g_k^{(n,l)} := \mathbb{P}(G^{(n,l)} = k)$ for all $k = 0, 1, ..., C$. Assuming that the steady state probabilities for all variables exist (the stability condition will be discussed in Section 2.4.4), we have $v_k^{(n)} := \lim_{l\to\infty} v_k^{(n,l)} = \mathbb{P}(V^{(n)} = k)$, $s_k^{(n)} := \lim_{l\to\infty} s_k^{(n,l)} = \mathbb{P}(S^{(n)} = k)$, and $g_k^{(n)} := \lim_{l\to\infty} g_k^{(n,l)} = \mathbb{P}(G^{(n)} = k)$, where $V^{(n)} = \lim_{l\to\infty} V^{(n,l)}$, $S^{(n)} = \lim_{l\to\infty} S^{(n,l)}$, and $G^{(n)} = \lim_{l\to\infty} G^{(n,l)}$.

Proposition 1. *$\forall$ n = 1,..,N, given the distribution of $V^{(n-1)}$ (i.e., $v^{(n)} := [v_0^{(n-1)}, ..., v_C^{(n-1)}] \in \mathbb{R}^{C+1}$), the distribution of $S^{(n)}$ (i.e., $s^{(n)} := [s_0^{(n-1)}, ..., s_C^{(n-1)}] \in \mathbb{R}^{C+1}$) is given as:*

$$s_k^{(n)} = g_{C-k}^{(n)} \qquad \forall k = 0, 1, ..., C, \tag{2.13}$$

where $g^{(n)} := [g_0^{(n)}, ..., g_C^{(n)}] \in \mathbb{R}^{C+1}$ and

$$g^{(n)} = v^{(n-1)} A^{(n)}, \tag{2.14}$$

$A^{(n)}$ is a $(C+1)$ by $(C+1)$ matrix with the element in row i and column j equal to

$a_{ij}^{(n)}$, and $a_{ij}^{(n)}$ is defined as

$$a_{ij}^{(n)} = \begin{cases} 1, & \textit{if } i = 0 \textit{ and } j = 0 \\ \binom{i}{i-j}(\alpha^{(n)})^{i-j}(1-\alpha^{(n)})^{j}, & \textit{if } i \geq j \textit{ and } i,j \neq 0 \\ 0, & \textit{otherwise} \end{cases} \qquad \forall\, i,j = 0,1,...,C \tag{2.15}$$

Proof. When vehicle l arrives at station n, by definition, there are $V^{(n-1,l)}$ number of passengers in the vehicle. Given that there are i passengers on-board when vehicle l arrives at station n, let the probability that there are j passengers remaining on the vehicle be $a_{ij}^{(n)}$. $a_{ij}^{(n)}$ also represents the probability of $i-j$ passengers alighting, which follows a binomial distribution with parameters i and $\alpha^{(n)}$ (if $i \geq j$ and $i,j \neq 0$). Hence, $a_{ij}^{(n)}$ can be expressed as Eq. 2.15. Then we have

$$g^{(n,l)} = v^{(n-1,l)}A^{(n)} \tag{2.16}$$

where $v^{(n-1,l)} = [v_0^{(n-1,l)},...,v_C^{(n-1,l)}] \in \mathbb{R}^{C+1}$, $g^{(n,l)} = [g_0^{(n,l)},...,g_C^{(n,l)}] \in \mathbb{R}^{C+1}$. According to the relationship between $S^{(n,l)}$ and $G^{(n,l)}$ as shown in Eq. 2.5, the distribution of the number of available spaces after alighting is simply

$$s_k^{(n,l)} = g_{C-k}^{(n,l)} \qquad \forall k = 0,1,...,C \tag{2.17}$$

Note that Eq. 2.16 and 2.17 hold for all l. Since we assume the steady state distributions exist, letting $l \to \infty$ on both sides of Eq. 2.16 and 2.17 completes the proof. □

2.4.2 Vehicle load steady-state distribution

In this section, we derive the steady-state distribution of $V^{(n,l)}$ given the steady-state distribution of $G^{(n,l)}$ and $Q^{(n,l)}$. Define $q_k^{(n)} := \lim_{l\to\infty} q_k^{(n,l)} = \mathbb{P}(Q^{(n)} = k)$, where $Q^{(n)} = \lim_{l\to\infty} Q^{(n,l)}$ and $q_k^{(n,l)} = \mathbb{P}(Q^{(n,l)} = k)$. Denote the first C elements of the steady-steady queue length distribution as $q_{0:C-1}^{(n)}$, where $q_{0:C-1}^{(n)} = [q_0^{(n)},...,q_{C-1}^{(n)}] \in \mathbb{R}^C$.

Proposition 2. *$\forall$ n = 1,..,N, given the distribution of $G^{(n)}$ (i.e., $g^{(n)}$) and $q^{(n)}_{0:C-1}$, the distribution of $V^{(n)}$ can be expressed as:*

$$v^{(n)} = g^{(n)} B^{(n)} \tag{2.18}$$

where $B^{(n)}$ is a matrix with the element in row i and column j equal to $b^{(n)}_{ij}$:

$$b^{(n)}_{ij} = \begin{cases} q^{(n)}_{j-i}, & \textit{if}\ 0 \le i \le j < C \\ 1 - \sum\limits_{k=0}^{C-i-1} q^{(n)}_k, & \textit{if}\ j = C \ \textit{and}\ 0 \le i < C \\ 1, & \textit{if}\ i = j = C \\ 0, & \textit{otherwise} \end{cases} \qquad \forall\, i, j = 0, 1, ..., C \tag{2.19}$$

Proof. Let $b^{(n,l)}_{ij}$ be the probability that the load of vehicle l is j after passenger boarding given that there are i passengers on-board after alighting (i.e., $G^{(n,l)} = i$) at station n. Hence, if $0 \le i \le j < C$, $b^{(n,l)}_{ij}$ is simply the probability that there are $j - i$ passengers in the queue (such that after boarding there are j passengers on the vehicle):

$$b^{(n,l)}_{ij} = q^{(n,l)}_{j-i}, \quad \text{if } 0 \le i \le j < C. \tag{2.20}$$

If $j = C$ and $0 \le i < C$, the vehicle reaches capacity after boarding. Then $b^{(n,l)}_{ij}$ should be the probability that the number of passengers in the queue is greater than or equal to $C - i$ (i.e., one minus the probability that there are less than or equal to $C - i - 1$ passengers in the queue). This leads to:

$$b^{(n,l)}_{ij} = 1 - \sum_{k=0}^{C-i-1} q^{(n,l)}_k, \quad \text{if } j = C \text{ and } 0 \le i < C. \tag{2.21}$$

When $i = j = C$, we simply have $b^{(n,l)}_{ij} = 1$ because regardless the number of waiting passengers in the queue, nobody can board as the vehicle is full. Given Eq. 2.19, the vehicle load distribution can be calculated as

$$v_j^{(n,l)} = \sum_{i=1}^{C} g_i \cdot b_{ij}^{(n,l)} \qquad \forall j = 0, 1, ..., C \tag{2.22}$$

Notice that Eq. 2.22 holds for all l. As we assume that the steady state distributions exist, taking $l \to \infty$ for both sides of Eq. 2.22 and rewriting it in a matrix form completes the proof. □

2.4.3 Queuing analysis at a station

In this section, assuming that we know the distribution of $S^{(n)}$ (i.e., $s^{(n)} = [s_0^{(n)}, ..., s_C^{(n)}] \in \mathbb{R}^{C+1}$), our goal is to derive $q_{0:C-1}^{(n,l)}$ and the mean and variance of passenger queue length and waiting time.

Probability generating function of queue length

We start with deriving the probability generating function (PGF) for $Q^{(n)}$, where $Q^{(n)} = \lim_{l\to\infty} Q^{(n,l)}$.

Proposition 3. *$\forall$ n = 1,..,N, given the distribution of $S^{(n)}$ (i.e., $s^{(n)}$), the PGF of $Q^{(n)}$ can be expressed as:*

$$Q(z) = \frac{\sum_{u=0}^{C} s_u^{(n)} \left[\sum_{i=0}^{u} q_i^{(n)} (z^C - z^{C-u+i})\right]}{\frac{z^C}{Y(z)} - \sum_{u=0}^{C} s_u^{(n)} z^{C-u}}, \tag{2.23}$$

where $Y(z)$ is the PGF of $Y^{(n)}$ and $Y^{(n)} = \lim_{l\to\infty} Y^{(n,l)}$ is the number of arrival passengers at station n within a headway at the steady state.

Proof. The proof follows a similar idea in Powell [48] and is attached in 2.7.1. The difference from Powell [48] is that we consider an arbitrary vehicle capacity distribution $s_0^{(n)}, ..., s_C^{(n)}$, while in Powell [48] the capacity is fixed. Note that Powell [41] provided an equivalent formulation as Eq. 2.23 with variable vehicle capacities using the transform of $S^{(n)}$. □

In Eq. 2.23, there are C unknown variables, $q_0^{(n)}, ..., q_{C-1}^{(n)}$. Note that $q_C^{(n)}$ does not appear in $Q(z)$ because when $u = C$ and $i = C$, we have $q_C^{(n)}(z^C - z^{C-u+i}) \equiv 0$. To quantify $Q(z)$, Rouche's theorem is used [97]. Let $\textsc{Num}(z)$ and $\textsc{Den}(z)$ be the numerator and denominator of $Q(z)$ (i.e., $Q(z) = \frac{\textsc{Num}(z)}{\textsc{Den}(z)}$). As shown in Powell [48], one can prove that $\textsc{Den}(z)$ (i.e., $\frac{z^C}{Y(z)} - \sum_{u=0}^{C} s_u^{(n)} z^{C-u}$) has exactly C complex roots within (or on) the unit circle on a complex plane using Rouche's theorem. Notice that for any $z \in \mathbb{C}$ that satisfies $|z| \leq 1$, where $\mathbb{C}$ is the set of complex numbers, the generating function $Q(z)$ must be analytic. Therefore, if z^* is the root of $\textsc{Den}(z)$ (i.e., $\textsc{Den}(z^*) = 0$), it should also be the root of $\textsc{Num}(z)$ (i.e., $\textsc{Num}(z^*) = 0$) such that $Q(z)$ is analytic [98]. Hence, one can solve for $q_0^{(n)}, ..., q_{C-1}^{(n)}$ using the following two steps:

- **Step 1**: Solve $\textsc{Den}(z) = 0$ for C different roots $z_0^*, ..., z_{C-1}^* \in \mathbb{C}$ that satisfy $|z_i^*| \leq 1$, $\forall\ 0 \leq i \leq C-1$. Note that $z = 1$ is always a root of $\textsc{Den}(z)$. But it does not give information about $q_0^{(n)}, ..., q_{C-1}^{(n)}$ as $\textsc{Num}(1) = 0$ is naturally satisfied. Hence, we adopt the convention that $z_0^* = 1$.

- **Step 2**: Combining $\textsc{Num}(z_i^*) = 0$ ($\forall\ 1 \leq i \leq C-1$) and $Q(1) = 1$, solve for $q_0^{(n)}, ..., q_{C-1}^{(n)}$ (there are C system equations and C unknown variables). Note that when $z \to 1$, both $\textsc{Num}(z)$ and $\textsc{Den}(z)$ approach 0. Therefore, using L'Hopital's rule,

$$\lim_{z\to 1} Q(z) = \lim_{z \to 1} \frac{\textsc{Num}'(z)}{\textsc{Den}'(z)} = \frac{\sum_{u=0}^{C} s_u^{(n)} \left[\sum_{i=0}^{u} q_i^{(n)}(u-i)\right]}{\bar{S}^{(n)} - \bar{Y}^{(n)}} = 1 \tag{2.24}$$

where $\bar{S}^{(n)} = \sum_{u=0}^{C} u s_u^{(n)} = \mathbb{E}[S^{(n)}]$, $\bar{Y}^{(n)} = Y'(1) = \mathbb{E}[Y^{(n)}]$. Eq. 2.24 is the equation used to solve for $q_{0:C-1}^{(n)}$ (instead of directly using $Q(1) = 1$).

Queue length distribution

Though $q_0^{(n)}, ..., q_{C-1}^{(n)}$ can be obtained by solving C system equations as mentioned in Section 2.4.3, we provide a simpler way to calculate $q_{0:C-1}^{(n)}$, which is known as matching the polynomial coefficients.

Proposition 4. *$\forall$ n = 1,..,N, given the distribution of $S^{(n)}$ (i.e., $s^{(n)}$), all complex roots of* $\textsc{Den}(z)$ *(i.e., $z_0^*, ..., z_{C-1}^*$), and $\bar{Y}^{(n)}$, if $s_C^{(n)} > 0$, then $q_{0:C-1}^{(n)}$ can be solved as:*

$$q_0^{(n)} = \frac{1}{s_C^{(n)}}(\bar{S}^{(n)} - \bar{Y}^{(n)}) \prod_{i=1}^{C-1} \frac{z_i^*}{z_i^* - 1}, \tag{2.25}$$

and

$$q_{0:C-1}^{(n)} = \tilde{\eta}^{(n)}(\Lambda^{(n)})^{-1}, \tag{2.26}$$

where $\tilde{\eta}^{(n)} = [s_C^{(n)} q_0^{(n)} \eta_0^{(n)}, s_C^{(n)} q_0^{(n)} \eta_1^{(n)}, ..., s_C^{(n)} q_0^{(n)} \eta_{C-1}^{(n)}] \in \mathbb{R}^C$ and

$$\Lambda^{(n)} = \begin{bmatrix} s_C^{(n)} & s_{C-1}^{(n)} & s_{C-2}^{(n)} & ... & s_1^{(n)} \\ 0 & s_C^{(n)} & s_{C-1}^{(n)} & ... & s_2^{(n)} \\ ... & 0 & s_C^{(n)} & ... & s_3^{(n)} \\ 0 & ... & 0 & ... & s_4^{(n)} \\ 0 & 0 & ... & ... & ... \\ 0 & 0 & 0 & ... & s_C^{(n)} \end{bmatrix} \in \mathbb{R}^{C \times C}. \tag{2.27}$$

$\eta_j^{(n)}$ is the polynomial coefficient of z^j in $\prod_{i=0}^{C-1}\left(1 - \frac{z}{z_i^*}\right)$ (i.e., $\sum_{j=0}^{C} \eta_j^{(n)} z^j := \prod_{i=0}^{C-1}\left(1 - \frac{z}{z_i^*}\right)$). As z_i^* is specified for station n, a superscript n is added to the coefficients.

Proof. The derivation is shown in 2.7.2. □

Note that assuming $s_C^{(n)} > 0$ in Proposition 4 is not restrictive because otherwise we can reduce C such that $s_C^{(n)} > 0$ always holds.

Analytical formulation of mean and variance of queue length and waiting time

After solving for $q_0^{(n)}, ..., q_{C-1}^{(n)}$, $Q(z)$ is determined. The expectation and variance of the queue length at station n can be written by definition as:

$$\mathbb{E}[Q^{(n)}] = \sum_{k=0}^{\infty} k q_k^{(n)} = \frac{dQ(z)}{dz}\bigg|_{z=1} \tag{2.28}$$

$$\text{Var}[Q^{(n)}] = \mathbb{E}[(Q^{(n)})^2] - \mathbb{E}[Q^{(n)}]^2 = \frac{d^2Q(z)}{dz^2}\bigg|_{z=1} + \mathbb{E}[Q^{(n)}] - \mathbb{E}[Q^{(n)}]^2. \tag{2.29}$$

Proposition 5. *∀ n = 1,..,N, given the distribution of $S^{(n)}$ and the expression of $Y(z)$, $\mathbb{E}[Q^{(n)}]$ and Var$[Q^{(n)}]$ can be calculated as:*

$$\mathbb{E}[Q^{(n)}] = \frac{\bar{\bar{S}}^{(n)} + \bar{\bar{Y}}^{(n)} + (\bar{S}^{(n)} - \bar{Y}^{(n)})[1 + 2(\bar{S}^{(n)} - C)] - (\bar{S}^{(n)} - \bar{Y}^{(n)})^2}{2(\bar{S}^{(n)} - \bar{Y}^{(n)})} + \sum_{i=1}^{C-1} \frac{1}{1 - z_i^*} \tag{2.30}$$

$$\begin{aligned} Var[Q^{(n)}] = \frac{1}{12(\bar{S}^{(n)} - \bar{Y}^{(n)})^2}\bigg[& - 4(\bar{\bar{\bar{S}}}^{(n)} - \bar{\bar{\bar{Y}}}^{(n)})(\bar{S}^{(n)} - \bar{Y}^{(n)}) + 3(\bar{\bar{S}}^{(n)} + \bar{\bar{Y}}^{(n)})^2 \\ & - [6(\bar{\bar{S}}^{(n)} - \bar{\bar{Y}}^{(n)}) - 1](\bar{S}^{(n)} - \bar{Y}^{(n)})^2 - (\bar{S}^{(n)} - \bar{Y}^{(n)})^4\bigg] - \sum_{i=1}^{C-1} \frac{z_i^*}{(1 - z_i^*)^2} \end{aligned} \tag{2.31}$$

where $\bar{\bar{S}}^{(n)}$ and $\bar{\bar{\bar{S}}}^{(n)}$ (resp. $\bar{\bar{Y}}^{(n)}$ and $\bar{\bar{\bar{Y}}}^{(n)}$) are the second and third central moments of $S^{(n)}$ (resp. $Y^{(n)}$).

Proof. The derivation follows the same idea in Powell [48]. Details are mathematical tedious and are thus attached in 2.7.3. These results are equivalent to Powell [41] who considered the general bulk-service queue model (but Powell [41] did not provide the detailed proof in the paper). □

Proposition 6. *∀ n = 1,..,N, given the distribution of $S^{(n)}$ and the expression of $Y(z)$, the mean and variance of waiting time at station n (denoted as $W^{(n)}$) is given* *as:*

$$\mathbb{E}[W^{(n)}] = \frac{\bar{Q}_t^{(n)}}{\lambda^{(n)}} \tag{2.32}$$

$$Var[W^{(n)}] = \frac{\bar{\bar{Q}}_t^{(n)} - \bar{Q}_t^{(n)}}{(\lambda^{(n)})^2} \tag{2.33}$$

where $Q_t^{(n)}$ is the queue length at an arbitrary time point (as opposed to $Q^{(n)}$ which is the queue length at the time of vehicle arrival). $\bar{Q}_t^{(n)}$ and $\bar{\bar{Q}}_t^{(n)}$ are defined as

$$\bar{Q}_t^{(n)} = \mathbb{E}[Q^{(n)}] - \bar{Y}^{(n)} + \frac{1}{2}\left(\bar{\bar{Y}}^{(n)}/\bar{Y}^{(n)} + \bar{Y}^{(n)} - 1\right) \tag{2.34}$$

$$\bar{\bar{Q}}_t^{(n)} = Var[Q^{(n)}] - \bar{\bar{Y}}^{(n)} + \frac{1}{12(\bar{Y}^{(n)})^2}\left[4\bar{Y}^{(n)}\bar{\bar{\bar{Y}}}^{(n)} + 6(\bar{Y}^{(n)})^2\bar{\bar{Y}}^{(n)} - (\bar{Y}^{(n)})^2 + (\bar{Y}^{(n)})^4 - 3(\bar{\bar{Y}}^{(n)})^2\right] \tag{2.35}$$

Eq. 2.32 is the application of Little's law. Proposition 6 is directly obtained from Powell [41].

Remark 1. The formulation of $\mathbb{E}[Q^{(n)}]$, $\text{Var}[Q^{(n)}]$, $\mathbb{E}[W^{(n)}]$, and $\text{Var}[W^{(n)}]$ in this study are equivalent to Powell [41] because in his paper the $M/G^{[S]}/1$ bulk queue model was considered, where $G^{[S]}$ represents a general (i.e., arbitrary) bulk-service distribution, which includes the service distribution incorporating random service suspension considered in this study. However, this does not lower the contribution of this study because to implement these equations, the formulation of $Y(z)$ needs to be specified. And in the next section 2.4.3 we show how random service suspension introduces a new distribution for $Y^{(n)}$, which has not been considered in the literature.

Headway distribution

According to Propositions 4 to 6, to calculate $q_{0:C-1}^{(n)}$ and the mean and variance of queue length and waiting time, it is essential to specify $Y(z)$ (i.e., the PGF of the number of passengers arriving within a headway). According to Eq. 2.4, taking $l \to \infty$ gives that $Y^{(n)}|H^{(n)}$ is a Poisson random variable with parameter $\lambda^{(n)}H^{(n)}$. Therefore, we first consider the distribution of $H^{(n)}$ under the random service suspension.

According to the discussion in Section 2.3.3, the actual headway for vehicle l at station n is $H^{(n,l)} = \bar{H} + \frac{2\cdot\mathbb{E}[I^{(N,l)}]}{\bar{F}} + I^{(n,l)} - I^{(n,l-1)}$, where $I^{(n,l)}$ is the total duration of incidents for vehicle l during its travel from the transportation hub to station n. Since $\bar{H}$ and $\mathbb{E}[I^{(n,l)}]$ are constants, obtaining the headway distribution is equivalent to quantifying the distribution of $I^{(n,l)} - I^{(n,l-1)}$.

Notice that $I^{(n,l)}$ and $I^{(n,l-1)}$ are i.i.d for all l by our assumption. It is useful to first consider the distribution of $I^{(n,l)}$.

Proposition 7. *The total incident duration for vehicle l during its travel from the transportation hub to station n (i.e., $I^{(n,l)}$) follows a compound Poisson-Exponential distribution with Poisson rate $\gamma T^{(n)}$ and exponential rate θ. Mathematically,*

$$I^{(n,l)} = \sum_{i=1}^{K} X_i \qquad \textit{where } X_i \sim \boldsymbol{Exp}(\theta) \;\; \forall i = 1, ..., K, \;\; \textit{and } K \sim \boldsymbol{Poi}(\gamma T^{(n)}) \quad (2.36)$$

Proof. When there are no incidents in the system, vehicle l reaches station n after $T^{(n)}$ time units. Since the system can only switch to the incident state from the normal state, the number of incident occurrences, K, follows a Poisson distribution with rate $\gamma T^{(n)}$. The vehicle stopping time for the i-th incident, X_i, follows an exponential distribution with rate θ (i.e., mean $\frac{1}{\theta}$). Therefore, the duration of all incidents is $I^{(n,l)} = \sum_{i=1}^{K} X_i$, where $X_i \sim \mathbf{Exp}(\theta)$ $\forall i = 1, ..., K$, and $K \sim \mathbf{Poi}(\gamma T^{(n)})$ □

The moment generating function (MGF) of a compound Poisson-Exponential variable can be written as [99]

$$M_{I^{(n,l)}}(t) = \mathbb{E}[e^{tI^{(n,l)}}] = e^{\gamma T^{(n)}(\frac{\theta}{\theta - t} - 1)} \qquad \forall\, t < \theta \quad (2.37)$$

Similarly, the MGF of $-I^{(n,l-1)}$ is

$$M_{-I^{(n,l-1)}}(t) = \mathbb{E}[e^{-tI^{(n,l-1)}}] = e^{\gamma T^{(n)}(\frac{\theta}{\theta + t} - 1)} \qquad \forall\, t > -\theta \quad (2.38)$$

From the MGF of $I^{(n,l)}$, we obtain $\mathbb{E}[I^{(N,l)}] = \frac{\gamma T^{(N)}}{\theta}$. Then the headway equation (Eq.

2.12) becomes

$$H^{(n,l)} = \bar{H} + \frac{2\gamma T^{(N)}}{\theta \bar{F}} + I^{(n,l)} - I^{(n,l-1)} \tag{2.39}$$

The following proposition provides the headway distribution:

Proposition 8. *Under the setting of this study, ∀ n = 1,..,N, the MGF of $H^{(n)}$ can be expressed as*

$$M_{H^{(n)}}(t) = e^{t(\bar{H}+\frac{2\gamma T^{(N)}}{\theta \bar{F}})} e^{\gamma T^{(n)}(\frac{2t^2}{\theta^2 - t^2})} \tag{2.40}$$

Proof.

$$\begin{aligned} M_{H^{(n,l)}}(t) &= \mathbb{E}[e^{tH^{(n,l)}}] = \mathbb{E}[e^{t(\bar{H}+\frac{2\gamma T^{(N)}}{\theta \bar{F}}} e^{tI^{(n,l)}} e^{-tI^{(n,l-1)}}] \\ &= e^{t(\bar{H}+\frac{2\gamma T^{(N)}}{\theta \bar{F}}} \mathbb{E}[e^{tI^{(n,l)}}]\mathbb{E}[e^{-tI^{(n,l-1)}}] \\ &= e^{t(\bar{H}+\frac{2\gamma T^{(N)}}{\theta \bar{F}}} e^{\gamma T^{(n)}(\frac{\theta}{\theta - t}-1)} e^{\gamma T^{(n)}(\frac{\theta}{\theta + t}-1)} \\ &= e^{t(\bar{H}+\frac{2\gamma T^{(N)}}{\theta \bar{F}})} e^{\gamma T^{(n)}(\frac{2t^2}{\theta^2 - t^2})} \end{aligned} \tag{2.41}$$

where Eq. 2.41 is because of the independence between $I^{(n,l)}$ and $I^{(n,l-1)}$. As this equation holds for all vehicles l, the MGF of $H^{(n)}$ (i.e., $l \to \infty$) is $M_{H^{(n)}}(t) = M_{H^{(n,l)}}(t)$. □

From the MGF of $H^{(n)}$, we can obtain the corresponding mean and variance of headway as:

$$\mathbb{E}[H^{(n)}] = \bar{H} + \frac{2\gamma T^{(N)}}{\theta \bar{F}} \tag{2.42}$$

$$\text{Var}[H^{(n)}] = \frac{4T^{(n)}\gamma}{\theta^2} \tag{2.43}$$

Remark 2. The results show that random suspensions can increase the mean and variance of headway. The impact on mean headway is through the increase in cycle time at the route planning stage. The headway variance will increase with a higher incident rate (γ) and higher average incident duration ($\frac{1}{\theta}$). Meanwhile, our model also captures the headway variance propagation along stations as observed in many previous studies [100, 75]: $\text{Var}[H^{(n)}]$ increase with the station index n (due to the increase in $T^{(n)}$).

However, the support of the derived headway distribution is $\mathbb{R}$, meaning that $H^{(n)}$ can be negative due to the overtaking of vehicles. The negative value of $H^{(n)}$ will cause problems in the definition of $Y^{(n)}$ (i.e., the number of arrival passengers within a headway). To address this problem, we assume that drivers are not allowed to overtake the preceding vehicles. This is true for the subway systems. Many transit agencies also use this policy for bus operations. Given this assumption, the support of $H^{(n)}$ becomes $[0, +\infty]$. Whenever $H^{(n)} < 0$, the actual headway will be 0 since the successor vehicle will not pass through the predecessor and they will arrive at the station simultaneously (i.e., bus bunching). Hence, the new truncated headway, denoted as $\hat{H}^{(n)}$, has a zero-inflation mixture distribution:

$$\hat{H}^{(n)} = \begin{cases} 0 & \text{if } H^{(n)} \leq 0 \\ H^{(n)} & \text{otherwise} \end{cases} \tag{2.44}$$

The zero-inflation truncated headway distribution is also observed in the previous empirical study assuming no overtaking [77].

However, to the best of the author's knowledge, there is no closed-form MGF for $\hat{H}^{(n)}$ because the difference between two compound Poisson-exponential random variables has no closed-form probability density function. Therefore, to have a tractable headway distribution, we have to approximate $H^{(n)}$ with other distributions for which the corresponding zero-inflation truncated distribution has analytical MGF.

$I^{(n,l)}$ can be seen as the summation of a large number of i.i.d random variables when the incident frequency is high (i.e., K is large, which is true for this study because we are considering high-frequency short random disturbance). Hence, from the Central Limit Theorem (CLT), we may approximate $I^{(n,l)}$ as a normal distribution, which leads to $H^{(n)}$ being a normal distribution as well. Approximating the headway disturbance as a normal random variable with the CLT was also used in Daganzo [101]. In fact, we observe the third central moment of $H^{(n)}$, which is a measure of skewness, is Skewness$[H^{(n)}] = 0$, implying that $H^{(n)}$ is symmetric. Moreover, the MGF of the normal distribution of $H^{(n)}_{\text{Normal}}$ with the same mean and variance is

$$M_{H^{(n)}_{\text{Normal}}}(t) = e^{t(\bar{H}+\frac{2\gamma T^{(N)}}{\theta \bar{F}})} e^{\gamma T^{(n)}(\frac{2t^2}{\theta^2})}, \tag{2.45}$$

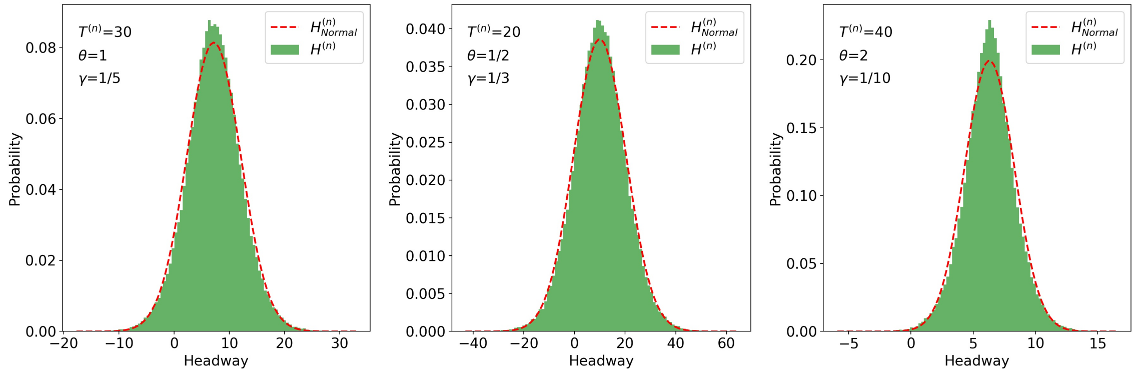
which is very similar to Eq. 2.40 (the MGF of $H^{(n)}$). Therefore, it is reasonable to approximate the distribution of $H^{(n)}$ as a normal distribution with the same mean and variance. Note that the first three moments of $H^{(n)}$ and $H^{(n)}_{\text{Normal}}$ are the same. And the corresponding forth moments (i.e., Kurtosis) are:

$$\text{Kurtosis}[H^{(n)}] = \frac{48(T^{(n)}\gamma)^2 + 48T^{(n)}\gamma}{\theta^4} \tag{2.46}$$

$$\text{Kurtosis}[H^{(n)}_{\text{Normal}}] = \frac{48(T^{(n)}\gamma)^2}{\theta^4} \tag{2.47}$$

which means that the distribution of $H^{(n)}$ may have heavier tails and peakedness compared to $H^{(n)}_{\text{Normal}}$.

Figure 2-6 empirically compares the distribution of $H^{(n)}$ and $H^{(n)}_{\text{Normal}}$ with various values of $T^{(n)}$, θ, and γ. The histogram of $H^{(n)}$ is generated by sampling variables from the associated exponential and Poisson distributions to get the compound distribution. Results show that the normal distribution approximates the original distribution well. As expected, $H^{(n)}$ shows more peakedness than $H^{(n)}_{\text{Normal}}$.



(a) Example 1

(b) Example 2

(c) Example 3

Figure 2-6: Empirical validation for approximating the headway distribution as normal

Next, let us consider a zero-inflation truncated distribution of $H^{(n)}_{\text{Normal}}$ with support $[0, +\infty]$ and a probability mass concentrated at zero. Denote the new random variable as $\hat{H}^{(n)}_{\text{Normal}}$.

Proposition 9. *Under the setting of this study, $\forall\ n = 1,..,N$, the MGF of $\hat{H}^{(n)}_{Normal}$ can be expressed as*

$$\begin{aligned} M_{\hat{H}^{(n)}_{Normal}}(t) =& \Phi\left(\frac{-(\bar{H}\theta + \frac{2\gamma T^{(N)}}{\bar{F}})}{2\sqrt{T^{(n)}\gamma}}\right) + \\ & e^{t(\bar{H}+\frac{2\gamma T^{(N)}}{\theta\bar{F}})} e^{\gamma T^{(n)}(\frac{2t^2}{\theta^2})}\left[1 - \Phi\left(\frac{-(\bar{H}\theta + \frac{2\gamma T^{(N)}}{\bar{F}})}{2\sqrt{T^{(n)}\gamma}} - \frac{2t\sqrt{T^{(n)}\gamma}}{\theta}\right)\right] \end{aligned} \tag{2.48}$$

where $\Phi(\cdot)$ is the cumulative density function (CDF) of a standard normal distribution.

Proof. Let μ and σ^2 be the mean and variance of $H^{(n)}_{\text{Normal}}$, respectively, where $\mu = \bar{H} + \frac{2\gamma T^{(N)}}{\theta\bar{F}}$ and $\sigma = \frac{2\sqrt{T^{(n)}\gamma}}{\theta}$. The MGF of $\hat{H}^{(n)}_{\text{Normal}}$ can be derived as

$$\begin{aligned} M_{\hat{H}^{(n)}_{\text{Normal}}}(t) &= \mathbb{E}[e^{t\hat{H}^{(n)}_{\text{Normal}}}] = \mathbb{P}[H^{(n)}_{\text{Normal}} \leq 0] \cdot e^0 + \int_0^{+\infty} e^{tz} \cdot \phi_{H^{(n)}_{\text{Normal}}}(z) \cdot dz \\ &= \Phi(\frac{-\mu}{\sigma}) + \frac{1}{\sigma\sqrt{2\pi}} \int_0^{+\infty} e^{tz + \frac{(z-\mu)^2}{-2\sigma^2}} dz \\ &= \Phi(\frac{-\mu}{\sigma}) + e^{\mu t + \frac{\sigma^2 t^2}{2}} \left[1 - \Phi(\frac{-\mu}{\sigma} - \sigma t)\right] \end{aligned} \tag{2.49}$$

where Eq. 2.49 follows the same derivation of a truncated normal distribution [102]. Subsisting the value of μ and σ completes the proof. □

Based on the MGF of $\hat{H}^{(n)}_{\text{Normal}}$, notice that $\left[1 - \Phi\left(\frac{-\mu}{\sigma}\right)\right] = \Phi\left(\frac{\mu}{\sigma}\right)$, we can get the corresponding mean and variance as follows.

$$\mathbb{E}[\hat{H}^{(n)}_{\text{Normal}}] = \mu \cdot \Phi\left(\frac{\mu}{\sigma}\right) + \sigma \cdot \phi\left(\frac{-\mu}{\sigma}\right) \tag{2.50}$$

$$\text{Var}[\hat{H}^{(n)}_{\text{Normal}}] = \mu\sigma\phi\left(\frac{-\mu}{\sigma}\right) + \Phi\left(\frac{\mu}{\sigma}\right)\left(\mu^2 + \sigma^2\right)\ - \left(\mu\Phi\left(\frac{\mu}{\sigma}\right) + \phi\left(\frac{-\mu}{\sigma}\right)\sigma\right)^2 \tag{2.51}$$

where $\phi(\cdot)$ is the probability density function (PDF) of a standard normal distribution.

It is not clear how incidents will affect the mean headway from Eq. 2.50 directly. However, the following proposition shows that the mean headway increases as incident frequency (γ) and average incident duration ($\frac{1}{\theta}$) increase.

Proposition 10. *The mean of the zero-inflation truncated headway (i.e, either $\mathbb{E}[\hat{H}^{(n)}]$ or $\mathbb{E}[\hat{H}^{(n)}_{Normal}]$) increases with the increase in incident intensity (i.e., increase in γ or $\frac{1}{\theta}$, or both).*

Proof. The strict mathematical proof can be done by taking derivative of $\mathbb{E}[\hat{H}^{(n)}_{\text{Normal}}]$ in terms of γ or $\frac{1}{\theta}$ and show that it is always positive. However, in this study, we adopt a more intuitive graphical proof, which is easier for understanding.

As shown in Figure 2-7, consider an arbitrary truncated headway distribution (shown in the red line, denoted the headway as $\hat{H}_{\text{Red}}$). When the incident intensity increases, according to Eqs. 2.42 and 2.43, both μ and σ increase. Let us first consider the increase in σ and assume μ does not change (which correspond to the scenario where $\bar{F} \to \infty$). Then the distribution will become the blue curve (denote the corresponding headway as $\hat{H}_{\text{Blue}}$). Note that $\hat{H}_{\text{Red}}$ and $\hat{H}_{\text{Blue}}$ have the same peak value, but since $\hat{H}_{\text{Blue}}$ has longer positive tail, we have $\mathbb{E}[\hat{H}_{\text{Blue}}] > \mathbb{E}[\hat{H}_{\text{Red}}]$. Next, let us also consider the incident's impact on the increase in μ as well. The distribution is shown by the green curve (denoted the headway as $\hat{H}_{\text{Green}}$). Since $\hat{H}_{\text{Blue}}$ and $\hat{H}_{\text{Green}}$ has the same σ, but $\hat{H}_{\text{Green}}$ has higher μ (shifted right), we have $\mathbb{E}[\hat{H}_{\text{Green}}] > \mathbb{E}[\hat{H}_{\text{Blue}}]$. Hence, $\mathbb{E}[\hat{H}_{\text{Green}}] > \mathbb{E}[\hat{H}_{\text{Red}}]$, showing that the increase in incident intensity will increase μ and σ, thus increase the mean of the truncated headway. □

Proposition 10 is useful for the analysis of system stability with respect to incidents, which is shown in Section 2.4.4.

Distribution of $Y^{(n)}$

The distribution of $Y^{(n)}$ is derived by assuming the headway is $\hat{H}^{(n)}_{\text{Normal}}$ (instead of $H^{(n)}$, which may be negative). To derive the PGF of $Y^{(n)}$, the following lemma is introduced.

Lemma 1. For two arbitrary random variable U and V, assume that

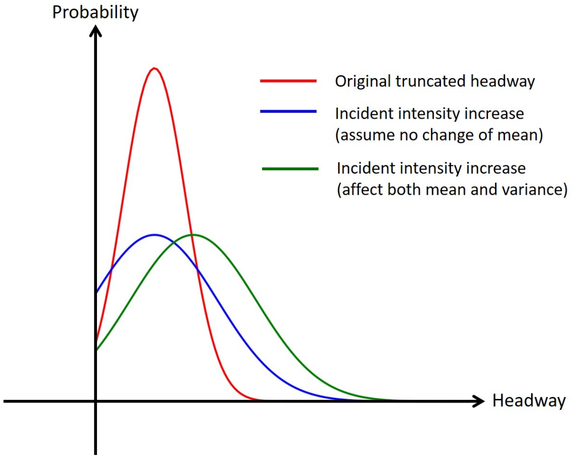


Figure 2-7: Illustration for the impact of incidents on expected headway. As the probability mass at zero does not contribute to the expectation calculation, it is not shown in the figure.

- there is a $\delta > 0$ such that for t in $(-\delta, \delta)$, the MGF of $U|V$ is $M_{U|V}(t) = C_1(t)e^{C_2(t)V}$, where $C_1(t)$ and $C_2(t)$ are finite functions of t that do not depend on V,
- and the MGF of V, $M_V(\cdot)$, exists and $M_V[C_2(t)]$ is finite for t in $(-\delta, \delta)$.

Then the MGF of U is given by

$$M_U(t) = C_1(t)M_V[C_2(t)], \qquad -\delta < t < \delta. \tag{2.52}$$

Proof. The proof of Lemma 1 can be found in Villa and Escobar [103] Result 1. □

Proposition 11. *Under the setting of this study, $\forall$ n = 1,..,N, the PGF of $Y^{(n)}$, $Y(z)$, can be expressed as*

$$Y(z) = \Phi\left(\frac{-\mu}{\sigma}\right) + e^{\mu\lambda^{(n)}(z-1)+\frac{\sigma^2(\lambda^{(n)}z-\lambda^{(n)})^2}{2}}\left[1 - \Phi\left(\frac{-\mu}{\sigma} - \sigma\lambda^{(n)}(z-1)\right)\right] \tag{2.53}$$

where $\mu = \bar{H} + \frac{2\gamma T^{(N)}}{\theta\bar{F}}$ and $\sigma = \frac{2\sqrt{T^{(n)}\gamma}}{\theta}$ are the mean and standard deviation of $H^{(n)}_{Normal}$, respectively.

Proof. Recall that $Y^{(n)}|\hat{H}^{(n)}_{\text{Normal}}$ is a Poisson random variable with parameter $\lambda^{(n)}\hat{H}^{(n)}_{\text{Normal}}$. So, the MGF of $Y^{(n)}|\hat{H}^{(n)}_{\text{Normal}}$ is $M_{Y^{(n)}|\hat{H}^{(n)}_{\text{Normal}}}(t) = \exp[\lambda^{(n)}\hat{H}^{(n)}_{\text{Normal}}(e^t - 1)]$. Based on

Lemma 1, setting $C_1(t) = 1$ and $C_2(t) = \lambda^{(n)}(e^t - 1)$, we conclude that the MGF of $Y^{(n)}$ is

$$M_{Y^{(n)}}(t) = \Phi\left(\frac{-\mu}{\sigma}\right) + e^{\mu\lambda^{(n)}(e^t-1)+\frac{\sigma^2(\lambda^{(n)}e^t-\lambda^{(n)})^2}{2}}\left[1 - \Phi\left(\frac{-\mu}{\sigma} - \sigma\lambda^{(n)}(e^t - 1)\right)\right] \tag{2.54}$$

Substituting $t = \log z$ in Eq. 2.54 completes the proof.

□

From Eq. 2.54, we can obtain $\bar{Y}^{(n)}$, $\bar{\bar{Y}}^{(n)}$, and $\bar{\bar{\bar{Y}}}^{(n)}$ by taking corresponding derivatives. The expression of $\bar{Y}^{(n)}$ is shown below. The expressions for $\bar{\bar{Y}}^{(n)}$ and $\bar{\bar{\bar{Y}}}^{(n)}$ are complicated and thus omitted.

$$\bar{Y}^{(n)} = \left(\mu \cdot \Phi\left(\frac{\mu}{\sigma}\right) + \sigma \cdot \phi\left(\frac{-\mu}{\sigma}\right)\right) \cdot \lambda^{(n)} \tag{2.55}$$

With the expression of $Y(z)$, the $z_0^*, ..., z_{C-1}^*$ can be obtained by solving the nonlinear equation $\textsc{Den}(z) = 0$ (see Section 2.4.3 for details). Then $q_{0:C-1}^{(n)}$ and other resilience indicators at station n can be obtained accordingly.

Solving for the roots

It is well known in the queuing literature that solving for the roots of $\textsc{Den}(z)$ is practically difficult because typical optimization algorithms usually only find only one root, while we need to find all C roots within the unit circle. This is especially changeling for $Y(z)$ with complex expressions because the objective function can be highly nonlinear (such as $Y(z)$ in this study). We propose an interpolation search algorithm to efficiently find all roots of $\textsc{Den}(z)$ within the unit circle.

Notice that $\textsc{Den}(z) = 0$ is equivalent to find $z_0^*, ..., z_{C-1}^*$, such that

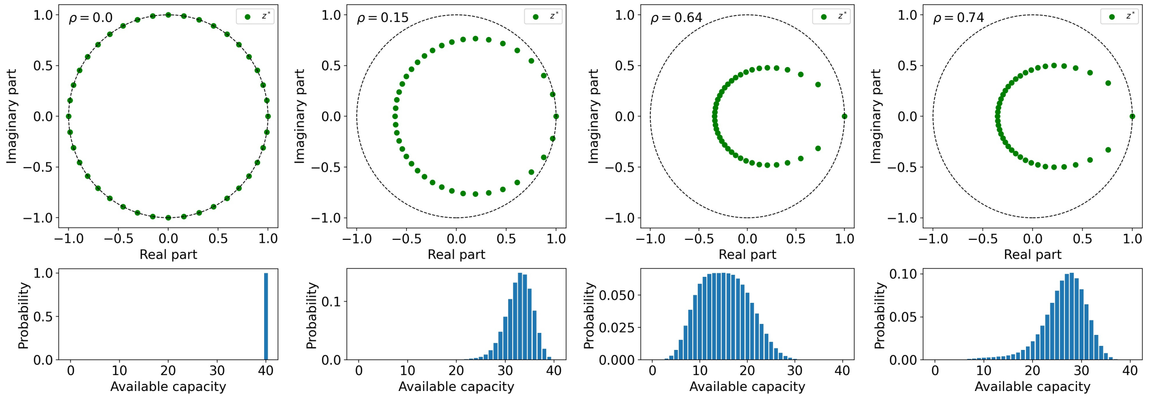
$$\frac{1}{Y(z_k^*)} - S(1/z_k^*) = 0 \Leftrightarrow J(z_k^*) = 1 \qquad \forall\, k = 0, ..., C-1 \tag{2.56}$$

where $J(z) := Y(z)S(1/z)$. Taking the logarithm of both sides of Eq. 2.56 and matching the real and imaginary parts gives:

$$\begin{cases} \mathrm{Re}[\log(J(z))] = 0 \\ \mathrm{Im}[\log(J(z))] = 0 \end{cases} \tag{2.57}$$

where Re[·] and Im[·] represent the real and imaginary part of a complex number. Eq. 2.57 can be solved efficiently with many optimization algorithms (such as trust-region and Levenberg-Marquardt algorithms). However, as there are C optimal solutions for this problem with $|z^*| \leq 1$, the challenge is how to select different initial values so as to find all solutions.

It can be empirically observed that the distribution of the C solutions has an oval-like shape. Figure 2-8 shows some examples of the solution distribution with different values of $\rho^{(n)}$ (where $\rho^{(n)} = \bar{Y}^{(n)}/\bar{S}^{(n)}$ is the utilization ratio of a bulk service queuing system) and $s^{(n)}$. It is found that the closer $\rho^{(n)}$ is to 1 (resp. 0), the closer the shape of the root distribution is to an ellipse (resp. circle). The value of $s^{(n)}$ (i.e., available capacity distribution) can also slightly affect the root distribution.



(a) $\rho^{(n)} = 0$, $C = 40$ (b) $\rho^{(n)} = 0.15$, $C = 40$ (c) $\rho^{(n)} = 0.64$, $C = 40$ (d) $\rho^{(n)} = 0.74$, $C = 40$

Figure 2-8: Examples of root distribution

We first express the complex number in polar coordinate system with $z = r\exp[\varphi i]$, where $i = \sqrt{-1}$, r is the length from z to the origin, and φ is the angle. Eq. 2.57 now has C optimal solutions (r_k^*, φ_k^*) for $k = 0, 1, ..., C-1$, where $0 \leq r_k^* \leq 1$ and $0 \leq \varphi_k^* < 2\pi$. Note that $z_0^* = 1$ corresponds to $r_0^* = 1$ and $\varphi_0^* = 0$. Another property is that the roots must appear as conjugate pairs. Hence, if (r^*, φ^*) is a root and $0 < \varphi^* < \pi$, then $(r^*, 2\pi - \varphi^*)$ is also a root.

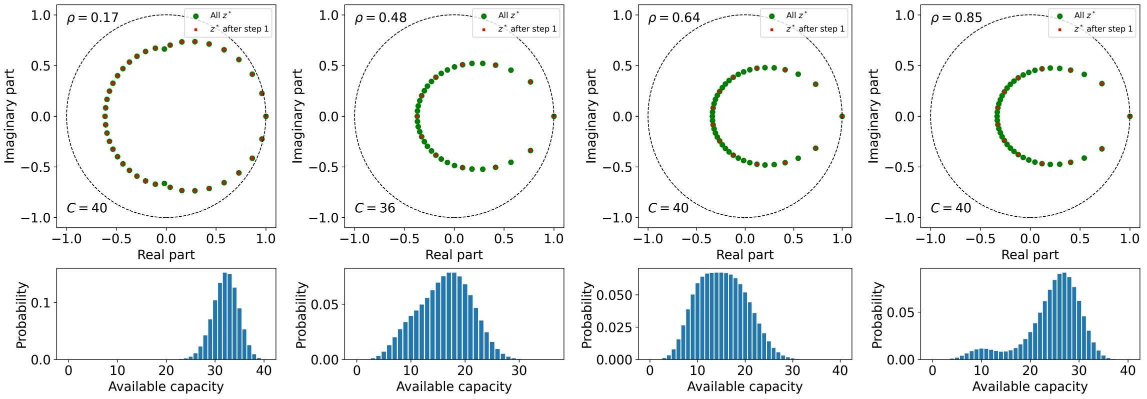
The proposed search algorithm has two steps. The first step is referred to as "clockwise searching", which is adapted from the numerical method in Powell [41]. The empirical observation (Figure 2-8) shows a rough relationship that $r^*_{k+1} - r^*_k \approx r^*_k - r^*_{k-1}$, especially for small $\rho^{(n)}$. This is equivalent to

$$r^*_{k+1} \approx 2r^*_k - r^*_{k-1} \tag{2.58}$$

Eq. 2.58 provides a way to determine the initial value for solving for the $k+1$-th root when the k-th and $k-1$-th roots are available. As we already know $r^*_0 = 1$ and $\varphi^*_0 = 0$, we first set the initial value for solving for the second root as $r^{\text{Ini}}_1 = 1 - 0.5\rho^{(n)}$ and $\varphi^{\text{Ini}}_1 = 3\pi/C$. This is motivated by the shape of the root distribution with respect to $\rho^{(n)}$. Then r^{Ini}_1 and φ^{Ini}_1 are used as the initial value to solve for r^*_1 and φ^*_1 based on Eq 2.57. For $k \geq 2$, the initial values for solving the for k-th root are set to $r^{\text{Ini}}_k = r^*_{k-1} + (r^*_{k-1} - r^*_{k-2})$, $\varphi^{\text{Ini}}_k = \varphi^*_{k-1} + (\varphi^*_{k-1} - \varphi^*_{k-2})$ according to Eq. 2.58.

However, only performing step 1 (i.e., Powell [41]'s method) may not find all C distinct roots. Figure 2-9 shows some examples of the comparison between roots found in step 1 and all roots. We observe that when $\rho^{(n)}$ is relatively large (i.e., the system is relatively congested), the clockwise search does not perform well because the approximate relationship in Eq. 2.58 does not hold. Even when $\rho^{(n)}$ is relatively small, it is also possible that some roots do not perfectly fit the oval-like shape (such as Figure 2-9a), resulting in the failure of step 1 to find all roots.

Therefore, we propose a second step called "interpolation search". Let the set of found roots from step 1 be $\mathcal{Z}^{(0)} = \{(r^{(0)}_0, \varphi^{(0)}_0), (r^{(0)}_1, \varphi^{(0)}_1), ..., (r^{(0)}_{M_0}, \varphi^{(0)}_{M_0})\}$, where $M_0 = |\mathcal{Z}^{(0)}|$ is the number of roots from step 1. Without loss of generality, assume that the elements in $\mathcal{Z}^{(0)}$ are clockwise ranked (i.e., $\varphi^{(0)}_0 < \varphi^{(0)}_1 < ... < \varphi^{(0)}_{M_0}$). The interpolation search is described in Algorithm 1. The main idea is to perform interpolation between any two adjacent roots that are already found. The interpolated points are set as initial values and fed into Eq. 2.57 to solve for new distinct roots. Then we update the



(a) $\rho^{(n)} = 0.17$, $C = 40$ (b) $\rho^{(n)} = 0.48$, $C = 36$ (c) $\rho^{(n)} = 0.64$, $C = 40$ (d) $\rho^{(n)} = 0.85$, $C = 40$

Figure 2-9: Comparison between roots found with clockwise search and all roots

set of roots with the new distinct roots or perform a finer (i.e., larger L) interpolation if no distinct roots are found. This process is repeated until there are C distinct roots found. In Algorithm 1, L is a parameter controlling how many points to interpolate between two known roots, and ϵ is a predetermined probability threshold to add randomness in the search process.

From the results of numerical testing, our algorithm allows us to find desired roots for all testing scenarios (Section 2.5.1). The methods in Powell [41] (i.e., only step 1) and Wilson [104] (which is used in Islam et al. [76]) fail to.

2.4.4 Stability condition

For all the derivations above, we assume that the steady-state distributions of all variables exist. This triggers the discussion about the stability condition, which is also an important indicator of the system's resilience. At the station level, the stability condition is described in Proposition 12.

Proposition 12. *Under the setting of this study, the bulk-service queuing system at station n is stable if and only if*

$$\rho^{(n)} = \frac{\bar{Y}^{(n)}}{\bar{S}^{(n)}} = \frac{\left(\mu \cdot \Phi\left(\frac{\mu}{\sigma}\right) + \sigma \cdot \phi\left(\frac{-\mu}{\sigma}\right)\right) \cdot \lambda^{(n)}}{\sum_{u=0}^{C} s_u^{(n)} u} = \frac{\lambda^{(n)} \cdot \mathbb{E}[\hat{H}^{(n)}_{Normal}]}{\sum_{u=0}^{C} s_u^{(n)} u} < 1 \tag{2.59}$$

Algorithm 1 Interpolation searching

```
Initialize $\mathcal{Z}^{(0)}$, $M_0$, $\epsilon$. Initialize $L = 2$, $k = 0$.
while $M_k < C$ do
    Initialize $\mathcal{Z}^{\text{Ini}}$ as an empty set.
    for $i = 1 : M_k$ do
        for $d = 1 : L - 1$ do
            $r^{\text{Ini}} = r_i^{(k)} + d \cdot \frac{r_{i+1}^{(k)} - r_i^{(k)}}{L}$; $\varphi^{\text{Ini}} = \varphi_i^{(k)} + d \cdot \frac{\varphi_{i+1}^{(k)} - \varphi_i^{(k)}}{L}$
            Draw a random value $w$ uniformly from $[0, 1)$
            if $w < \epsilon$ then
                Draw a random value $\delta_1$ uniformly from $[-\frac{|r_{i+1}^{(k)} - r_i^{(k)}|}{2L}, \frac{|r_{i+1}^{(k)} - r_i^{(k)}|}{2L}]$
                $r^{\text{Ini}} = r^{\text{Ini}} + \delta_1$
                Draw a random value $\delta_2$ uniformly from $[-\frac{|\varphi_{i+1}^{(k)} - \varphi_i^{(k)}|}{2L}, \frac{|\varphi_{i+1}^{(k)} - \varphi_i^{(k)}|}{2L}]$
                $\varphi^{\text{Ini}} = \varphi^{\text{Ini}} + \delta_2$
            Add $(r^{\text{Ini}}, \varphi^{\text{Ini}})$ into $\mathcal{Z}^{\text{Ini}}$.
    Initialize $\mathcal{Z}^{\text{temp}}$ as an empty set.
    for all $z^{\text{Ini}}$ in $\mathcal{Z}^{\text{Ini}}$ do
        Solve Eq. 2.57 using $z^{\text{Ini}}$ as the initial value, obtaining $z^*_{\text{temp}}$. Let its conjugate be $\bar{z}^*_{\text{temp}}$.
        If $z^*_{\text{temp}}$ $(\bar{z}^*_{\text{temp}})$ not in $\mathcal{Z}^{(k)}$, add it to $\mathcal{Z}^{\text{temp}}$, otherwise do nothing.
    $\mathcal{Z}^{(k+1)} = \mathcal{Z}^{(k)} \cup \mathcal{Z}^{\text{temp}}$ and rank all elements in $\mathcal{Z}^{(k+1)}$ clockwise
    Denote $\mathcal{Z}^{(k+1)}$ as $\{(r_0^{(k+1)}, \varphi_0^{(k+1)}), ..., (r_{M_{k+1}}^{(k+1)}, \varphi_{M_{k+1}}^{(k+1)})\}$
    $k = k + 1$
    if $M_{k+1} = M_k$ then
        $L = L + 1$
```

where $\rho^{(n)}$ is the utilization ratio for station n.

Proof. The stability condition is equivalent to $\mathbb{P}(Q^{(n)} = 0) = q_0^{(n)} > 0$. In Eq. 2.25, we notice that $\prod_{i=1}^{C-1} \frac{z_i^*}{z_i^* - 1}$ is always greater than 0 (see 2.7.2 for details), and $s_C^{(n)} > 0$ is a known condition. Therefore, $q_0^{(n)} > 0$ if and only if $\bar{Y}^{(n)} < \bar{S}^{(n)}$ (i.e., $\rho^{(n)} < 1$), which completes the proof. □

Proposition 12 is intuitive as it indicates that station n is stable if the average number of passengers arrived within a headway is smaller than the average available capacity for each arrival vehicle (after alighting). From Proposition 7, we know that a higher rate of incidents (i.e., larger γ) and higher duration of incidents (i.e., higher $\frac{1}{\theta}$) increase $\mathbb{E}[\hat{H}^{(n)}_{\text{Normal}}]$, which makes the system more likely to be unstable. There are some remarks for Proposition 12.

Remark 3. As $\rho^{(n)}$ depends on $s^{(n)}$ and $s^{(n)}$ depends on the roots (i.e., $z_0^*, ..., z_{C-1}^*$) at station n, there is no direct way to judge the stability at station n without iterating the previous $n-1$ stations. But for the first station ($n=1$), we have $s_C^{(1)} = 1$ and $s_u^{(1)} = 0$ for all $u = 0, ..., C-1$. Then Eq. 2.59 reduces to $\rho^{(1)} = \frac{\lambda^{(n)} \cdot \mathbb{E}[\hat{H}_{\text{Normal}}^{(n)}]}{C}$, which can be used to assess the stability directly.

Remark 4. Proposition 12 only discusses the stability at the station level. At the route level, a route is considered stable if "all stations in the route are stable". Mathematically, a route is stable if and only if $\rho^{(n)} < 1, \forall\, n = 1, 2, ..., N$.

Remark 5. It is worth discussing the relationship of stability of stations n and $n-1$. If station $n-1$ is stable, then $s^{(n)}$ can be calculated as described in Section 2.4.1, and the stability of station n can be evaluated accordingly. However, if station $n-1$ is not stable, station n may be stable because there may be passengers alighting at station n. For this situation, we have $v_C^{(n-1)} = 1$ and $v_k^{(n-1)} = 0$ for all $k = 0, 1, ..., C-1$. Then $s^{(n)}$ is determined by the alighting rate at station n. It is easy to verify that in this situation $\bar{S}^{(n)} = \alpha^{(n)} C$. And the stability condition is $\rho^{(n)} = \frac{\lambda^{(n)} \cdot \mathbb{E}[\hat{H}_{\text{Normal}}^{(n)}]}{\alpha^{(n)} C} < 1$.

2.4.5 Summary of calculation procedure

So far, we have derived the calculation process for all variables of interest. Algorithm 2 summarizes the calculation procedure, which iterates through the N stations of the route. This is more efficient and provides more analytical insights than a simulation model.

2.5 Numerical example

2.5.1 Experimental design

To test the proposed framework, we use an example bus route adapted from Islam et al. [76] and Hickman [75]. There are 10 stations and the attributes for each station are shown in Table 2.1. The layout of the bus route is shown in Figure 2-10, where

Algorithm 2 Resilience indicators calculation procedure

```
Initialize $v_0^{(0)} = 1$ and $v_k^{(0)} = 0 \quad \forall k = 1, ...C$.
for $n = 1 : N$ do
    $g^{(n)} = v^{(n-1)} A^{(n)}$                                        ▷ Eq. 2.14
    $s_k^{(n)} = 1 - g_{C-k}^{(n)} \quad \forall k = 0, 1, ..., C$                ▷ Eq. 2.13
    Calculate $\bar{S}^{(n)}, \bar{\bar{S}}^{(n)}$, and $\bar{\bar{\bar{S}}}^{(n)}$ based on $s^{(n)}$.
    Calculate $\bar{Y}^{(n)}, \bar{\bar{Y}}^{(n)}$, and $\bar{\bar{\bar{Y}}}^{(n)}$                       ▷ Section 2.4.3
    if $\bar{Y}^{(n)} < \bar{S}^{(n)}$ then                                 ▷ Station $n$ is stable
        Solve the roots $z_0^*, ..., z_{C-1}^*$ for the denominator of $Q(z)$ in Eq. 2.23   ▷ Section 2.4.3
        Calculate $q_0^{(n)}, ..., q_{C-1}^{(n)}$ based on $z_0^*, ..., z_{C-1}^*$          ▷ Section 2.4.3
        Calculate $\mathbb{E}[Q^{(n)}], \mathrm{Var}[Q^{(n)}], \mathbb{E}[W^{(n)}]$, and $\mathrm{Var}[W^{(n)}]$   ▷ Eq. 2.30 - 2.33
        $v^{(n)} = g^{(n)} B^{(n)}$                  ▷ Eq. 2.18. $B^{(n)}$ is a function of $q_{0:C-1}^{(n)}$
    else                                                       ▷ Station $n$ is not stable
        $q_k^{(n)} = 0 \quad \forall k = 0, 1, ..., C-1$
        Set $\mathbb{E}[Q^{(n)}], \mathrm{Var}[Q^{(n)}], \mathbb{E}[W^{(n)}]$, and $\mathrm{Var}[W^{(n)}]$ to infinity
        $v_C^{(n)} = 1$ and $v_k^{(n)} = 0 \quad \forall k = 0, 1, ..., C-1$
```

we assume the no-incident travel time between two consecutive stations is 5 minutes, the total cycle time without incident is $\bar{E} = 100$ min, and travel time from the transportation hub to the last station is $T^{(N)} = 50$ minutes.

Table 2.1: Example bus system parameters

Station ID	$\lambda^{(n)}$ (passengers/min)	$\alpha^{(n)}$	Station ID	$\lambda^{(n)}$ (passengers/min)	$\alpha^{(n)}$
1	0.75	0	6	1	0.8
2	1.5	0	7	0.75	0.5
3	0.75	0.1	8	0.5	0.1
4	3	0.25	9	0.2	0.75
5	1.5	0.25	10	0	1

To test the sensitivity of resilience indicators to different parameters, we consider different values of C, θ, γ, $\bar{H}$, and demand (Table 2.2). The demand is adjusted by a scaling factor that is applied to the arrival rates $\lambda^{(n)}$ in Table 2.1. The fleet size $\bar{F}$ is determined as $\frac{\bar{E}}{\bar{H}}$. When the sensitivity testing is conducted for one parameter (e.g., C), other parameters (e.g., θ, γ, $\bar{H}$, and the demand factor) are set to their reference values for comparison.

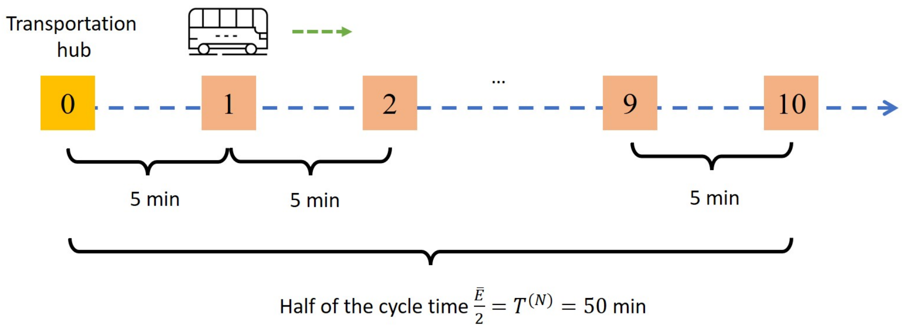

Transportation hub
0
1
2
...
9
10
5 min
5 min
5 min
Half of the cycle time $\frac{\bar{E}}{2} = T^{(N)} = 50$ min

Figure 2-10: Case study route layout

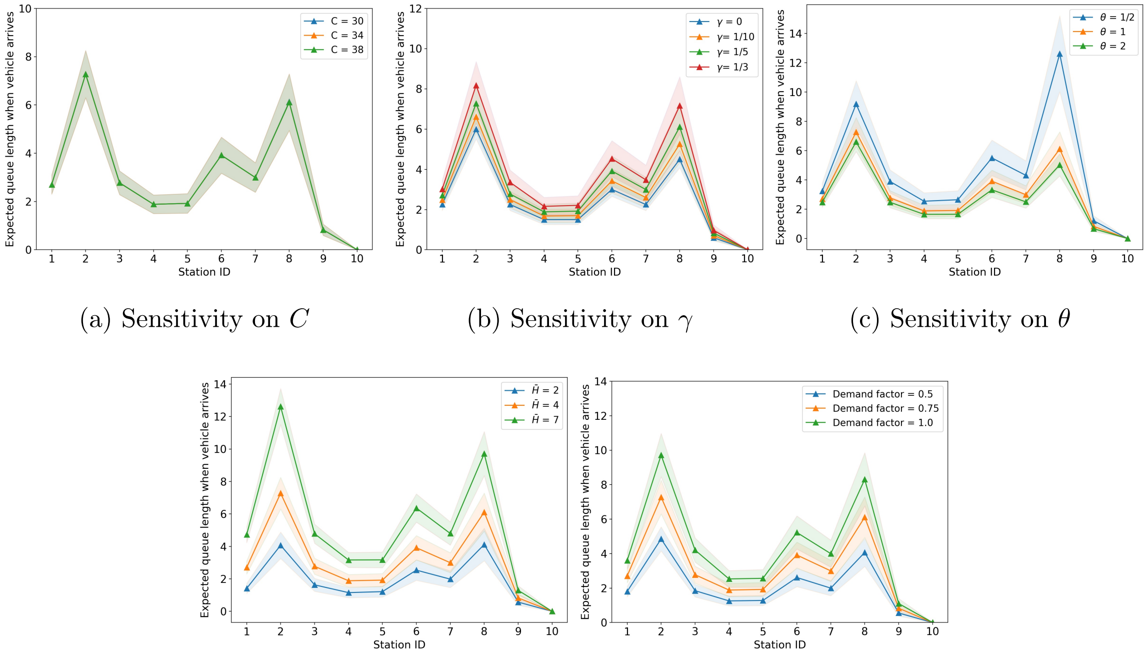
Table 2.2: Scenario design

Parameters	Value space	Reference value
C	{30, 34, 38}	34
γ (/min)	{0, 1/10, 1/5, 1/3}	1/5
θ (/min)	{2, 1 ,1/2}	1
$\bar{H}$ (min)	{2, 4, 7}	6
Demand factor	{0.2, 0.4, 0.6, 0.8, 1}	0.8

2.5.2 Resilience indicators

The mean and standard deviation of queue length for each station under different testing scenarios are shown in Figure 2-11. Generally, for all scenarios, the queue length patterns are consistent with the congestion patterns we expect given the passenger arrival and alighting rates. That is, the expected queue length is relatively higher at stations 2 and 8. The expected queue length at the last station is always zero as its passenger arrival rate is 0.

Figure 2-11a shows the queue length patterns with respect to bus capacity. The system is not very sensitive to bus capacity. The reason is that under the reference scenario, the system is not congested and capacity is not fully utilized. Thus, increasing capacity does not affect the queuing distribution. Figure 2-11b shows the impact of incident occurrence rate γ on queue length. When there is no random suspension in the system ($\gamma = 0$), the expected queue length at station 8 is 4.5. As the frequency of incidents increases, the system becomes more congested with longer expected queue length and higher variance. When the incident frequency increases to 1/3 per minute on average ($\gamma = 1/3$), the expected queue length at station 8 is increased to 8.3 units. Similar results can be observed for the duration of incidents (Figure 2-11c). When the average incident duration is 30 seconds ($\theta = 2$), $\mathbb{E}(Q^{(8)}) = 5.0$. When the average incident duration is 2 minutes ($\theta = 1/2$), $\mathbb{E}(Q^{(8)})$ increases to 12.6. The impacts of θ and γ on queue length are both more significant at crowded stations. The impact of $\bar{H}$ is shown in Figure 2-11d. As expected, higher headway means a lower service rate and thus a higher expected queue length. As $\bar{H}$ increases from 2 minutes to 7 minutes, the queue length at station 8 increases from 4.1 to 9.7. The impact of the demand factor (Figure 2-11e) shows the similar patterns. As the demand factor increases from 0.5 to 1.0, the queue length at station 8 increases from 4.1 to 8.3. The impact of $\bar{H}$ and the demand factor are relatively similar for crowded and uncrowded stations.



(a) Sensitivity on C

(b) Sensitivity on γ

(c) Sensitivity on θ

(d) Sensitivity on $\bar{H}$

(e) Sensitivity on demand factor

Figure 2-11: Mean and standard deviation of queue length (the shaded part is 0.2×standard deviation)

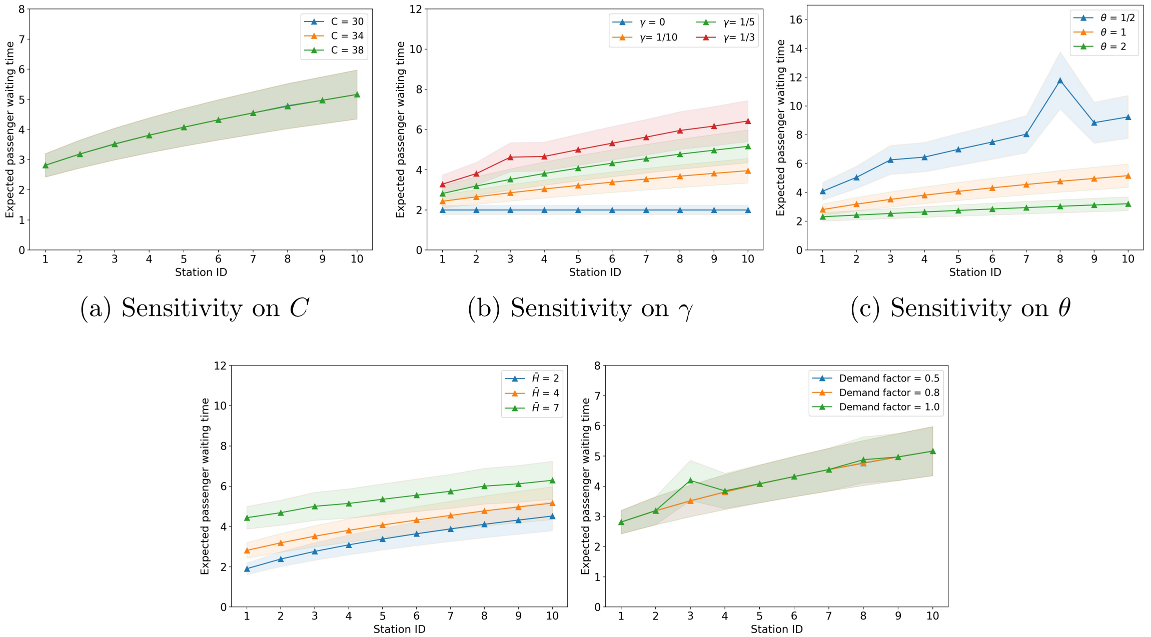
Figure 2-12 shows the mean and standard deviation of passenger waiting time for the different scenarios. We observe that the downstream stations generally have higher waiting time expectations and variances due to the headway variance propagation. For congested stations, such as stations 3 and 8, extra waiting times are observed due to passengers left behind with capacity constraints.

Figure 2-12a shows the impact of capacity on waiting time. Similar to the results on queue length, the impact is not very significant. The impacts of γ and θ on waiting times are shown in Figure 2-12b and 2-12c, respectively. As increases in γ and $1/\theta$ result in an increase in expected headway, the mean waiting times at all stations are increased. The impacts on crowding stations are more significant. When $\gamma = 0$, there is no incident in the system. In this case, there are no left behind or headway irregularity at any stations and their expected waiting times are all equal to 2 minutes (i.e., $\frac{1}{2}\bar{H}$, as no incidents means all stations have the same fixed headway). When γ increases to 1/5, station 3 has left behind passengers and the waiting time is increased to 4.6 minutes. When θ decreases (i.e., mean incident duration increases) from 2 to 1/2, the expected waiting time at station 8 increases from 3.0 to 11.8 minutes. Changes in $\bar{H}$ have the most direct impact on the expected waiting time. The increase in planned headway causes an increase in waiting time for all stations. There are a few left behind passengers observed at stations 3 and 8 when $\bar{H} = 7$ min. Finally, as demand increases, the waiting time increases only if there are left behind (e.g., when demand factor = 1) because it does not change the headway distribution. At station 3, the increase in the demand factor from 0.5 to 1.0 results in an increase in the expected waiting time from 3.5 to 4.2 minutes.

2.5.3 Comparison between simulated and theoretical results

Simulation model

To validate the theoretical results, we develop a simulation model to calculate the expectation and variance of queue length and waiting time. The simulation procedure is shown in Algorithm 3. For each vehicle l at each station n, we generate the total



(a) Sensitivity on C (b) Sensitivity on γ (c) Sensitivity on θ

(d) Sensitivity on $\bar{H}$ (e) Sensitivity on demand factor

Figure 2-12: Mean and standard deviation of waiting time (the shaded part is 0.2×standard deviation)

duration of incidents $I^{(n,l)}$ as a compound Poisson exponential variable to get the arrival time. Since no overtaking is allowed, the arrival time at station n cannot be earlier than vehicle $l-1$. When a vehicle arrives at a station, passengers board based on the first-come-first-serve (FCFS) principle up to the vehicle's capacity C. Queue lengths at vehicle arrival and passenger waiting times are recorded during the simulation. To ensure the system reaches steady-state conditions, the first 10% records are dropped.

Results

We compare the simulation and theoretical results for the reference parameter setting (Table 2.2). A total of $L = 50,000$ vehicle runs are simulated. The comparisons of mean and standard deviation for queue length and waiting time are shown in Figure 2-13. We observe that the simulation and theoretical results match well, validating the theoretical model's correctness. However, the theoretical results slightly overestimate the mean and variance of the queue length and waiting time. The main reason may

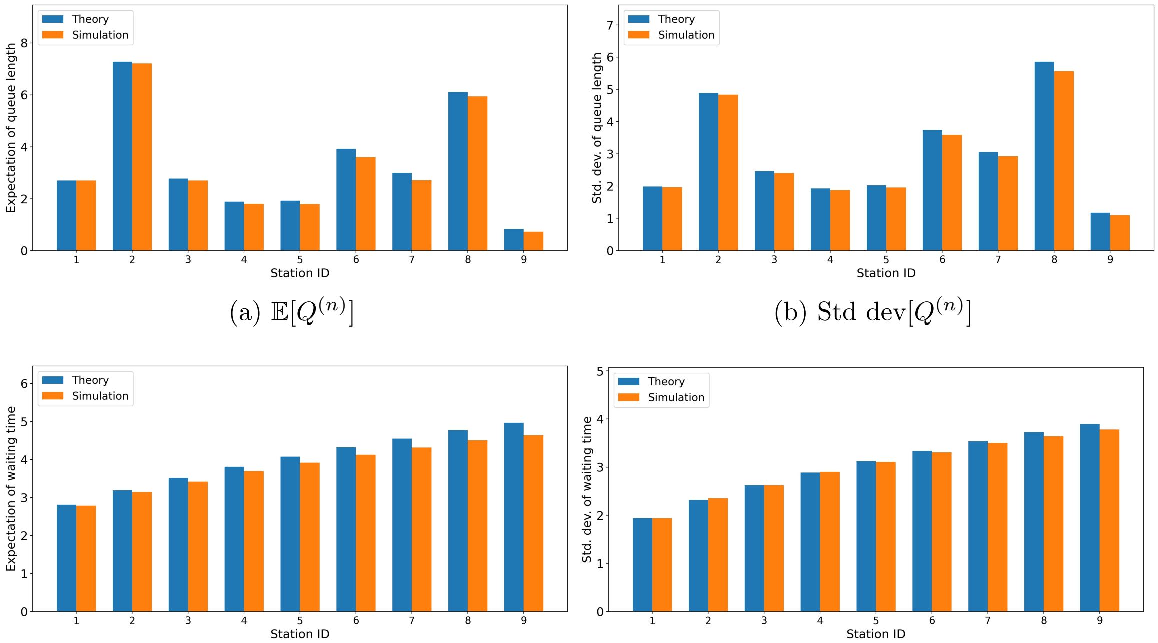
Algorithm 3 Simulation procedure

```
Initialize model parameters: $C$, $\gamma$, $\theta$, $\bar{H}$, Demand factor. Set the total number of vehicles $L$.
for $l = 1{:}L$ do
    Get vehicle dispatch time as $DT^{(l)}$
    for $n = 1 : N$ do
        Sample total incident duration $I^{(n,l)}$ from a compound Poisson exponential distribution
        $t_D^{(n,l)} = \min\{DT^{(l)} + T^{(n)} + I^{(n,l)}, t_D^{(n,l-1)}\}$
        Headway for vehicle $l$ at station $n$ is $t_D^{(n,l)} - t_D^{(n,l-1)}$
        Sample the arrival passengers within the headway as a Poisson process based on $\lambda^{(n)}$.
        Record queue length (including left behind passengers from the last run)
        Alight passengers based on the binomial distribution with parameter $\alpha^{(n)}$
        Board passengers based on FCFS principle up to the vehicle capacity, record left behind passengers
        Record boarding passengers' waiting time
Drop the first 10% records. Calculate $\mathbb{E}[Q^{(n)}], \mathrm{Var}[Q^{(n)}], \mathbb{E}[W^{(n)}]$, and $\mathrm{Var}[W^{(n)}]$ based on the recorded samples for $n = 1, ..., N$
```

be the approximation of headway distribution as normal. As shown in Figure 2-6, the actual headway has more probability density concentrated at the mean (i.e., more peakedness), implying that the actual headway has less probability of deviating from the planned one, thus the simulation scenario may have a smaller queue length and waiting time.

2.6 Conclusion and discussion

This study proposes a stochastic framework to model the resilience of public transit systems under short random service suspensions. Specifically, we analyze the system stability conditions and derive closed-form formulations for the mean and variance of queue length and waiting time at each station. The derived stability conditions are intuitive and imply that the system is more likely to be unstable with high incident rates, high incident duration, high demand, low service frequency, and low vehicle capacity. The proposed model is implemented using an example bus network adapted from the literature. A sensitivity analysis of different parameters (such as incident



(a) $\mathbb{E}[Q^{(n)}]$

(b) Std dev$[Q^{(n)}]$

(c) $\mathbb{E}[W^{(n)}]$

(d) Std dev$[W^{(n)}]$

Figure 2-13: Comparison between simulation and theoretical results (reference scenario)

rate, incident duration, vehicle capacity, etc.) was conducted. The results show that the congested stations (i.e., stations with high demand rates) are more vulnerable to random service suspensions. The results are validated with a simulation model, showing consistency between theoretical and simulation outcomes.

The proposed model has several potential applications. 1) It can facilitate the design and planning of public transit systems with the consideration of random system interruptions, such as the design of headways and determination of vehicle capacity. Moreover, the estimated queue length can be used to evaluate the layout and capacity of congested stations. 2) The model can be used to monitor system performance and identify critical stations by inputting the historical demand and incident information. 3) The model can support efficient cost-benefit analysis of approaches to improve services using estimates of waiting time and queue length. For example, the model can answer that, to control the waiting time within a threshold, what is the most cost-effective way (e.g., increase vehicle size, decrease headway, or increase maintenance frequency to reduce the random suspension rate). In summary, the efficient calculation of the system's resilience indicators can be used in public transit planning, operations, and management applications.

Future studies can address a number of aspects. First, the model can be extended from route-level to network-level. The main difference between route level and network level is the consideration of transfer passengers. A straightforward way is to incorporate the transfer demand as part of the arrival demand. But additional transfer-related parameters (which should be connected with the alighting rate) need to be specified in the model. Second, like many previous random service disruption papers (Section 2.2.2), the model can be extended to consider partial interruptions (as opposed to fully stopped as assumed in this study). With partial interruptions, vehicles can still have positive speed at the failure state. The headway distribution assumption needs to be revised. Third, as mentioned before, this study assumes no balking and reneging behavior of passengers. Future studies may extend the model by considering a more complicated passenger-side behavior.

2.7 Appendix

2.7.1 Derivation of probability generating function of $Q^{(n)}$

From the relationship in Eq. 2.3, we have

$$q_k^{(n,l+1)} = \sum_{i=0}^{k} r_i^{(n,l)} y_{k-i}^{(n,l)} \tag{2.60}$$

where $r_k^{(n,l)} := \mathbb{P}(R^{(n,l)} = k)$ and $y_k^{(n,l)} := \mathbb{P}(Y^{(n,l)} = k)$ for all non-negative integers k. Let $S^{(n,l)}$ be the number of available spaces for train l when it arrives at station n. Given that $S^{(n,l)} = u$, if u is greater than or equal to $Q^{(n,l)}$, all passengers can board and there is no left-behind. Then we have

$$r_0^{(n,l)}\Big|_{S^{(n,l)}=u} = \mathbb{P}(u \geq Q^{(n,l)}) = \sum_{k=0}^{u} q_k^{(n,l)} \tag{2.61}$$

where $r_k^{(n,l)}\Big|_{S^{(n,l)}=u} = \mathbb{P}(R^{(n,l)} = k \mid S^{(n,l)} = u)$. If u is less than $Q^{(n,l)}$, only u passengers can board and there are $Q^{(n,l)} - u$ number of left-behind passengers. So

$$r_i^{(n,l)}\Big|_{S^{(n,l)}=u} = \mathbb{P}(Q^{(n,l)} - u = i) = q_{u+i}^{(n,l)} \qquad i = 1, 2, ... \tag{2.62}$$

Based on Eq. 2.61 and 2.62, Eq 2.60 can be reformulated as

$$q_k^{(n,l+1)} = \sum_{u=0}^{C} s_u^{(n,l)} \left(\sum_{i=0}^{u} q_i^{(n,l)} y_k^{(n,l)} + \sum_{i=u+1}^{u+k} q_i^{(n,l)} y_{k-i+u}^{(n,l)} \right) \tag{2.63}$$

Assume the steady state probabilities for all variables exist, we have $\lim_{l\to\infty} q_k^{(n,l)} = q_k^{(n)}$, $\lim_{l\to\infty} s_k^{(n,l)} = s_k^{(n)}$, $\lim_{l\to\infty} y_k^{(n,l)} = y_k^{(n)}$. Taking the limit of l for both sides of Eq. 2.63 leads to

$$q_k^{(n)} = \sum_{u=0}^{C} s_u^{(n)} \sum_{i=0}^{u} q_i^{(n)} y_k^{(n)} + \sum_{u=0}^{C} s_u^{(n)} \sum_{i=u+1}^{u+k} q_i^{(n)} y_{k-i+u}^{(n)} \tag{2.64}$$

Assume the probability generating function (PGF) for $Q^{(n)}$, $R^{(n)}$ and $Y^{(n)}$ are $Q(z)$, $R(z)$, and $Y(z)$, respectively, where $Q^{(n)}$, $R^{(n)}$ and $Y^{(n)}$ and the steady state random variables of $Q^{(n,l)}$, $R^{(n,l)}$ and $Y^{(n,l)}$. Hence,

$$Q(z) = \sum_{k=0}^{\infty} q_k^{(n)} z^k \tag{2.65}$$

$$R(z) = \sum_{k=0}^{\infty} r_k^{(n)} z^k = \sum_{k=0}^{\infty} z^k \sum_{u=0}^{C} s_u^{(n)} \cdot r_k^{(n)}\Big|_{S^{(n)}=u} \tag{2.66}$$

$$Y(z) = \sum_{k=0}^{\infty} y_k^{(n)} z^k \tag{2.67}$$

Substituting Eq. 2.61 and 2.62 into 2.66 results in

$$R(z) = \sum_{u=0}^{C} s_u^{(n)} \sum_{i=0}^{u} q_i^{(n)} + \sum_{k=1}^{\infty} z^k \sum_{u=0}^{C} s_u^{(n)} q_{k+u}^{(n)} \tag{2.68}$$

$$= \sum_{u=0}^{C} s_u^{(n)} \left[\sum_{i=0}^{u} q_i^{(n)} + \frac{1}{z^u} Q(z) - \frac{1}{z^u} \sum_{i=0}^{u} q_i^{(n)} z^i \right] \tag{2.69}$$

Notice that $Q^{(n)} = R^{(n)} + Y^{(n)}$ (this is obtained by taking the limit of l for Eq. 2.3). Since $R^{(n)}$ and $Y^{(n)}$ are independent, we have

$$Q(z) = R(z)Y(z) \tag{2.70}$$

Combining Eq. 2.69 and 2.70 obtains

$$Q(z) = \frac{Y(z) \sum_{u=0}^{C} s_u^{(n)} \left[\sum_{i=0}^{u} q_i^{(n)} (1 - \frac{z^i}{z^u}) \right]}{1 - \sum_{u=0}^{C} s_u^{(n)} \frac{Y(z)}{z^u}}$$

$$= \frac{\sum_{u=0}^{C} s_u^{(n)} \left[\sum_{i=0}^{u} q_i^{(n)} (z^C - z^{C-u+i}) \right]}{\frac{z^C}{Y(z)} - \sum_{u=0}^{C} s_u^{(n)} z^{C-u}} \tag{2.71}$$

2.7.2 Derivation of $q_{0:C-1}^{(n)}$ by matching polynomial coefficients

Though $q_0^{(n)}, ..., q_{C-1}^{(n)}$ can be obtained by solving C system equations as mentioned in Section 2.4.3, we attempt to provide a more direct way to calculate $q_{0:C-1}^{(n)}$ in this section.

Using the fact that the numerator of $Q(z)$ is in the polynomial order of C, $Q(z)$ can be reformulated in terms of $z_1^*, ..., z_{C-1}^*$ as:

$$Q(z) = \frac{(z-1) \prod_{i=1}^{C-1} (z - z_i^*) \sum_{u=0}^{C} s_u^{(n)} \sum_{i=0}^{u} q_i^{(n)}}{\frac{z^C}{Y(z)} - \sum_{u=0}^{C} s_u^{(n)} z^{C-u}} \tag{2.72}$$

When $z \to 1$, both $\textsc{Num}(z)$ and $\textsc{Den}(z)$ approach 0. We also have the fact that

$\lim_{z\to 1} Q(z) = 1$. Therefore, using L'Hopital's rule we have

$$\lim_{z\to 1} Q(z) = 1 = \lim_{z\to 1} \frac{\textsc{Num}'(z)}{\textsc{Den}'(z)} = \frac{\prod_{i=1}^{C-1}(1-z_i^*)\sum_{u=0}^{C} s_u^{(n)} \sum_{i=0}^{u} q_i^{(n)}}{\sum_{u=0}^{C} s_u^{(n)} u - Y'(1)} \tag{2.73}$$

$$\Rightarrow \sum_{u=0}^{C} s_u^{(n)} \sum_{i=0}^{u} q_i^{(n)} = \frac{\sum_{u=0}^{C} s_u^{(n)} u - Y'(1)}{\prod_{i=1}^{C-1}(1-z_i^*)} \tag{2.74}$$

Define $\bar{Y}^{(n)} := Y'(1) = \mathbb{E}[Y^{(n)}]$ as the mean number of arrival passengers within a headway at station n, $\bar{S}^{(n)} := \sum_{u=0}^{C} s_u^{(n)} u =$ as the mean number of available spaces in an arriving bus at station n. Substituting Eq. 2.74 into 2.72, $Q(z)$ can be rewritten as

$$Q(z) = \frac{(\bar{S}^{(n)} - \bar{Y}^{(n)})(z-1)\prod_{i=1}^{C-1}\frac{z-z_i^*}{1-z_i^*}}{\frac{z^C}{Y(z)} - \sum_{u=0}^{C} s_u^{(n)} z^{C-u}} \tag{2.75}$$

Comparing Eq. 2.75 and 2.23, let the numerators of two equations be equal, we have

$$(\bar{S}^{(n)} - \bar{Y}^{(n)})(z-1)\prod_{i=1}^{C-1}\frac{z-z_i^*}{1-z_i^*} = \sum_{u=0}^{C} s_u^{(n)} \left[\sum_{i=0}^{u} q_i^{(n)}(z^C - z^{C-u+i})\right] \tag{2.76}$$

As the LHS and RHS of Eq. 2.76 are both polynomials about z, the coefficients of each polynomial in z must be equal. By matching the coefficients of z^0, we have

$$(-1)(\bar{S}^{(n)} - \bar{Y}^{(n)})\prod_{i=1}^{C-1}\frac{z_i^*}{z_i^*-1} = -q_0^{(n)} s_C^{(n)} \tag{2.77}$$

which leads to

$$q_0^{(n)} = \frac{1}{s_C^{(n)}}(\bar{S}^{(n)} - \bar{Y}^{(n)})\prod_{i=1}^{C-1}\frac{z_i^*}{z_i^*-1} \tag{2.78}$$

Note that $\prod_{i=1}^{C-1}\frac{z_i^*}{z_i^*-1}$ is always greater than 0 because 1) when C is odd, as the complex roots appear as conjugates, $\prod_{i=1}^{C-1}\frac{z_i^*}{z_i^*-1} > 0$. 2) when C is even, besides $z_0^* = 1$, there exists another real root on the negative real axis (denoted as $z_{\frac{C}{2}}^*$, where

$-1 \leq z^*_{\frac{C}{2}} < 0$). So, we have $\frac{z^*_{\frac{C}{2}}}{z^*_{\frac{C}{2}}-1} > 0$, which leads to $\prod_{i=1}^{C-1} \frac{z^*_i}{z^*_i-1} > 0$.

To validate Eq. 2.76, consider the fixed capacity situation where $s^{(n)}_C = 1$ and $\bar{S}^{(n)} = C$. Then Eq. 2.76 reduces to

$$q^{(n)}_0 \Big|_{s^{(n)}_C=1} = (C - \bar{Y}^{(n)}) \prod_{i=1}^{C-1} \frac{z^*_i}{z^*_i - 1} \tag{2.79}$$

This is the same as Chaudhry et al. [105]. Now we will derive $q^{(n)}_{1:C-1}$. Observing that the numerator of Eq. 2.75 can be rewritten as

$$\begin{aligned}(\bar{S}^{(n)} - \bar{Y}^{(n)})(z-1) \prod_{i=1}^{C-1} \frac{z - z^*_i}{1 - z^*_i} &= \frac{1}{s^{(n)}_C}(\bar{S}^{(n)} - \bar{Y}^{(n)}) \prod_{i=1}^{C-1} \frac{z^*_i}{z^*_i - 1} \prod_{i=1}^{C-1} \frac{z^*_i - z}{z^*_i}(z-1)s^{(n)}_C \\ &= s^{(n)}_C q^{(n)}_0 (z-1) \prod_{i=1}^{C-1} \left(1 - \frac{z}{z^*_i}\right) \\ &= -s^{(n)}_C q^{(n)}_0 \prod_{i=0}^{C-1} \left(1 - \frac{z}{z^*_i}\right) \end{aligned} \tag{2.80}$$

Define $\prod_{i=0}^{C-1} \left(1 - \frac{z}{z^*_i}\right) := \sum_{j=0}^{C} \eta_j z^j$, where η_j is the polynomial coefficient of z^j. For the RHS of Eq. 2.76, the polynomial coefficient of z^{C-k} is $-\sum_{u=k}^{C} s^{(n)}_u q^{(n)}_{u-k}$. And from Eq. 2.80, the polynomial coefficient of z^{C-k} is $-s^{(n)}_C q^{(n)}_0 \eta_{C-k}$. Matching the coefficient of the same order of z leads to

$$s^{(n)}_C q^{(n)}_0 \eta_{C-k} = \sum_{u=k}^{C} s^{(n)}_u q^{(n)}_{u-k} \qquad k = 1, 2, ..., C-1 \tag{2.81}$$

To validate Eq. 2.81, consider the fixed capacity situation where $s^{(n)}_C = 1$ and $s^{(n)}_k = 0, \forall\, 0 \leq k < C$. then Eq. 2.81 reduces to

$$q^{(n)}_{C-k} = q^{(n)}_0 \eta_{C-k} \qquad k = 1, 2, ..., C-1 \quad \text{if } s^{(n)}_C = 1 \tag{2.82}$$

which is the same as Chaudhry et al. [105].

Eq. 2.81 can be expressed in a matrix form by adding $s^{(n)}_C q^{(n)}_0 \eta^{(n)}_0 = s^{(n)}_C q^{(n)}_0$ (note that $\eta_0^{(n)} = 1$ by definition):

$$\tilde{\eta}^{(n)} = q_{0:C-1}^{(n)} \Lambda^{(n)} \tag{2.83}$$

where $\tilde{\eta}^{(n)} = [s_C^{(n)} q_0^{(n)} \eta_0^{(n)}, s_C^{(n)} q_0^{(n)} \eta_1^{(n)}, ..., s_C^{(n)} q_0^{(n)} \eta_{C-1}^{(n)}] \in \mathbb{R}^C$ and

$$\Lambda^{(n)} = \begin{bmatrix} s_C^{(n)} & s_{C-1}^{(n)} & s_{C-2}^{(n)} & ... & s_1^{(n)} \\ 0 & s_C^{(n)} & s_{C-1}^{(n)} & ... & s_2^{(n)} \\ ... & 0 & s_C^{(n)} & ... & s_3^{(n)} \\ 0 & ... & 0 & ... & s_4^{(n)} \\ 0 & 0 & ... & ... & ... \\ 0 & 0 & 0 & ... & s_C^{(n)} \end{bmatrix} \in \mathbb{R}^{C \times C} \tag{2.84}$$

As $s_C^{(n)} > 0$ is a known condition, the triangular matrix $\Lambda^{(n)}$ is invertible. Thus, we have

$$q_{0:C-1}^{(n)} = \tilde{\eta}^{(n)} (\Lambda^{(n)})^{-1} \tag{2.85}$$

2.7.3 Derivation of queue length mean and variance

Here we try to provide analytical formulations of $\mathbb{E}[Q^{(n)}]$ and $\mathrm{Var}[Q^{(n)}]$. The key is to find $Q'(1)$ and $Q''(1)$. The derivation follows the similar idea in Powell [48].

Let $A(z) = \frac{(\bar{S}^{(n)} - \bar{Y}^{(n)})(z-1)}{\frac{z^C}{Y(z)} - \sum_{u=0}^{C} s_u^{(n)} z^{C-u}}$ and $B_i(z) = \frac{z - z_i^*}{1 - z_i^*}$, then $Q(z) = A(z) \prod_{i=1}^{C-1} B_i(z)$. Based on the fact that $B_i(1) = 1$ and $Q(z) = 1$, we must have $A(1) = 1$. Hence,

$$\begin{aligned} Q'(1) &= A'(1) B_1(1) ... B_{C-1}(1) + A(1) B_1'(1) ... B_{C-1}(1) + ... + A(1) B_1(1) ... B_{C-1}'(1) \\ &= A'(1) + \sum_{i=1}^{C-1} B_i'(1) \end{aligned} \tag{2.86}$$

Since $B_i'(1) = \frac{1}{1 - z_i^*}$, the problem now becomes finding $A'(1)$. Again, let $A(z) = \frac{A_1(z)}{A_2(z)}$.

Then,

$$A'(z) = \frac{A_1'(z)A_2(z) - A_1(z)A_2'(z)}{(A_2(z))^2} \tag{2.87}$$

Notice that when $z \to 1$, the numerator and denominator of $A'(z)$ approach 0 (because $A_1(1) = 0$ and $A_2(1) = 0$). Therefore, applying L'Hopital's rule yields:

$$A'(z) = \frac{A_1''(z)A_2(z) - A_1(z)A_2''(z)}{2A_2(z)A_2'(z)} \tag{2.88}$$

Again we have 0/0 when $z \to 1$ because $A_1''(z) = 0$ and $A_2(1) = 0$. Applying L'Hopital's rule once more gives:

$$A'(z) = \frac{-A_1'(z)A_2''(z) - A_1(z)A_2'''(z)}{2A_2'(z)A_2'(z) + 2A_2(z)A_2''(z)} \tag{2.89}$$

Substituting $z = 1$ leads to

$$A'(1) = \frac{-A_1'(1)A_2''(1)}{2(A_2'(1))^2} \tag{2.90}$$

Based on the fact that $Y(1) = 1$, $Y'(1) = \bar{Y}^{(n)}$, $Y''(1) = \mathbb{E}[(Y^{(n)})^2] - \bar{Y}^{(n)}$, we have

$$A_1'(1) = \bar{S}^{(n)} - \bar{Y}^{(n)} \tag{2.91}$$

$$A_2'(1) = \frac{Cz^{C-1}Y(z) - Y'(z)z^C}{Y(z)^2} - \sum_{u=0}^{C}(C-u)s_u^{(n)}z^{C-u-1}\Bigg|_{z=1} = \bar{S}^{(n)} - \bar{Y}^{(n)} \tag{2.92}$$

$$\begin{aligned} A_2''(1) = &\frac{C(C-1)z^{C-2}}{Y(z)} - \frac{2Y'(z)Cz^{C-1}}{Y(z)^2} - \frac{Y''(z)z^C}{Y(z)^2} + \frac{2Y'(z)^2z^C}{Y(z)^3} \\ &- \sum_{u=0}^{C}(C-u)(C-u-1)s_u^{(n)}z^{C-u-2}\Bigg|_{z=1} \\ &= C(C-1) - 2\bar{Y}^{(n)}C - \mathbb{E}[(Y^{(n)})^2] + 2(\bar{Y}^{(n)})^2 + \bar{Y}^{(n)} - C^2 + C + 2C\bar{S}^{(n)} \\ &- \bar{S}^{(n)} - \mathbb{E}[(S^{(n)})^2] \\ &= -2\bar{Y}^{(n)}C - \mathbb{E}[(Y^{(n)})^2] + 2(\bar{Y}^{(n)})^2 + \bar{Y}^{(n)} + 2C\bar{S}^{(n)} - \bar{S}^{(n)} - \mathbb{E}[(S^{(n)})^2] \end{aligned} \tag{2.93}$$

Substituting Eq. 2.91, 2.92, and 2.93 into Eq. 2.90 results in

$$A'(1) = \frac{2\bar{Y}^{(n)}C + \mathbb{E}[(Y^{(n)})^2] - 2(\bar{Y}^{(n)})^2 - \bar{Y}^{(n)} - 2C\bar{S}^{(n)} + \bar{S}^{(n)} + \mathbb{E}[(S^{(n)})^2]}{2(\bar{S}^{(n)} - \bar{Y}^{(n)})} \tag{2.94}$$

Therefore, we have

$$\mathbb{E}[Q^{(n)}] = \frac{2\bar{Y}^{(n)}C + \mathbb{E}[(Y^{(n)})^2] - 2(\bar{Y}^{(n)})^2 - \bar{Y}^{(n)} - 2C\bar{S}^{(n)} + \bar{S}^{(n)} + \mathbb{E}[(S^{(n)})^2]}{2(\bar{S}^{(n)} - \bar{Y}^{(n)})} + \sum_{i=1}^{C-1} \frac{1}{1 - z_i^*} \tag{2.95}$$

To validate this formulation, let us consider a fixed capacity situation with $s_C^{(n)} = 1$. Then $\bar{S}^{(n)} = C$, $\mathbb{E}[(S^{(n)})^2] = C^2$. Then Eq. 2.95 reduces to

$$\mathbb{E}[Q^{(n)}]\big|_{s_C^{(n)}=1} = \frac{C - C^2 + 2\bar{Y}^{(n)}C + \mathbb{E}[(Y^{(n)})^2] - 2(\bar{Y}^{(n)})^2 - \bar{Y}^{(n)} +}{2(C - \bar{Y}^{(n)})} + \sum_{i=1}^{C-1} \frac{1}{1 - z_i^*} \tag{2.96}$$

which is equivalent to Powell [48]'s.

According to Eq. 2.29, the key to obtain $\mathrm{Var}[Q^{(n)}]$ is to calculate $Q''(1)$. Taking the logarithm of $Q(z) = A(z) \prod_{i=1}^{C-1} B_i(z)$ gives

$$\log Q(z) = \log A(z) + \sum_{i=1}^{C-1} \log B_i(z) \tag{2.97}$$

Taking derivatives of both sides leads to

$$\frac{Q'(z)}{Q(z)} = \frac{A'(z)}{A(z)} + \sum_{i=1}^{C-1} \frac{B_i'(z)}{B_i(z)} \tag{2.98}$$

Taking derivatives again:

$$\frac{Q''(z)}{Q(z)} - \frac{Q'(z)^2}{Q(z)^2} = \frac{A''(z)}{A(z)} - \frac{A'(z)^2}{A(z)^2} + \sum_{i=1}^{C-1} \left(\frac{B_i''(z)}{B_i(z)} - \frac{B_i'(z)^2}{B_i(z)^2} \right) \tag{2.99}$$

Solving for $Q''(z)$ and letting $z = 1$ gives:

$$Q''(1) = \mathbb{E}[Q^{(n)}]^2 + A''(1) - A'(1)^2 + \sum_{i=1}^{C-1} \left(B_i''(1) - B_i'(1)^2 \right) \tag{2.100}$$

Notice that $B_i''(1) = 0$ ($\forall i = 1, ..., C-1$) and $\mathbb{E}[Q^{(n)}] = Q'(1)$. Substituting Eq. 2.86 and 2.100 into Eq. 2.29 gives

$$\mathrm{Var}[Q^{(n)}] = A''(1) - A'(1)^2 + A'(1) + \sum_{i=1}^{C-1} \left(B_i'(1) - B_i'(1)^2 \right) \tag{2.101}$$

Now we only need to solve for $A''(1)$. The process is similar to finding $A'(1)$. Applying L'Hopital's rule five times to Eq. 2.89 and substituting $z = 1$ leads to

$$A''(1) = \frac{-2A_2'(1)A_2'''(1) + 3A_2''(1)^2}{6A_2'(1)} \tag{2.102}$$

Notice that the derivation process uses $A_1''(z) = 0$, $A_1(1) = 0$, $A_2(1) = 0$, and $A_1'(1) = A_2'(1)$. Details are omitted due to the tedious mathematical manipulation. To obtain $A_2'''(1)$, taking derivative of Eq. 2.93 gives:

$$\begin{aligned}
A_2'''(1) = & \left[\frac{C(C-1)(C-2)z^{C-3}}{Y(z)} - \frac{3Y'(z)C(C-1)z^{C-2}}{Y(z)^2} - \frac{3Y''(z)Cz^{C-1}}{Y(z)^2} + \frac{4Y'(z)^2Cz^{C-1}Y(z)}{Y(z)^4} \right. \\
& - \frac{Y'''(z)z^C}{Y(z)^2} + \frac{2Y(z)Y'(z)Y''(z)z^C}{Y(z)^4} + \frac{4Y''(z)Y'(z)z^C + 2Cz^{C-1}Y'(z)^2}{Y(z)^3} - \frac{6Y(z)Y'(z)^3z^C}{Y(z)^6} \\
& \left. - \sum_{u=0}^{C} (C-u)(C-u-1)(C-u-2)s_u^{(n)} z^{C-u-3} \right]_{z=1} \\
& = C(C-1)(C-2) - 3\bar{Y}^{(n)}C(C-1) - 3Y''(1)C + 6(\bar{Y}^{(n)})^2C - Y'''(1) + 6\bar{Y}^{(n)}Y''(1) \\
& - 6(\bar{Y}^{(n)})^3 - (C^3 - 3C^2 + 2C) + (2 + 3C^2 - 6C)\bar{S}^{(n)} + (3 - 3C)\mathbb{E}[(S^{(n)})^2] + \mathbb{E}[(S^{(n)})^3]
\end{aligned} \tag{2.103}$$

Notice that $Y'''(1) = \mathbb{E}[(Y^{(n)})^3] - 3\mathbb{E}[(Y^{(n)})^2] + 2\bar{Y}^{(n)}$. Hence,

$$\begin{aligned}
A_2'''(1) = {}& 3C^2\bar{S}^{(n)} - 3C^2\bar{Y}^{(n)} - 6C\bar{S}^{(n)} - 3C\mathbb{E}[(S^{(n)})^2] - 3C\mathbb{E}[(Y^{(n)})^2] + 6C(\bar{Y}^{(n)})^2 \\
& + 6C\bar{Y}^{(n)} + 2\bar{S}^{(n)} + 3\mathbb{E}[(S^{(n)})^2] + \mathbb{E}[(S^{(n)})^3] + 6\mathbb{E}[(Y^{(n)})^2]\bar{Y}^{(n)} + \\
& 3\mathbb{E}[(Y^{(n)})^2] - \mathbb{E}[(Y^{(n)})^3] - 6(\bar{Y}^{(n)})^3 - 6(\bar{Y}^{(n)})^2 - 2\bar{Y}^{(n)}
\end{aligned} \tag{2.104}$$

Substituting Eq. 2.92, 2.93, and 2.104 into Eq. 2.102 results in

$$\begin{aligned}
A''(1) = {}& \Bigg[6C^2(\bar{S}^{(n)})^2 - 12C^2(\bar{S}^{(n)})(\bar{Y}^{(n)}) + 6C^2(\bar{Y}^{(n)})^2 - 6C(\bar{S}^{(n)})\mathbb{E}[(S^{(n)})^2] \\
& - 6C(\bar{S}^{(n)})\mathbb{E}[(Y^{(n)})^2](\bar{Y}^{(n)})^2 + 12C(\bar{S}^{(n)}) + 6C\mathbb{E}[(S^{(n)})^2](\bar{Y}^{(n)}) + 6C\mathbb{E}[(Y^{(n)})^2](\bar{Y}^{(n)}) \\
& - 12C(\bar{Y}^{(n)})^3 - (\bar{S}^{(n)})^2 - 2(\bar{S}^{(n)})\mathbb{E}[(S^{(n)})^3] - 12(\bar{S}^{(n)})\mathbb{E}[(Y^{(n)})^2](\bar{Y}^{(n)}) + 2(\bar{S}^{(n)})\mathbb{E}[(Y^{(n)})^3] \\
& + 12(\bar{S}^{(n)})(\bar{Y}^{(n)})^3 + 2(\bar{S}^{(n)})(\bar{Y}^{(n)}) + 3\mathbb{E}[(S^{(n)})^2]^2 + 6\mathbb{E}[(S^{(n)})^2]\mathbb{E}[(Y^{(n)})^2] \\
& - 12\mathbb{E}[(S^{(n)})^2](\bar{Y}^{(n)})^2 + 2\mathbb{E}[(S^{(n)})^3](\bar{Y}^{(n)}) + 3\mathbb{E}[(Y^{(n)})^2]^2 \\
& - 2\mathbb{E}[(Y^{(n)})^3](\bar{Y}^{(n)}) - (\bar{Y}^{(n)})^2\Bigg]\Bigg/6(\bar{S}^{(n)} - \bar{Y}^{(n)})
\end{aligned} \tag{2.105}$$

Now with Eq. 2.105 and 2.94 we have

$$\begin{aligned}
& A''(1) - A'(1)^2 + A'(1) = \\
& \Bigg[(\bar{S}^{(n)})^2 - 4\bar{S}^{(n)}\mathbb{E}[(S^{(n)})^3] - 24\bar{S}^{(n)}\mathbb{E}[(Y^{(n)})^2]\bar{Y}^{(n)} + 4\bar{S}^{(n)}\mathbb{E}[(Y^{(n)})^3] \\
& + 24\bar{S}^{(n)}(\bar{Y}^{(n)})^3 - 2\bar{S}^{(n)}\bar{Y}^{(n)} + 3\mathbb{E}[(S^{(n)})^2]^2 + 6\mathbb{E}[(S^{(n)})^2]\mathbb{E}[(Y^{(n)})^2] - 12\mathbb{E}[(S^{(n)})^2](\bar{Y}^{(n)})^2 \\
& + 4\mathbb{E}[(S^{(n)})^3]\bar{Y}^{(n)} + 3\mathbb{E}[(Y^{(n)})^2]^2 + 12\mathbb{E}[(Y^{(n)})^2](\bar{Y}^{(n)})^2 - 4\mathbb{E}[(Y^{(n)})^3]\bar{Y}^{(n)} - 12(\bar{Y}^{(n)})^4 + (\bar{Y}^{(n)})^2\Bigg] \\
& \Bigg/ 12(\bar{S}^{(n)} - \bar{Y}^{(n)})^2
\end{aligned} \tag{2.106}$$

Slight manipulation of Eq. 2.106 leads to

$$
\begin{aligned}
&A''(1) - A'(1)^2 + A'(1) \\
&= \Big[-4(\bar{\bar{\bar{S}}}^{(n)} - \bar{\bar{\bar{Y}}}^{(n)})(\bar{S}^{(n)} - \bar{Y}^{(n)}) + 3(\bar{\bar{S}}^{(n)} + \bar{\bar{Y}}^{(n)})^2 - [6(\bar{\bar{S}}^{(n)} - \bar{\bar{Y}}^{(n)}) - 1](\bar{S}^{(n)} - \bar{Y}^{(n)})^2 \\
&- (\bar{S}^{(n)} - \bar{Y}^{(n)})^4 \Big] \Big/ 12(\bar{S}^{(n)} - \bar{Y}^{(n)})^2 \qquad (2.107)
\end{aligned}
$$

Observe that $B_i'(1) - B_i'(1)^2 = \frac{-z_i^*}{(1-z_i^*)^2}$. Therefore, substituting Eq. 2.106 into 2.101 gives the final results:

$$
\begin{aligned}
\mathrm{Var}[Q^{(n)}] = \frac{1}{12(\bar{S}^{(n)} - \bar{Y}^{(n)})^2} \Big\{ &-4(\bar{\bar{\bar{S}}}^{(n)} - \bar{\bar{\bar{Y}}}^{(n)})(\bar{S}^{(n)} - \bar{Y}^{(n)}) + 3(\bar{\bar{S}}^{(n)} + \bar{\bar{Y}}^{(n)})^2 - \\
&[6(\bar{\bar{S}}^{(n)} - \bar{\bar{Y}}^{(n)}) - 1](\bar{S}^{(n)} - \bar{Y}^{(n)})^2 - (\bar{S}^{(n)} - \bar{Y}^{(n)})^4] \Big\} - \sum_{i=1}^{C-1} \frac{z_i^*}{(1 - z_i^*)^2} \\
&\qquad (2.108)
\end{aligned}
$$

Chapter 3

Empirical analysis for the impact of service disruptions

3.1 Introduction

Urban public transit systems play a crucial role in cities worldwide, transporting people to jobs, homes, outings, and a variety of other activities. Millions rely on urban transit systems to provide them with transportation. However, transit systems are susceptible to unplanned delays and service disruptions caused by equipment, weather, passengers, or other internal and external factors.

Mitigating the impact of unplanned service disruptions is an important task for urban transit agencies. For this reason, it is important to recognize how a transit system is affected by service disruptions. The analysis framework for incident impacts can be summarized in Table 3.1. The two main dimensions of analysis, supply and demand, can further be broken down into "network performance" and "service" for supply analysis and "passenger flow" and "individual behavior" for demand analysis.

The network performance analysis usually uses graph theory-based techniques to calculate indicators related to incidents [80, 79, 78, 30], such as network resilience, vulnerability, redundancy, and a variety of other properties. The service analysis focuses on changes in an agency's operations during the incident period including headway, routing, staffing, and other operator-controlled factors designed to mitigate

Table 3.1: Analysis framework for incident impacts

	Analysis tasks	Description
Supply	Network performance	Indicators such as resilience, vulnerability, redundancy
	Service	Changes in agency's operations (e.g., headway, routing)
Demand	Passenger flow	Demand changes at different stations, lines
	Individual behavior	Passengers' mode choices under incidents

the incidents. From the demand point of view, passenger flow analysis investigates the demand changes at different stations, lines, or regions of a network, presenting passenger choices and flow redistribution after service disruptions. The individual behavior analysis focuses on studying the individual's response (such as mode choices, waiting time tolerance) to the incident and its relationship to the individual's characteristics (e.g., travel histories, demographics) [106, 96]. Surveys are usually used for such studies.

Previous research has used a variety of methods to analyze the impact of service disruption, including graph theory-based, survey-based, and simulation-based. Graph theory-based methods usually derive resilience or vulnerability indicators based on the network topology [78, 30, 79, 80]. These methods are effective for understanding high-level network properties related to incidents. Survey-based methods investigate passenger behavior during and opinions about the incident [81, 82, 83, 84, 85]. Passengers' individual-level behavior is analyzed and understood using econometric models. Simulation-based methods simulate passenger flows on the transit network under incident scenarios [86, 87, 88]. These studies can output many metrics of interest such as vehicle load changes, additional travel delays caused by incidents, distribution of the impact, etc.

Recently, automated data collection systems in transit networks enable a data-driven analysis of the impacts of service disruptions. The two major sources are automatic fare collection (AFC) and automatic vehicle location (AVL) data. AFC data is collected when passengers tap their transit cards on smart card readers (in buses or rail station gates). The records include times, locations, and card IDs. Depending on whether the fare system requires passengers to tap out, AFC data may only include tap-in records or both tap-in and tap-out records. AVL data records vehicle's (bus and train) time-dependent locations based on GPS and train tracking systems. From the AVL records, information such as headways can be inferred. Recently, a limited number of studies have been conducted using AFC and AVL data to look at unplanned transit disruptions. For example, Sun et al. [89] analyzed three types of abnormal passenger flows during unplanned rail disruptions using AFC data with both tap-in and tap-out records. Tian and Zheng [90] proposed a classification model to predict whether commuters switch from rail to other transportation modes because of unexpected travel delays using six months of AFC data.

However, despite numerous studies on incident analysis, there are still research gaps. First, for the graph theory-based approaches, the network indicators such as redundancy are usually defined for the whole network, an OD pair, or a link, and do not consider the influence of the disruption duration. Incidents usually cause service interruptions at multiple links depending on the power system and rail track configuration. And the duration of an incident can vary from 5 minutes to several hours, resulting in various impacts on the network. An incident-based indicator that reflects the network's redundancy under an actual incident with a specific location and duration is needed. Second, the studies that leverage AFC data to analyze passengers' mode choices under disruptions are very limited. Such approaches would require the inference of both individual choices and socio-demographic information from the AFC data. Third, most of the previous studies on incident analysis only addressed one or two aspects in Table 3.1 using case studies of a single incident. A comprehensive study that analyzes all four dimensions of the problem with comparable case studies using AFC and AVL data is missing from the literature.

The chapter aims to fill these research gaps by developing a data-driven methodology for the comprehensive analysis of the impact of unplanned rail disruptions on passengers and operations. Specifically, on the supply side, we propose an incident-based network redundancy index to analyze the ability of bus and rail networks to provide alternative services under a specific rail disruption. The impacts on operations are evaluated through headway changes across the systems. On the demand side, we calculate the demand changes at different rail lines, rail stations, bus routes, and bus stops to better understand the passenger flow redistribution under incidents. Individual behavior is analyzed using a binary logit model based on inferred passengers' mode choices and socio-demographics using AFC data. The public transit system of the Chicago Transit Authority (CTA) is used for a case study with two rail disruptions, one of which has high network redundancy and the other low.

The main contributions of this chapter are as follows:

- Propose an incident-based network redundancy index to reflect the system's ability to provide alternative services considering the integrated bus and rail systems. The index leverages the proposed concept of path throughput to incorporate the impact of the incident duration on the redundancy calculation.

- Develop an incident analysis framework using AFC and AVL data and apply it to incidents with different characteristics. Specifically, we analyze two types of incidents with high and low redundancy separately from both demand and supply perspectives.

- Propose an individual mode choice analysis method using AFC data. The approach includes a travel mode inference model and a passenger demographics extraction model. To the best of our knowledge, this is the first study that adopts AFC data for individual mode choice analysis during incidents.

- Conduct an empirical study to demonstrate the proposed framework using AFC and AVL data from two real-world incidents in the CTA system. The corresponding policy implications and operation suggestions are also discussed.

The remainder of this chapter is organized as follows. Section 3.2 reviews the literature. Section 3.3 presents the methodology used in this study. Case studies and data are described in Section 3.4 and results are discussed in Section 3.5. Section 3.6 concludes the chapter and discusses the policy implications.

3.2 Literature review

There are generally four methods researchers use to analyze the impact of disrupted operations: graph theory-based, survey-based, AFC data-based, and simulation-based. Graph theory-based analysis is majorly used for supply network performance and supply service analysis. Survey and AFC data-based methods are primarily used for passenger flows and individual behavior analysis. Lastly, simulation-based analysis can be used for both supply and demand analysis. Each method has strengths and weaknesses depending on the context.

3.2.1 Supply analysis

Network performance

Network performance analysis usually uses graph theory-based techniques to identify key aspects of the network's properties related to incidents, such as resilience, redundancy, and vulnerability based on graph theory (or complex network theory). For example, Yin et al. [78] studied subway networks with respect to disruptions, finding the weakness or critical locations of the network using "network betweenness" and "global efficiency" metrics. Similarly, Zhang et al. [30] built a general framework to assess the resilience of large and complex metro networks by quantitatively analyzing their vulnerability and recovery time using graph theory-based definitions.

Simulation is also used to evaluate the impact of incidents on networks. Usually, different hypothetical incident scenarios are tested. System performance metrics, such as travel delays and vehicle loads, are output to analyze the incident effects. For example, Suarez et al. [87] looked at the effects of climate change on Boston's transportation system performance using a simulation model, suggesting almost a doubling in delays and lost transit trips due to a variety of climate change effects.

Redundancy is an important indicator for analyzing the network performance under incidents. Redundancy is best defined as "the extent to which elements, systems, or other units of analysis exist that are substitutable, i.e., capable of satisfying functional requirements in the event of a disruption, degradation, or loss of function"

[26]. Redundancy has been widely studied, not just for transportation networks, but also in other areas including reliability engineering [107], communications [108], water distribution systems [109], and supply chain and logistics [110]. In terms of transportation-specific resiliency and redundancy, Berdica [80] developed a qualitative framework and basic concepts for vulnerability, resilience, and redundancy for transportation systems. Wilson-Goure et al. [111], Murray-Tuite [112], and Goodchild et al. [113], defined redundancy in the context of a specific transportation application areas such as evacuation, traffic network, and freight network.

However, nearly all previous studies defined redundancy at the level of networks, links, or OD pairs. Redundancy can also be defined for a specific incident to assess the system's ability to provide alternative services under a specific condition. This study proposes such an incident-based redundancy index to evaluate the network's ability to satisfy functional requirements under a specific incident. Both incident location and duration impact the redundancy. Moreover, the bus system, which is an important alternative for rail but rarely considered in previous studies, is included in the redundancy calculation.

Service Analysis

Service analysis mainly focuses on changes in an agency's operations during an incident period. This type of analysis looks at headways, routing, staffing, shuttle services, and other operator-controlled factors designed to mitigate the incident. For example, Nash and Huerlimann [114] developed a simulation model to analyze service variables such as headways and routing in the wake of disruptions. Schmöcker et al. [115] evaluated different operating strategies in six metro systems under service disruptions. Service delays and recovering times are treated as performance indicators. Similarly, Mo et al. [4] proposed an event-based simulation model that is capable of analyzing the impacts of incidents on service performance (e.g., headways).

3.2.2 Demand analysis

Passenger flow

Passenger flow analysis focuses on understanding how passengers choose alternative services **at an aggregated level**. Simulation-based methods can be applied to passenger flow analysis. For example, Hong et al. [88] simulated passenger flows in a metro station during an emergency. Using AFC data, Sun et al. [89] quantified three types of passenger flows: leaving the system, taking a detour, and continuing the journey but being delayed. This model was applied to the Beijing metro network. Tian and Zheng [90] looked at unexpected train delay effects on Singapore's MTR customers. Using AFC data, they built a classification model to predict whether commuters switch from MRT to other transportation modes because of unexpected train delays. Wu et al. [116] used AFC data to detect passenger flow volumes and travel time increases under station closures. Liu et al. [117] uses AFC data to comprehensively analyze unplanned disruption impacts, especially on passenger flows with trip cancellation, station changes, etc.

Individual behavior

Individual behavior analysis usually focuses on individual responses, like mode choice, waiting time tolerance, and a variety of other variables. These studies are usually conducted using surveys. Surveys are a good means to understand individual choices. Revealed preference (RP) and stated preference (SP) are two major types of survey design. Examples of transit-oriented RP studies include Currie and Muir [81], who conducted an RP survey to understand rail passengers' behavior, perceptions, and priorities in response to unplanned urban rail disruptions in Melbourne, Australia. Murray-Tuite et al. [82] used a web-based RP survey to understand the long-term impacts of a deadly metro rail collision in Washington DC. Tsuchiya et al. [118] conducted an RP survey in Japan that looked at passenger choices of four alternative routes. Pnevmatikou and Karlaftis [119] used RP survey data to analyze the effect of a pre-announced closure of an Athens Metro Line. SP survey studies include

Kamaruddin et al. [120], who studied the modal shift behavior of rail users after incidents. Fukasawa et al. [83] investigated the effect of providing information such as estimated arrival time, arrival order, and congestion level on passengers' modal shift behavior in response to an unplanned transit disruption. Similar research was conducted by Bai and Kattan [121], who found that various socioeconomic attributes and experience with the systems had strong influences on travelers' behavioral responses in the context of real-time information. Additionally, Rahimi et al. [96, 106] used a failure time model and a discrete choice model to analyze individuals' waiting time tolerances and mode choices, respectively, during unplanned service disruptions in Chicago using survey data.

The major drawback of survey-based methods is that they are time-consuming and labor-intensive. Hence, it is important to develop individual behavior analysis methods using AFC data as an alternative.

3.2.3 Comparison between our study and the literature

Table 3.2 summarizes the various studies in the literature from three aspects: study methods, data sources, and research focus. The main methodologies include graph theory-based methods (GTB), simulation-based (SB), optimization models (OM), descriptive analysis (DA), statistical inference (SI), machine learning (ML), and econometric models (EM).

Our study presents a comprehensive analysis focusing on four aspects: travel mode choice, passenger flow, redundancy, and service. It is also exclusive based on AFC and AVL data.

3.3 Methodology

In this section, we present the building blocks and methods used to support the analysis framework of an unplanned incident. On the supply side, a method to calculate the network redundancy index under a certain incident is proposed, which reflects the network's ability to provide alternative routes when incidents occur. To analyze

Table 3.2: Summary of literature review

Study	**Study Method**	**Data Sources**	**Research Focus**
Yin et al. [78]	GTB	Network, AFC	[1]NP - Efficiency
Zhang et al. [30]	GTB	Network	NP - Resilience
Balakrishna et al. [86]	SB	Network, Survey	NP - Efficiency, Passenger flow, Service
Hong et al. [88]	SB	Synthetic	Passenger Flow
Suarez et al. [87]	SB	Network, Geographical	NP - Resilience
Mo et al. [4]	SB, OM	Network, AFC	Passenger Flow, Service
Jenelius and Cats [122]	SB	Network, AFC	NP - Redundancy
Adnan et al. [123]	SB	Survey	NP - Efficiency, Service
Schmöcker et al. [115]	DA	AVL, AFC, Survey	NP - Resilience, Service
Sun et al. [89]	DA, SI	AFC	Passenger Flow
Tian and Zheng [90]	ML	AFC	Delay
Wu et al. [116]	SI	AFC	Passenger Flow
Liu et al. [117]	DA	AFC, AVL, Network	Passenger Flow, Delay
Currie and Muir [81]	EM	Survey	Travel mode choice, User satisfaction
Murray-Tuite et al. [82]	EM	Survey	Travel mode choice
Tsuchiya et al. [118]	EM	Survey	Travel mode choice
Pnevmatikou and Karlaftis [119]	EM	Survey	Travel mode choice
Kamaruddin et al. [120]	EM	Survey	Travel mode choice
Fukasawa et al. [83]	EM	Survey	Travel mode choice
Bai and Kattan [121]	EM	Survey	Travel mode choice
Lin et al. [85]	EM	Survey	Travel mode choice
Pnevmatikou et al. [124]	EM	Survey	Travel mode choice
Rahimi et al. [96]	EM	Survey	Travel mode choice
Rahimi et al. [106]	EM	Survey	User waiting behavior
Current study	GTB, EM, DA	AFC, AVL, Network	Travel mode choice, Passenger flow, NP - Redundancy, Service

[1]NP: Network performance.

the agency's service, we calculate the headway distribution using AVL data. On the demand side, we describe how to analyze passenger flows under incidents using AFC data, and how to use AFC data to analyze passengers' mode choice using a binary logit model.

To infer the effect of an incident, we compare data from the incident day to corresponding data from normal days. A "normal day" is defined as a recent day with the same day of the week and there are no incidents occurring in the incident line or nearby region during the incident period on that day. For example, if an incident happened on Friday 9:00-10:00 AM at a station, normal days can be all Fridays in recent months without incidents from 8:00-11:00 AM (a buffer is added to ensure normal services on a normal day) on the same line. Headways and passenger flow during the incident day are compared to those of normal days to reveal their difference.

3.3.1 Supply analysis

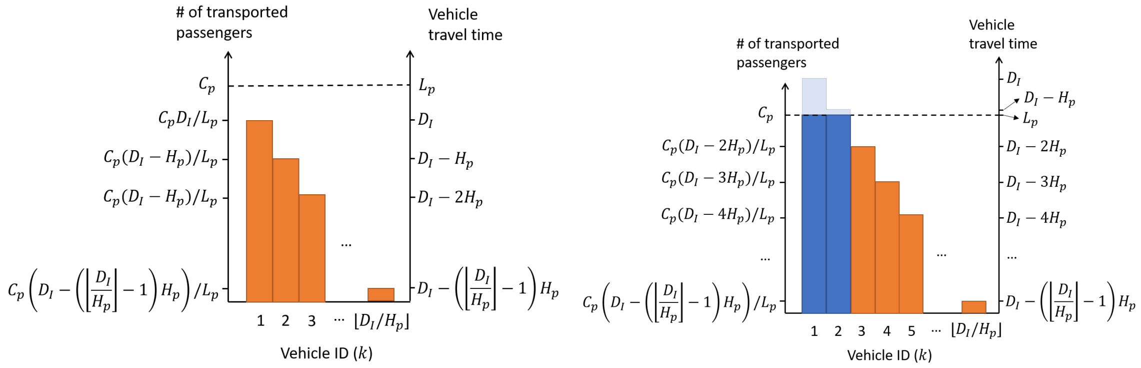
Network redundancy under incidents

As mentioned in Section 3.2, since redundancy is used to evaluate a network's functional response in the event of disruptions, it is important to develop an incident-specific redundancy index (opposite to network, link, or OD pair-specific in the literature). For a given incident, such an index can be used to evaluate the network's ability to provide alternative services under this incident. Furthermore, given the substitutional relationship between bus and urban rail systems, the proposed redundancy index in this study also explicitly considers the complementary role of bus and rail systems during the incident.

Redundancy is usually a function of the number of available paths for each OD pair because more available paths correspond to more opportunities of realizing the impacted trips when encountering service disruptions [79]. Hence, network redundancy under incidents (NRUI) should capture the transport capacity of alternative paths during the incident. Typical path capacity is defined as the maximum number of passengers transported per time unit (i.e. service frequency times the vehicle capacity). It is a time-insensitive value, which means the travel times of paths are not considered. However, for the redundancy calculation, path travel times are also important because passengers may not successfully finish their trips during the incident period if they choose paths with a long travel time. This means that the time-insensitive path capacity does not reflect the actual ability of paths to move passengers. Hence, a time-sensitive path capacity should be used for the calculation. In this study, we propose a new metric for the calculation of NRUI. The basis of the analysis uses the concept of throughput, instead of the classic definition of path capacity. In our approach, throughput explicitly takes into account the travel time on each alternative path. Throughput is defined as the number of "equivalent" passenger trips that have been completed per time unit during the incident. If a passenger has completed half of the trip on an alternative path by the time an incident is over, the "equivalent" trip count is 0.5.

More specifically, let $\mathcal{W}$ be the set of all OD pairs of the rail network. For an OD pair $w \in \mathcal{W}$, let $\mathcal{P}_w$ be the set of available paths for w *before the incident.* As we consider both bus and urban rail systems, a path $p \in \mathcal{P}_w$ may include segments of bus trips. $\mathcal{P}_w$ can be obtained in several ways, such as route choice surveys, Google Map API, and k-shortest paths. In this study, k-shortest path is used to obtain $\mathcal{P}_w$ with additional manual tuning to filter out unrealistic paths (e.g., too many transfers). Let D_I be the duration of incident I, H_p the headway of path p (defined as the maximum headway of each segment of path p), C_p be the vehicle (i.e. train or bus) capacity of path p (defined as the minimum vehicle capacity over all segments of path p), and L_p the travel time of path p. Then $\lfloor D_I/H_p \rfloor$ is the total number of vehicles dispatched on path p during the incident period. The throughput aims to capture the number of passengers at various stages of completing their trips *during the incident.* Figure 3-1 illustrates how the equivalent number of passengers completing trips are calculated during the incident period. If $D_I < L_p$ (Figure 3-1a), all vehicles in the path cannot reach the final destination. Therefore, the number of transported passengers is counted proportionally based on their travel time in the vehicle. For example, the first vehicle has traveled for D_I during the incident period (i.e. $\frac{D_I}{L_p}$ of the total path length). We assume this is equivalent to $C_p\frac{D_I}{L_p}$ completed passenger trips. And it is easy to show that the k-th vehicle's travel time is $D_I - (k-1)H_p$, which corresponds to $C_p\frac{D_I-(k-1)H_p}{L_p}$ equivalent completed passenger trips during the incident period. If $D_I \geq L_p$ (Figure 3-1b), the first vehicle can reach the destination. So it accounts for C_p completed passenger trips. In the example shown in Figure 3-1b, the second vehicle can also reach the destination during the incident (accounting for C_p passenger trips), while the third cannot (accounting for $C_p\frac{D_I-2H_p}{L_p}$ passenger trips). Therefore, combining these two scenarios, the number of equivalent completed passenger trips for vehicle k can be calculated as $\frac{\min\{D_I-(k-1)H_p, L_p\}}{L_p} \cdot C_p$.

Let A_p be the throughput of path p under incident I. From the analysis above, it



(a) $D_I < L_p$

(b) $D_I \geq L_p$

Figure 3-1: Illustration of path throughput. The bars show the number of equivalent completed passenger trips during the incident period (unfinished trips are counted proportionally based on their travel time). The orange (blue) bars represent vehicles that cannot (can) finish the trips.

can be formulated as

$$A_p = \frac{1}{D_I} \sum_{k=1}^{\lfloor D_I/H_p \rfloor} \frac{\min\{D_I - (k-1)H_p, L_p\}}{L_p} \cdot C_p \tag{3.1}$$

Eq. 3.1 counts the total number of equivalent passenger trips along path p that have been completed per time unit during the incident (passengers who did not finish their trips are counted proportionally).

A larger value of D_I implies that A_p is less sensitive to the path travel time. On the extreme situation where $D_I \to \infty$, $A_p \to \frac{C_p}{H_p}$ (proof in Section 3.7.1), which corresponds to the typical definition of capacity where L_p does not matter. The intuition behind this is that the proposed A_p limits the capacity calculation *in the incident period*. When D_I is large, even if passengers have a longer travel time on a path, the majority of the passengers impacted by the incident will have their trips completed. On the contrary, if D_I is small, most of the passengers using paths with long travel times cannot finish their trips. The typical definition of $\frac{C_p}{H_p}$ does not capture this important aspect. Hence, considering travel time in the redundancy calculation is much more representative of the actual conditions.

In summary, A_p is an indicator reflecting a path's ability to serve impacted trips during the incident period. And $\sum_{p\in\mathcal{P}_w} A_p$ reflects the ability of the network to provide services for OD pair w. Actually, one of the network-level definitions of redundancy in the literature is $\sum_{w\in\mathcal{W}}\sum_{p\in\mathcal{P}_w} A_p$ (where A_p is defined differently), which measures the total path capacity in the network [125].

In this study, we want to capture the incident-specific characteristics of redundancy. Let $\mathcal{W}_I$ be the set of all OD pairs with at least one path blocked due to incident I. Mathematically, $\mathcal{W}_I = \{\ w \in \mathcal{W} : \exists p \in \mathcal{P}_w$ s.t. p is blocked due to incident $I\ \}$. Then, only passengers with OD in $\mathcal{W}_I$ are affected by the incident. Let the total path throughput of w before the incident be T_w.

$$T_w = \sum_{p\in\mathcal{P}_w} A_p \qquad \forall w \in \mathcal{W}_I \tag{3.2}$$

Because of the incident, passengers may augment their typical path choice alternatives with paths that were not considered before the incident. Hence, we define $\tilde{\mathcal{P}}_w$ as the set of available paths for $w \in \mathcal{W}_I$ *during the incident.* $\tilde{\mathcal{P}}_w$ can be seen as $\mathcal{P}_w$ without the blocked paths and adding the augmented paths. Usually, augmented paths are longer and less preferred by passengers. For a specific OD pair $w \in \mathcal{W}_I$, the total path throughput of w after the incident, denoted as $\tilde{T}_w$, should be less than or equal to that before. Therefore, we define $\tilde{T}_w$ as:

$$\tilde{T}_w = \min\{\sum_{p\in\tilde{\mathcal{P}}_w} A_p,\ T_w\} \qquad \forall w \in \mathcal{W}_I \tag{3.3}$$

This corresponds to our assumption that the throughput during the incident cannot exceed that before the incident for a specific OD pair. Hence, the NRUI for incident I is formulated as:

$$R_I = \frac{\sum_{w\in\mathcal{W}_I} \tilde{T}_w}{\sum_{w\in\mathcal{W}_I} T_w} \tag{3.4}$$

where the numerator (denominator) is the total throughput of available paths after (before) the incident. Since $T_w \geq \tilde{T}_w$ for all $w \in \mathcal{W}_I$ by definition, we have $0 \leq R_I \leq 1$.

$R_I = 1$ means the capacities before and after the incident are the same, suggesting that the incident does not deteriorate the function of the network (i.e. the network is fully redundant under incident I). $R_I = 0$ means no alternative paths are available during the incident (i.e. the network has no redundancy under incident I)

For a better understanding of the index, we present a small numerical example to show how R_I is calculated. As shown in Figure 3-2, consider a system with only one OD pair w. Path 1 and 2 are primary and alternative paths, respectively, where $\mathcal{P}_w = \{1\}$ and $\tilde{\mathcal{P}}_w = \{2\}$ (i.e. before the incident path 2 is not chosen by passengers). The attributes of the two paths are shown in the figure. For path 1, there are $\lfloor D_I/H_p \rfloor = 2$ vehicles dispatched. The first vehicle has traveled for $\min\{D_I, L_1\} = 20$ minutes and reached the destination. The second vehicle, which was dispatched 30 minutes later, can also travel for $\min\{D_I - H_1, L_1\} = 20$ minutes and reach the destination. Therefore, these two vehicles successfully carried 400 passengers to the destination. According to Eq. 3.1, we have

$$A_1 = \frac{\min\{D_I, L_1\}}{L_1 D_I} C_1 + \frac{\min\{D_I - H_1, L_1\}}{L_1 D_I} C_1 = 400 \text{ passengers/hour} \tag{3.5}$$

In terms of path 2, similarly, there are two vehicles dispatched during the incident. The first vehicle can reach the destination and the second one can only finish $\frac{30}{60}$ of its journey. Therefore,

$$A_2 = \frac{\min\{D_I, L_2\}}{L_2 D_I} C_2 + \frac{\min\{D_I - H_2, L_2\}}{L_2 D_I} C_2 = 300 \text{ passengers/hour} \tag{3.6}$$

The two terms in Eq. 3.6 represent the number of passengers carried (successfully and partially) by the two vehicles in path 2 per time unit.

The redundancy index for this single network under incident I is

$$R_I = \frac{\tilde{T}_w}{T_w} = \frac{A_2}{A_1} = \frac{300}{400} = 0.75 \tag{3.7}$$

which means during the incident when passengers start to use path 2, the system maintains 75% of its original capacity. For comparison purpose, if one follows the typical definition of path capacity and calculate A_p as $\frac{C_p}{H_p}$, the redundancy index will be 1 because $C_1 = C_2$ and $H_1 = H_2$ in this example, which is obviously unreasonable because it implies that the system maintains 100% capacity and the incident has no impact.

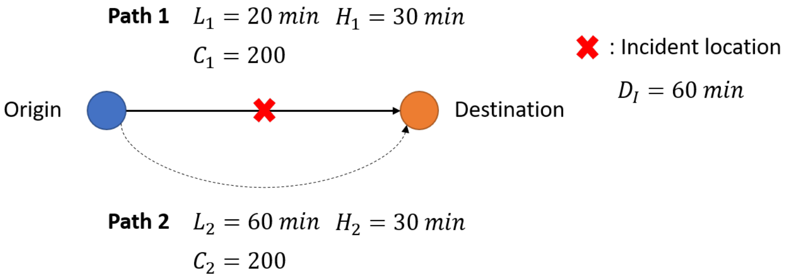

Path 1 $L_1 = 20\ min$ $H_1 = 30\ min$
$C_1 = 200$
✖ : Incident location
$D_I = 60\ min$
Origin
Destination
Path 2 $L_2 = 60\ min$ $H_2 = 30\ min$
$C_2 = 200$

Figure 3-2: Example of network redundancy calculation

It is worth noting that we illustrate the definition of the NRUI assuming a fixed vehicle capacity C_p. Actually, C_p can be defined as the "available capacity" in a vehicle considering the onboard passengers. In this way, the NRUI can also capture the **impact of demand**. The "available capacity" can be calculated for the specific day of the incident using a transit assignment model [4]. Alternatively, the average "available capacity" can also be used. In the case study, due to lack of data, the total capacity of a vehicle is used.

Headway analysis

Headways are important indicators of the level of service for transit systems. Analyzing headway patterns during an incident can provide direct information about how services are reduced by the incident. As mentioned in Section 3.1, AVL data provide the headway of each station in the urban rail system. In this study, we calculate the headway temporal distribution for lines of interest to capture the impact of incidents.

Let us divide the analysis time period into several intervals with equal length. Denote the headway on station i of trip j as $H_{i,j}$ (i.e. the length of interval between trip j and $j-1$ and $H_{i,1} = 0$). Suppose line l has two directions, inbound and outbound. The headway of line l *outbound* at time interval τ is calculated as

$$H_{l,\tau}^{\text{out}} = \frac{\sum_{i \in \mathcal{S}_l^{\text{out}}} \sum_{j \in \mathcal{R}_{i,\tau}} H_{i,j}}{\sum_{i \in \mathcal{S}_l^{\text{out}}} |\mathcal{R}_{i,\tau}| - 1} \tag{3.8}$$

where $\mathcal{S}_l^{\text{out}}$ is set of outbound stations in line l. $\mathcal{R}_{i,\tau}$ is the set of trips passing through station i during the time interval τ. Eq. 3.8 implies the headway of a line is calculated as the mean of all stations along the line. The inbound headway, $H_{l,\tau}^{\text{in}}$, is calculated in a similar way by replacing $\mathcal{S}_l^{\text{out}}$ with $\mathcal{S}_l^{\text{in}}$. The headway distributions of both normal days and the incident day are calculated for comparison.

3.3.2 Demand analysis

Passenger flow analysis

AFC data record passengers' tap-in information in bus and rail systems (tap-out is not available in this study). These transactions can capture passengers' route choices during an incident if they use the transit system again [126]. Therefore, analyzing AFC data can help understand passenger flow redistribution during an incident.

At the station level, we calculate the number of tap-in passengers at the stations in the incident area, and compare the values on the incident day and normal days. The difference in this number is an indicator of the impact of the incident on passenger flow redistribution. Stations with high demand increases reflect passengers' choices after the incident. Similarly, at the line level, we calculate the number of tap-in passengers for lines near the incident area for both the incident day and normal days. Line-level demands are calculated as the sum of all station-level demands in corresponding lines.

Note that we assume the number of tap-in passengers is approximately normally distributed. Hence, if the incident day demand is beyond the ±2×standard deviation of the normal day demand, we say that a significant difference is observed (i.e., the impact of the incident is significant).

Individual behavior analysis

Passengers may make different mode choices after the incident. One important question is how the characteristics of the passengers influence their mode choices. This is typically using data from surveys. In this study, we propose a method based on conveniently available AFC data for individual behavior analysis.

The proposed approach consists of two steps: a) inferring individual's mode choice and b) extracting samples' characteristics. We infer individual choices using AFC data. In this study, only two choices are considered: 1) using transit and 2) other (including canceling trips and using other travel modes). This is because these two options can be confidently identified using AFC data and they are important for transit operators. Since passenger travel patterns in transit systems show high irregularity [127], it is more convenient to identify the behavioral changes of *regular passengers* [128]. In this study, we define regular passengers as those who use the public transit system every normal day and have the same travel trajectories. Note that, as normal days have the same day of week as the incident day, regular passengers are not necessarily frequent users as they may only use the system on a specific day of week. For example, if an incident happened on Friday, a passenger who only uses the public transit system on Friday (i.e. on each normal day) is a regular passenger. But since he/she only uses the transit once a week, he/she may not be a frequent user. Mathematically, let us denote the trajectories of passenger i in normal day k as $\mathcal{T}^{i,k} = \{(o_1^{i,k}, d_1^{i,k}, t_1^{i,k}), ..., (o_{N^{i,k}}^{i,k}, d_{N^{i,k}}^{i,k}, t_{N^{i,k}}^{i,k})\}$, where $o_n^{i,k}$, $d_n^{i,k}$, and $t_n^{i,k}$ are the origin, destination and start time of the n-th trip, respectively ($t_1^{i,k} < ... < t_{N^{i,k}}^{i,k}$). $N^{i,k}$ is the total number of trips in normal day k for passenger i. The set of regular passengers is defined as $\{i \in \mathcal{I} \mid o_n^{i,k} = o_n^{i,k'}, d_n^{i,k} = d_n^{i,k'}, t_n^{i,k} \in [\bar{t}_n^i - \sigma_n^i, \bar{t}_n^i + \sigma_n^i], N^{i,k} = N^{i,k'}, \forall k, k' \in \mathcal{K}, k \neq k'\}$, where $\mathcal{I}$ is the set of all passengers, $\bar{t}_n^i$ and σ_n^i are the mean and standard deviation of the start time of trip n for passenger i over all normal days. $\mathcal{K}$ is the set of all normal days considered in this study. This means regular passengers have the same number of trips and corresponding origins and destinations in each normal day (for tap-in only AFC systems, destinations are not considered).

And the corresponding trip start times in each normal day are stable (i.e. within a standard deviation). Hence, if these passengers had different travel patterns on the incident day, most likely they would be affected by the incident and chose a new travel mode. Denote the trip sequence of passenger i on the incident day as $\mathcal{T}^{i,\text{In}} = \{(o_1^{i,\text{In}}, d_1^{i,\text{In}}, t_1^{i,\text{In}}), ..., (o_{N^{i,\text{In}}}^{i,\text{In}}, d_{N^{i,\text{In}}}^{i,\text{In}}, t_{N^{i,\text{In}}}^{i,\text{In}})\}$. And let $[T_e, T_s]$ be the incident period, where T_e and T_s is the incident start and end time. The mode choice during the incident for a regular passenger i is denoted as Y_i. We infer Y_i as follows:

- Y_i = "Transit" if 1) there are additional transit trips (compared to that in a normal day) during the incident period or 2) there are changes of tap-in stations during the incident period. The first condition implies that the regular passenger may have transferred to a nearby rail station or bus stop, with more transit trips than usual. The second condition implies that the regular passenger may have changed to a different rail line or bus route in response to the incident. Let $\mathcal{T}_I^{i,k} = \{(o_n^{i,k}, d_n^{i,k}, t_n^{i,k}) \in \mathcal{T}^{i,k} \mid T_s \leq \bar{t}_n^i \leq T_e\}$ and $\mathcal{T}_I^{i,\text{In}} = \{(o_n^{i,\text{In}}, d_n^{i,\text{In}}, t_n^{i,\text{In}}) \in \mathcal{T}^{i,\text{In}} \mid T_s \leq t_n^{i,\text{In}} \leq T_e\}$ be sub-sequences of trips within the incident period (i.e. $[T_s, T_e]$) on a normal and the incident day, respectively. Mathematically, the first condition can be expressed as: $|\mathcal{T}_I^{i,\text{In}}| > |\mathcal{T}_I^{i,k}|$ and the second: $\exists n$ s.t. $o_n^{i,k} \neq o_n^{i,\text{In}}$, where $(o_n^{i,k}, d_n^{i,k}, t_n^{i,k}) \in \mathcal{T}_I^{i,k}$ and $(o_n^{i,\text{In}}, d_n^{i,\text{In}}, t_n^{i,\text{In}}) \in \mathcal{T}_I^{i,\text{In}}$. Note that $k \in \mathcal{K}$ can be any normal day because the trajectories for all normal days are the same for a regular passenger by definition.

- Y_i = "Other" if the transit trips that are supposed to happen during the incident period on the normal days disappear on the incident day. This means that the regular passengers may change to other modes or cancel their trips. Mathematically, this can be expressed as $|\mathcal{T}_I^{i,\text{In}}| < |\mathcal{T}_I^{i,k}|$.

Other regular passengers without the above behavior may not be affected by the incident or have other choices that are hard to be identified (e.g., transfer to another line without leaving the system), which are not considered in the analysis.

In the second step, the characteristics of each regular passenger (i.e. demographics and trip information) are extracted. We aim to use information that is available in

AFC and sale transaction data as a proxy to passengers' socio-demographics.

Since regular passengers have consistent travel trajectories, we can infer their home locations as the tap-in rail station or bus stop of the first trip on a normal day (i.e. $o_1^{i,k}$ for any $k \in \mathcal{K}$). Given the station/stop location, we can obtain the median household income in passenger i's neighborhood or census tract using census data. Living in a high-income or low-income neighborhood can be a proxy of passengers' income. AFC data can also provide passengers' fare status information, such as whether the passenger is in a reduced fare status. Reduced fare status users are usually students, seniors, and people with disabilities. This information is also a proxy for socio-demographic characteristics.

Sale transaction data provide the historical add-value transactions of passengers. We extract three variables in this study: total added value per year, add-value frequency (i.e. number of add-value transactions per year), and maximum single added value in a year. The first two variables reflect the passenger's dependence on and familiarity with public transit and part of their income information. The last variable can also be used to some extent as a proxy for income because low-income people may not be able to deposit a large amount of money in the smart card at once. We denote all this "proxy" demographic information for passenger i as X_i.

The characteristics of passenger i's trip (denoted as Z_i) during the incident may also affect mode choices. We define the incident-related trip (trip ID denoted as n^*) as the first trip with $\bar{t}^{i,k}$ in the incident period. Mathematically, $n^* = \arg\min_n \{n = 1, ..., N^{i,k} \mid \bar{t}_n^i \in [T_s, T_e]\}$. Since regular passengers are supposed to have stable travel patterns, $o_{n^*}^{i,k}$ and $d_{n^*}^{i,k}$ should be the intended origin and destination for passenger i on the incident day. Based on $(o_{n^*}^{i,k}, d_{n^*}^{i,k})$, two trip-related variables are considered. The first is whether the $d_{n^*}^{i,k}$ is downtown, which is a proxy for work trips. Note that for a tap-in only system, $d_{n^*}^{i,k}$ can be inferred from a destination estimation model [129, 130, 131]. The second variable is *OD-based redundancy*, defined as

$$R_i^{\text{OD}} = \frac{\tilde{T}_{w_i}}{T_{w_i}} \tag{3.9}$$

where R_i^{OD} is the OD-based redundancy for passenger i, measuring the availability of alternative transit services for the specific OD pair during an incident. $w_i = (o_{n^*}^{i,k}, d_{n^*}^{i,k})$ is passenger i's OD pair for the incident-related trip. It is worth noting that R_i^{OD} can be seen as the NRUI for the case of a single OD pair.

In this study, we use a binary logit model [132] to better understand the main factors that impact choice Y_i. Let the utility of mode j for passenger i be U_{ij}.

$$U_{ij} = \text{ASC}_j + \alpha_j X_i + \beta_j Z_i + \epsilon_j \tag{3.10}$$

where ASC_j is the alternative specific constant (ASC) for mode j. ϵ_j is the error term that is assumed to be Gumbel distributed. α_j and β_j are the vectors of parameters to be estimated. The probability of passenger i choosing mode j is

$$\mathbb{P}(Y_i = j) = \frac{\exp(\text{ASC}_j + \alpha_j X_i + \beta_j Z_i)}{\sum_{j' \in \mathcal{C}} \exp(\text{ASC}_j + \alpha_j X_i + \beta_j Z_i)} \qquad \forall j \in \mathcal{C} \tag{3.11}$$

where $\mathcal{C} = \{$“Transit”,“Other”$\}$ is the choice set.

The approach of the individual behavioral analysis model is summarized in Figure 3-3.

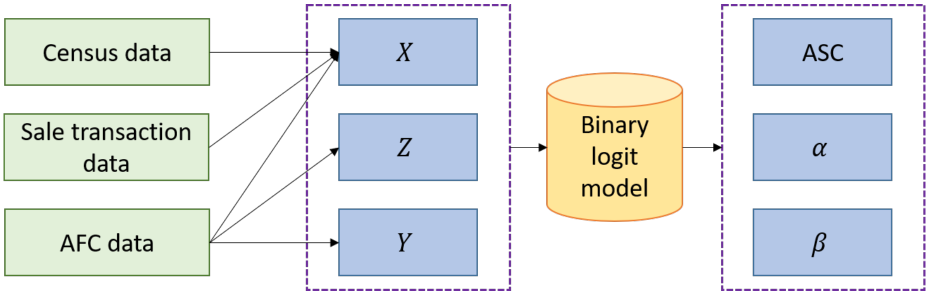


Figure 3-3: Summary of the individual behavioral analysis model

3.4 Application

3.4.1 Chicago Transit System

We use incident data from the Chicago Transit Authority (CTA) public transit system for the model application in this section. CTA is the second-largest transit system in the United States, providing services in Chicago, Illinois, and some of its surrounding suburbs. It operates 24 hours each day and is used by 0.84 million bus and 0.81 million train passengers per weekday on average [40]. The map of the CTA rail system is shown in Figure 3-4. The rail system consists of eight lines (named after their color) and the "Loop". The Loop, located in the Chicago downtown area, is a 2.88 km long circuit of elevated rail that forms the hub of the Chicago rail system. Its eight stations account for around 10% of the weekday boardings of the CTA trains.

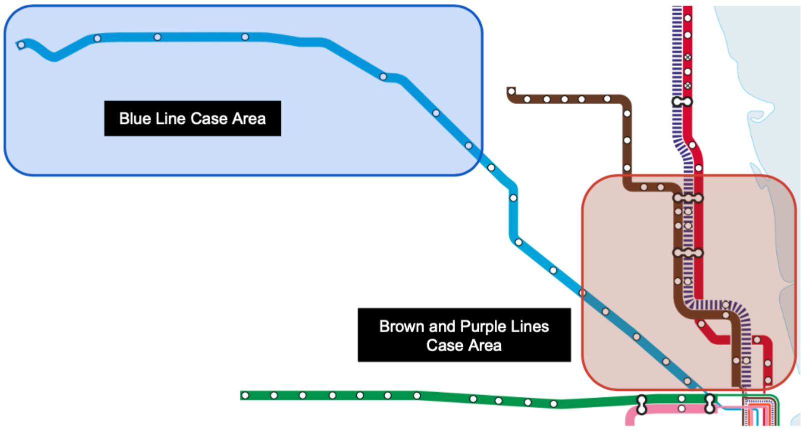


Figure 3-4: CTA rail system map

CTA's AFC system is entry-only, meaning passengers use their farecards only when entering a rail station or boarding a bus, and so no information about a trip's destination is directly provided. The train tracking system provides train arrival and departure times at each station.

According to the control center data, CTA experienced a total of 27,198 incidents in 2019. However, around 80% percent of the incidents have a duration of fewer than 10 minutes. Since small incidents may not affect the system significantly, this study focuses on substantial incidents that lasted longer than 1 hour. Passengers who leave the rail system because of service disruptions need to re-tap in if they decide to use other CTA services (buses or rails). They are only charged a transfer fee. However, no tap-in is needed for shuttle service that may have been deployed in response to the incident. Hence, there is no information for passengers using shuttle buses.

3.4.2 Redundancy index

Prior to analyzing actual incidents, we first present an overview of the CTA system redundancy. As the NRUI is defined based on each incident, for the purpose of this analysis, we assume that a hypothetical incident takes place at a station in the system (one at a time), blocking the track segment that connects the station for 1 hour. Note that if a station has two separate tracks, each track is blocked independently and there will be two incident cases for this station. For example, the Roosevelt station has two different tracks for the Red Line and Purple/Yellow Lines. So two hypothetical incident cases are generated, each corresponding to the interruption of a track. Considering the infrastructure layout of urban rail systems, assuming that incidents occur at the track level is more realistic than simply assuming an incident blocks the whole station as in many previous studies that used graph-based methods [133, 134].

Besides the incident-specific redundancy index, the occurrence frequency of incidents at various stations is also of importance. Figure 3-5 shows the redundancy index at each station against the number of incidents taking place per year at that station (only incidents with a duration greater than 10 minutes are counted). The combination of the two metrics divides the figure into four sections: 1) Stations in the red section (upper left) have high incident occurrence frequency and low redundancy. These are critical stations in the system where alternative public transit services are limited and service disruptions happen frequently. Transit operators need to prepare strategies in advance for these stations. 2) Stations in the yellow section (upper right) have high incident occurrence frequency and high redundancy. In these stations, passengers are able to seek alternative services during a disruption. Operators need to provide direct information to passengers with suggestions regarding alternatives. 3)

Stations in the blue section (lower left) have low incident occurrence frequency and low redundancy. Though incidents may not happen frequently, mitigation plans need to be prepared as there are limited substitutional services. 4) Stations in the green section (lower right) have low incident occurrence frequency and high redundancy. These stations are less critical in terms of incident management compared to stations in other sections.

Figure 3-5 shows that most of the stations in the CTA system are in the blue or green sections. And only a limited number of stations are in the red section. This implies that CTA can focus more on some critical stations with adequate incident management strategies. In terms of critical stations (red section), most of them are terminal stations (such as Howard, Forest Park). This is expected as terminal stations usually have more complex infrastructure layouts (i.e. more prone to failures) and are usually located in suburban areas (i.e. fewer alternative services and low redundancy index). Backup shuttle services can be provided in these stations.

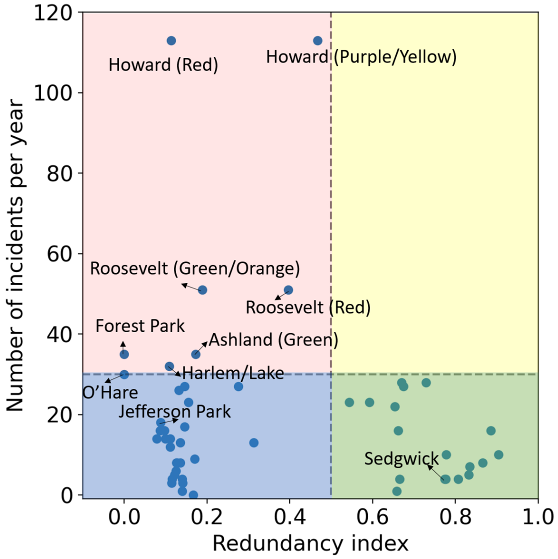


Figure 3-5: Redundancy index v.s. incident occurrence rate.

3.4.3 Rail disruption cases

Since the location of the incidents may influence their impact, we selected two incidents at locations with high and low redundancy, respectively, for comparative analysis.

Brown and Purple Lines Sedgwick incident

On September 24 (Tuesday), 2019, at 9:09 AM, a Purple Line train collided with a Brown Line train at the Sedgwick station. The incident caused a number of stations to be blocked and closed in both Brown and Purple Lines since these two lines share the same track in this area. The impacted stations were Fullerton and Armitage to the north and Chicago and Merchandise Mart (MM) to the south. Southbound trains short turned at Fullerton, while northbound trains short turned at MM. At 9:28 AM, 19 minutes after the incident started, bus substitution service began between Fullerton to MM. Service resumed at all blocked stations at 10:19 AM, 70 minutes after the start of the incident. The incident on the Brown and Purple Lines is a high redundancy case because the Red Line is a good substitution for the incident location (See Figure 3-6).

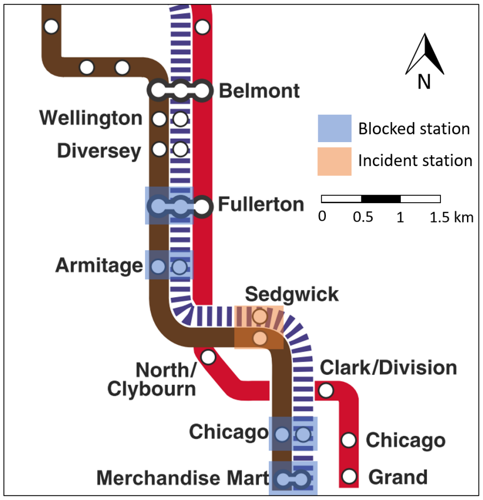


Figure 3-6: Incident diagram of Brown Line Sedgwick case

Blue Line Jefferson Park incident

On February 1 (Friday), 2019, at 8:14 AM, the inbound track Blue Line between Harlem and Jefferson Park was closed due to infrastructure problems. All trains in the Blue Line were suspended. CTA used the remaining single-direction track to serve trains from both directions in the incident link. At 9:03 AM, 49 minutes after the incident, single track operations commenced between Harlem and Jefferson Park, with shuttle service starting 7 minutes later. At 9:40 AM, all inbound trains succeeded to move under the single-track operation. At 12:09 PM, the full line was reopened. The entire incident lasted 4 hours and 9 minutes. The incident on the Blue Line is a low redundancy case because the Blue Line is far away from other rail lines with limited alternative services (see Figure 3-7).

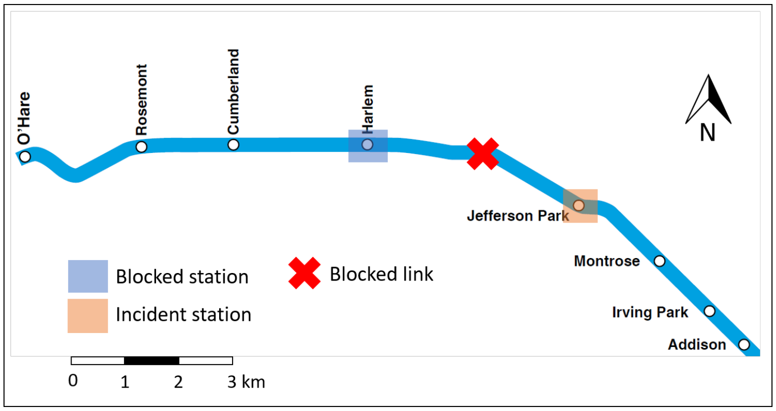


Figure 3-7: Incident diagram of Blue Line Jefferson Park case

3.5 Analysis

The framework discussed in Section 3.3 was used for the analysis of the cases. For each case, the results are organized from supply to demand analysis. The individual choice analysis is conducted based on samples of affected passengers from two incidents.

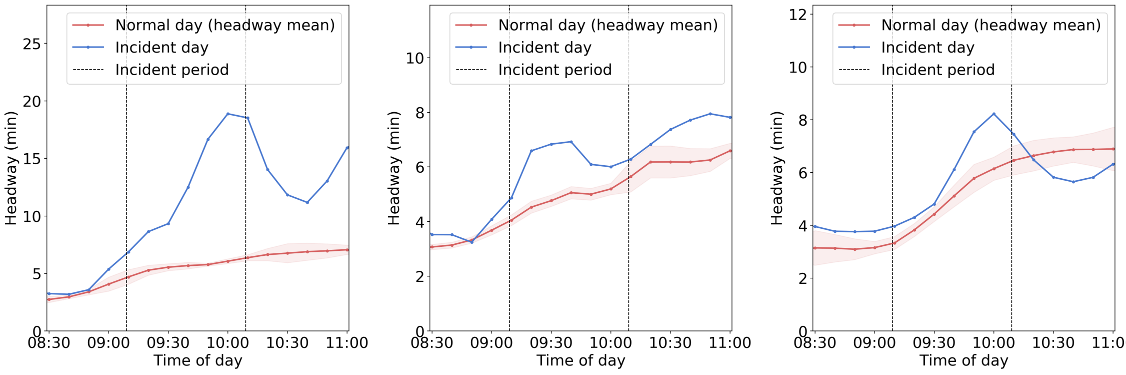
3.5.1 Brown and Purple Line incident analysis

Redundancy index

The NURI (Eq. 3.4) for the Brown and Purple Line case is 0.732, meaning that the transit system maintains 73.2% transporting capacity for the Brown and Purple Lines incident during the incident period. The high redundancy of the Brown and Purple Lines incident is as expected. In the incident area, the Red Line is almost parallel with the Brown and Purple Lines. In addition, there exist many south-bound bus routes going to Downtown Chicago (see Figure 3-11). This implies that during the Brown and Purple Line incident, CTA can focus on guiding passengers to find alternative services. Some information dissemination strategies need to be applied, such as route and transfer recommendations.

Headway analysis

The headway analysis results from the Brown and Purple Line Sedgwick incident are summarized in Figure 3-8. The shade around normal day lines indicates ±standard deviation (same for all the following figures with shades around normal day lines). The line-level headway is calculated as Eq. 3.8. We selected three lines with directions of interest to analyze. Recall that the Brown and Purple Lines share tracks in the incident area, while the Red Line runs on separate tracks in the incident area but shares tracks further north of the line.



(a) Brown Line (southbound) (b) Purple Line (southbound) (c) Red Line (southbound)

Figure 3-8: Headway temporal distribution (Brown and Purple Lines Sedgwick incident). The shade around normal day lines indicates ±standard deviation (same for all following figures)

A rise in southbound headways for both the Brown and Purple Lines are observed (Figures 3-8a and 3-8b) and the changes are significant (i.e., beyond the two standard deviation ranges). This is as expected because these two lines are blocked due to the incident. On average, headway increases from 5 minutes to 15 minutes in the Brown Line and from 5 minutes to 7 minutes in the Purple Line, implying a reduction of service frequency by 66.7% and 28.6% for the Brown and Purple Lines, respectively. The Brown Line experiences a continuous increase in headways towards the end of the incident. And we see a decrease in headways once the incident clears. The Purple Line, which has most of its local stops farther away from the incident area, has less disrupted service at the line level, despite sharing tracks with the Brown Line. So its headways deviated less from the normal-day average.

As shown in Figure 3-8c, the Red Line experiences little deviation from its normal day service for the first half of the incident (before 9:30 AM), largely because it does not share tracks at the incident location and could run largely uninterrupted. However, halfway through the incident, there is a headway increase spike. This could be caused by two possible reasons: 1) Because of the bad service on Brown and Purple Lines, passengers chose to take the Red Line southbound instead, leading to more passengers and thus the delays at the stations when loading and unloading passengers. 2) The unusual operation (e.g., short-turn) of the trains on Brown and Purple Lines may occupy facilities in the Red Line, resulting in congestion and longer headways. The headway increase in nearby lines implies that the transit operator should pay attention to both incident lines and nearby lines to better serve passengers.

Passenger flow analysis

Passenger flows can be examined at multiple levels, including system-wide, line level, and station level. Figure 3-9 shows the total number of tap-in passengers for the bus and rail systems during the Brown and Purple Lines incident. The results show that there is no significant difference between the incident day and normal days for both bus and rail because the demand lines on the incident day are within the ±2 standard deviation range. This implies that though the incident lasted for more than 1 hour and blocked several stations, the impact on the whole system demand is still negligible (i.e., as influential as the inherent demand variations).

The line-level demand changes for the Brown and Purple Line incident are shown in Figure 3-10. As expected, demand on the Brown and Purple Lines (interrupted by the incident) both decreased during the incident and returned to normal after the

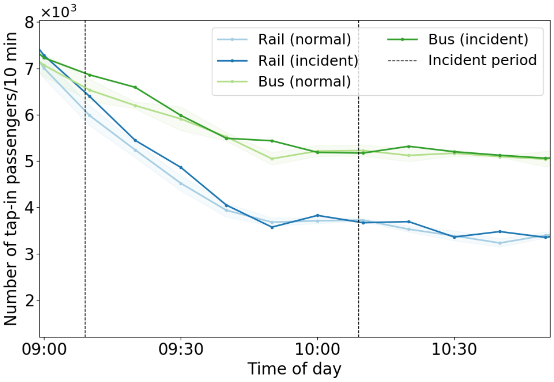
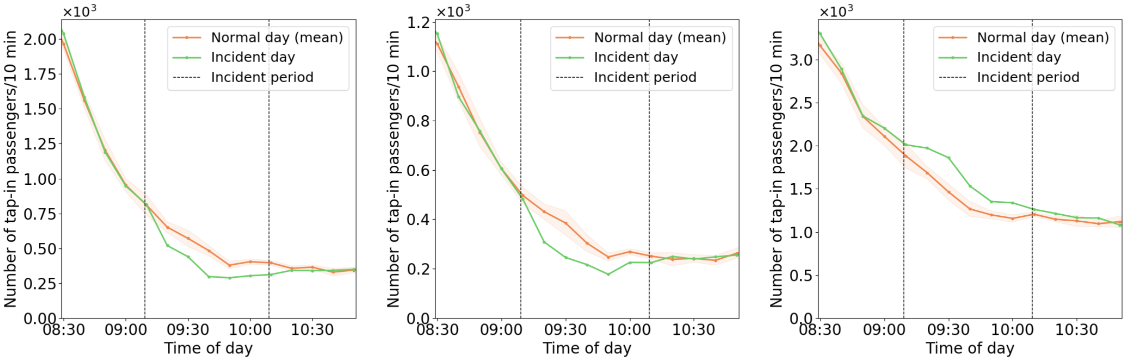


Figure 3-9: System level passenger flow analysis (Brown and Purple Lines incident).

incident. And the decrease is significant. As the Red Line runs adjacent to the Brown and Purple Lines for a significant portion and is not suspended, we see a significant increase in demand during the incident period with a return to normal after the incident is over.



(a) Brown Line (blocked) (b) Purple Line (blocked) (c) Red Line (open)

Figure 3-10: Line level passenger flow analysis (Brown and Purple Lines incident)

We further examine the demand changes at rail and bus stations close to the incident area (shown in Figure 3-11). During the incident, we see an increase in rail demand at Fullerton and Belmont stations that have direct connections to the uninterrupted Red Line. We also see clusters of increased bus demand near the incident lines. Of note are the clusters outlined in red and blue squares. The red clusters represent increased bus demand proximal to blocked stations. These passengers may have transferred directly to nearby bus stops from the blocked line. Additionally, the blue clusters represent increases in bus demand for routes that connect directly to downtown. The increase may be attributed to passengers who live in nearby neighborhoods and change to buses during the incident.

The total decrease in the number of tap-in passengers in the Brown and Purple Lines is 1,141, while the increase in nearby bus stations and the Red Line is 696 and 1,414, respectively. The demand decrease in the Brown and Purple lines is smaller than the corresponding increase in the Red Line and bus stations. This is probably because some passengers may first tap in the Brown and Purple Lines and then leave (this phenomenon will be illustrated in Figure 3-12), which leads to the underestimation of demand decrease in the Brown and Purple Lines. For all the 2,110 observed passengers using the alternative services, around one-third of them (696) transfer to buses and two-thirds (1,414) to the Red Line. Note that there may also be many passengers with direct transfers without leaving the system, which cannot be observed from the AFC data.

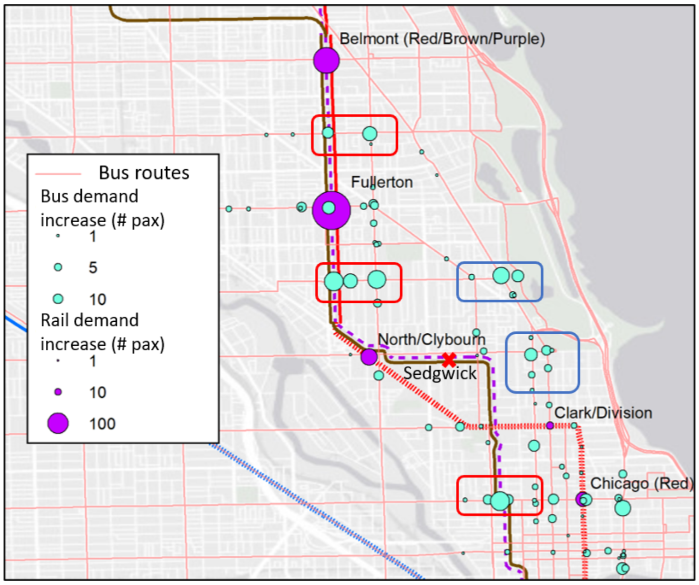
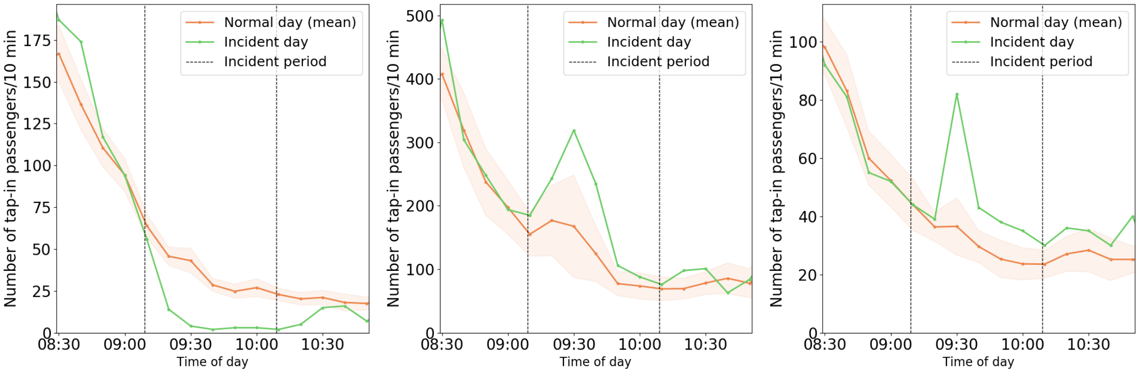

Belmont (Red/Brown/Purple)
Fullerton
North/Clybourn
Sedgwick
Clark/Division
Chicago (Red)
Bus routes
Bus demand increase (# pax)
1
5
10
Rail demand increase (# pax)
1
10
100

Figure 3-11: Station demand increase patterns (Brown and Purple Lines incident)

Additionally, Figure 3-12 shows the temporal demand distribution at three stations: Sedgwick (the incident station), Fullerton (a nearby partially blocked station), and North/Clybourn (a nearby station in the Red Line that is open). The illustrated trends align with the incident pattern. At Sedgwick (Figure 3-12a), the center of the incident, we see a drastic decrease in demand once the incident starts. As some passengers may not be aware of the incident and accidentally tapped into the station, the demand is not zero during the incident period. After the incident is over, we see a quick recovery in demand. In terms of the Fullerton (Figure 3-12b) station, despite it being partially blocked (the tracks of the Brown/Purple Lines are blocked but the tracks of the Red Line are not), we see an immediate rise in demand. This indicates that passengers used Fullerton station for the Red Line. Lastly, we see a sharp increase in the number of tap-in passengers at the North/Clybourn station in the Red Line (Figure 3-12c), which is within walking distance from Sedgwick station. This implies that passengers from the Brown and Purple Lines may also walk to the Red Line to finish their journey. Additionally, this station gives passengers access to Fullerton station, where they can switch to Brown or Purple Lines trains going northbound. The sharp increase may represent the first wave of transfer passengers.

The demand analysis is helpful for transit operators to identify passengers' choices, supplement transit on other lines, and inform passengers of better alternatives.



(a) Sedgwick (incident station, blocked)

(b) Fullerton (partially blocked)

(c) North/Clybourn (Red Line, open)

Figure 3-12: Station level passenger flow analysis (Brown and Purple Lines incident)

3.5.2 Blue Line incident analysis

Redundancy index

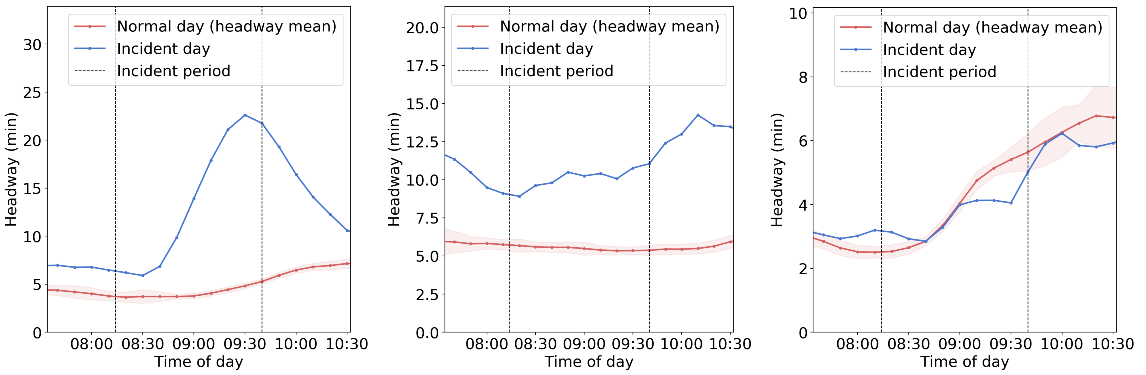
The NURI (Eq. 3.4) for the Blue Line Jefferson Park case is 0.093, meaning that the transit system maintains 9.3% transporting capacity during the incident period. The relatively low redundancy, in this case, is due to the lack of alternative rail lines. Though there are some nearby bus services (see Figure 3-16), the capacity of buses is much lower than that of the metro lines. Also, most of the bus routes are not directly connected to downtown, which increases the travel time for passengers using buses. The low NURI indicates that during the Blue Line incident, CTA needs to provide more alternative services, such as dispatching shuttle buses, increasing the frequency of substitutional bus routes.

Headway analysis

The headway analysis results from the Blue Line incident are shown in Figure 3-13. Looking at the Blue Line southbound (Figure 3-13b), the headway was a little bit longer than usual at the start of the day for unknown reasons. As the incident starts, the headway increases immediately for southbound trips. The increase is steeper before 9:30 AM, which is understandable since before that time CTA was working on changing the system to single-track operation. Once the single-track operation successfully deployed for all southbound trains, the headway plateaued, and then gradually decreased after 9:30 AM. On average, headway increases from 7 minutes to 17 minutes in the Blue Line southbound, indicating a 58.8% reduction in service frequency.

Figure 3-13a shows the headway change for the Blue Line northbound. Similarly, the headway was a little bit longer than usual at the start of the day. As the incident started, headways gradually increased. However, though the single-track operation starts at 9:30 AM, the northbound headway still remains higher than normal. This may be because CTA allowed more southbound trains to cross the single track area as they serve the major demand in the morning peak, which caused delays for the northbound trains. On average, headway increases from 8 minutes to 12 minutes in the Blue Line southbound, indicating a 33.3% reduction in service frequency.

The headway for the Brown Line southbound is also shown in Figure 3-13c as the Brown Line may be a possible alternative for passengers in the south part of the Blue Line. The headway remains relatively unchanged throughout the Blue Line incident, which means the incident did not affect the Brown Line operations.



(a) Blue Line (southbound) (b) Blue Line (northbound) (c) Brown Line (southbound)

Figure 3-13: Headway temporal distribution (Blue Line incident)

Passenger flow analysis

We first look at the system level demand change during the Blue Line incident in Figure 3-14. Similar to the results from the Brown and Purple Lines incident, there is no significant difference between incident day and normal days for both bus and rail systems because the incident demand is within the 2 standard deviation range, implying that the incident did not significantly change the demand patterns for the whole system.

The demand patterns of the Blue, Brown, and Red Lines during the Blue Line incident are shown in Figure 3-15. The Blue Line (Figure 3-15a) initially experiences a drop in the number of tap-in passengers immediately after the incident, which is as expected because passengers were informed of the incident and chose to not tap in. As the single-track operation started, the number of tap-ins gradually returned to regular levels as the system's backlog slowly began to clear. By 9:40 AM, single

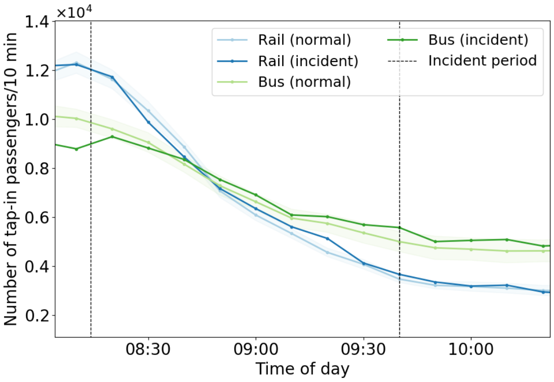
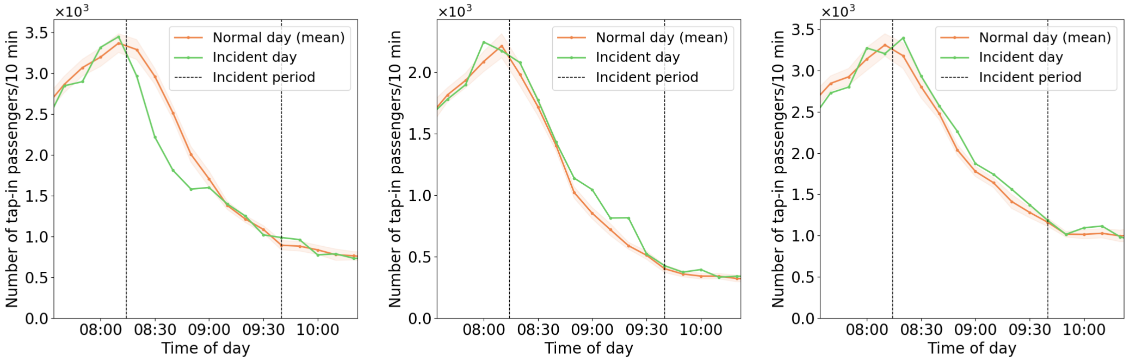


Figure 3-14: System level demand analysis (Blue Line incident)

tracking is in full operation. Hence, the number of tap-in passengers is closer to average.

For the Brown Line (Figure 3-15b), we see a slight spike of demand about 30 minutes after the incident. This is because the Brown Line is not within the walking distance from the Blue Line. Passengers need to take the eastbound bus routes and then transfer onto the Brown Line to continue their journeys, which takes around 30 minutes. We also observe a consistent (though not significant) demand increase in the Red Line for the entire major incident period (Figure 3-15c). The reason may be that Red and Brown Lines are largely overlapped near the incident area and can both be alternatives for the Blue Line.



(a) Blue Line (blocked)

(b) Brown Line (open)

(c) Red Line (open)

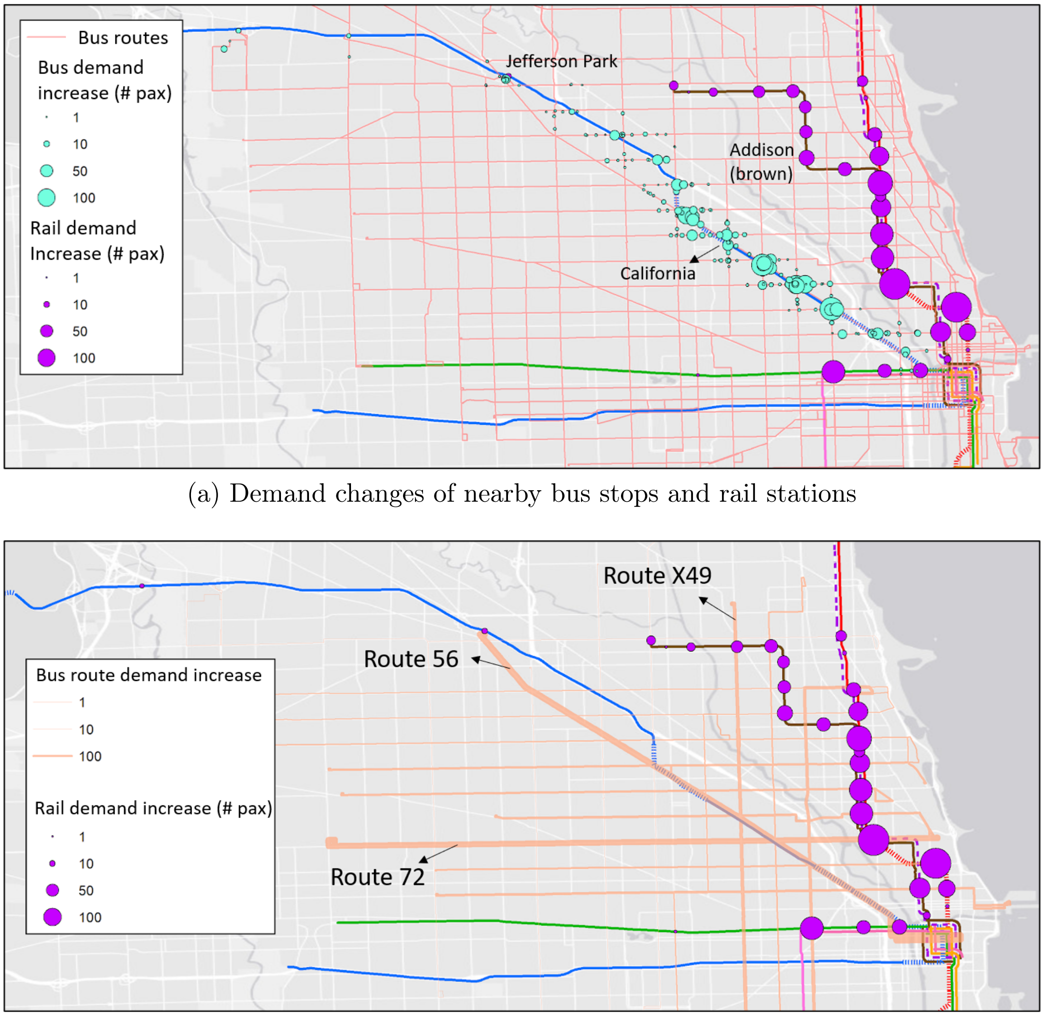
Figure 3-15: Line level passenger flow analysis (Blue Line incident)

Demand changes at individual rail stations and bus stops near the incident area are shown in Figure 3-16. Figure 3-16a shows that demand rises at the bus stations that are close to the Blue Line, which means many passengers switched to bus services during the incident. We also observe a substantial increase in ridership on the nearby Brown and Red Lines. This is presumably from passengers taking buses from the Blue Line and transferring to the Brown and Red Lines. However, we see little increase in ridership on the Green Line in comparison. Since the Green Line is close to the Blue Line and provides service to downtown as well, it should be a good alternative. But the small number of passengers using it implies that some passengers did not make good choices.

Figure 3-16b illustrates the demand changes for nearby bus routes. The demand for several bus lines increased, with routes 56, 72, and X49 being the top 3. Route 56 demand increased most because it runs parallel to much of the Blue Line and connects directly to downtown. The increase in route 72 may be due to passengers transferring to that route and using it to connect to the Brown and Red Lines. Since there is little increase in the Green Line where the X49 connects, most of the increased ridership in route X49 was probably passengers with destinations in the south that route X49 directly serves.

The total decrease of the number of tap-in passengers in the Blue Line is 2,219, while the increases in nearby bus stations, Brown Line, and Red Line are 2,426, 845, and 1,125, respectively. It is worth noting that passengers may tap in the Blue Line then get out to use buses due to long waiting times. This implies that the actual demand decrease in the Blue Line is larger than 2,219. For all passengers using nearby bus stations (2,426), most of them (845+1,125) transferred to Brown and Red Lines.

Further analysis can be done on specific key stations in terms of temporal demand patterns. Figure 3-17 summarizes the demand changes at the Jefferson Park (incident station, partially blocked) (Figure 3-17a), California, (partially blocked) (Figure 3-17b), and Addison (Brown Line, open) (Figure 3-17c) stations. At Jefferson Park, the number of tap-in passengers does not show a significant difference compared to that of normal days (i.e., within two standard deviations). Possible reasons are 1)

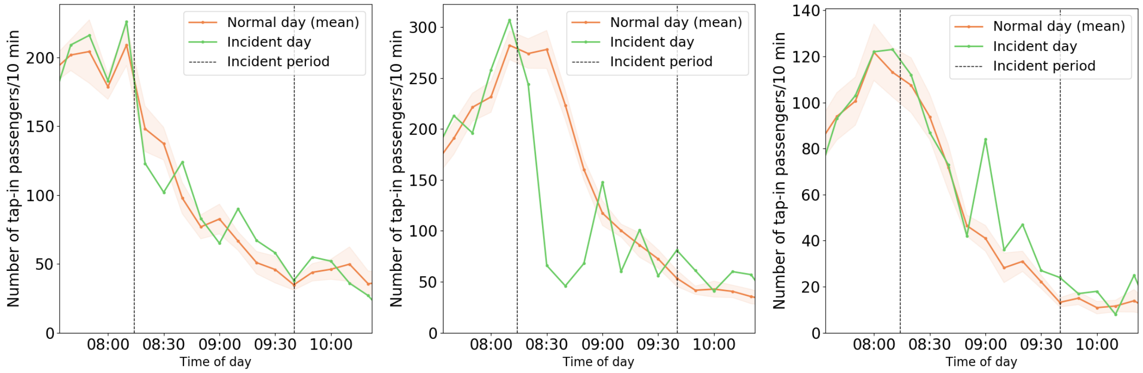

Bus routes
Bus demand increase (# pax)
1
10
50
100
Rail demand Increase (# pax)
1
10
50
100
Jefferson Park
Addison (brown)
California
Bus route demand increase
1
10
100
Rail demand increase (# pax)
1
10
50
100
Route X49
Route 56
Route 72

(a) Demand changes of nearby bus stops and rail stations

(b) Demand changes of nearby bus routes

Figure 3-16: Station and bus route demand increase patterns (Blue Line incident)

passengers were not well informed of the incident and entered the station during the service disruption; 2) there are not enough alternative services for passengers at the Jefferson Park station, so passengers chose to enter the station and wait for service. At the California station, we see a huge drop-off in ridership before 9:00 AM. This may be due to the fact that California is closer to downtown and has more bus options for riders, which corresponds to the results in Figure 3-16a. As the single track operations stabilize (around 9:00 AM), we see an increase in the number of tap-ins. Lastly, looking at Addison (Figure 3-17c), we observe normal ridership during the first part of the incident. Halfway through, a large spike in ridership takes place. This is most likely explained by Blue Line riders taking a bus to the Brown Line, as outlined in Figure 3-16a. And the spike is due to the fact that they arrived as a group.



(a) Jefferson Park (incident station, partially blocked) (b) California (partially blocked) (c) Addison (Brown Line, open)

Figure 3-17: Station level passenger flow analysis (Blue Line incident)

3.5.3 Individual passenger choice analysis

To analyze the individual-level passenger choices, we sampled 1,060 regular passengers who are affected by the incident (see Section 3.3.2 for method details) using the AFC data from the two incidents above, 533 of which are from the Brown and Purple Line incident case and 527 from the Blue Line incident case. Table 3.3 provides descriptive statistics related to various variables of interest. All transaction-related variables (such as total added value and total add-value times) are calculated based on smart card transaction data from January to December 2019.

The estimation results of the binary logit model are shown in Table 3.4. “Other” is set as the base travel mode. We observe that passengers with larger total added value and those who use a pass (as opposed to pay-as-you-go) are more likely to choose CTA during the incident. This is understandable because these passengers generally use the public transit system more frequently. They are familiar with the service and able to find alternative public transit routes during the incident. Passengers who live in high household income areas and have high max single added value are less likely to choose CTA (both are significant at 0.15 level). Note that both of

Table 3.3: Descriptive statistics of samples

Variables	Mean	Standard deviation
Total added values ($/year)	917.7	367.8
Add-value frequency (times/year)	31.38	32.76
Max single added value ($)	65.99	37.61
Living in high household income area[1] (Yes = 1)	0.093	0.291
Living in low household income area[2] (Yes = 1)	0.013	0.114
Using pass[3] (Yes = 1)	0.374	0.484
Reduced fare status (Yes = 1)	0.087	0.281
OD-based redundancy	0.929	0.237
Downtown destination (Yes = 1)	0.635	0.482

Number of observations: 1,060 (533 from Brown Line case and 527 from Blue Line case)
Choices: CTA: 268; Other: 792
[1]: Living in areas where the median annual household income is greater than $120,000
[2]: Living in areas where the median annual household income is less than $25,000
[3]: The fare type is “pass” on the incident day

these two variables are used as proxies for the high income. Hence, their choice of other options may be because they can afford alternative modes of transportation (such as Uber/Lyft). Passengers with reduced fare status are more likely to use CTA services. The reason may be that reduced fare status users are usually students, seniors, and disabled people likely on limited incomes. They usually rely primarily on CTA to travel. OD-based redundancy has a positive impact on choosing CTA, which is as expected because higher redundancy indicates better alternative public transit services. Another interesting result is that passengers with the destinations in the downtown area are less likely to use CTA. This may be because these passengers were going to work and they have a higher motivation to arrive on time, thus changing to alternative modes (such as Uber/Lyft).

We also evaluate the sensitivity of the probability of choosing CTA with respect to the OD-based redundancy (Figure 3-18). The probabilities in Figure 3-18 are calculated by fixing the remaining variables to the corresponding sample means. Similar to the results above, low-income passengers have a higher probability of using CTA than that of high income, and the difference increases with the increase in redundancy. This implies that low-income passengers have higher elasticity with respect to redundancy. Assuming OD-based redundancy equal to 0.5, a 1% increase in OD-

Table 3.4: Individual choice model estimation results

Parameters	Value (standard error)	
CTA: ASC	-3.27 (0.522)	***
CTA: Total added value (\$1000/year)	1.26 (0.277)	***
CTA: Add-value frequency (100 times/year)	-0.449 (0.365)	
CTA: Max single added value (\$1000)	-5.89 (3.72)	·
CTA: Living in high household income area (Yes = 1)	-0.396 (0.269)	·
CTA: Living in low household income area (Yes = 1)	1.380 (0.568)	**
CTA: Using pass (Yes = 1)	1.13 (0.215)	***
CTA: Reduced fare status (Yes = 1)	0.627 (0.297)	**
CTA: OD-based redundancy	1.26 (0.432)	***
CTA: Downtown destination (Yes = 1)	-0.335 (0.159)	**
Other: ASC	0 (fixed)	

Number of individuals: 1060. Adjusted $\rho^2 = 0.245$

***: $p < 0.01$; **: $p < 0.05$; *: $p < 0.1$; ·: $p < 0.15$

based redundancy can lead to a 0.03% increase in the probability of choosing CTA for high-income passengers, and a 0.11% increase for low-income passengers.

Understanding the impact of demographics on travel mode choices is helpful for transit operators to customize their operation strategies during the incident. For example, as low-income passengers are more likely to use CTA during the incident, alternative services can be provided to serve low-income areas first.

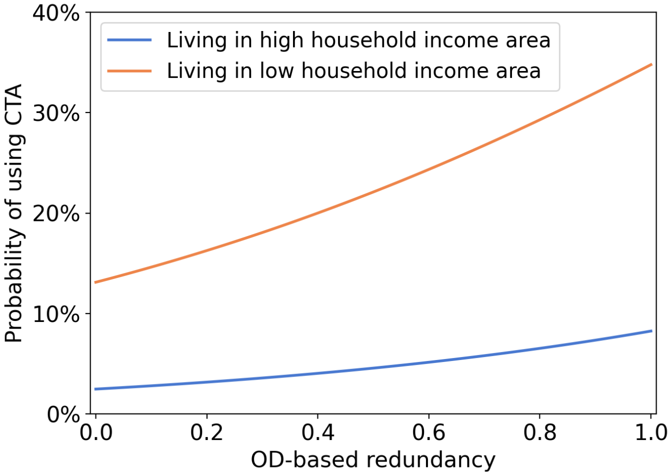


Figure 3-18: Impact of OD-based redundancy for passengers living in high and low income areas

3.6 Conclusion and Discussion

3.6.1 Conclusion

This study proposes a general incident analysis framework both from the supply and demand sides using automatically collected data (AFC and AVL) in public transit systems. Specifically, from the supply side, we propose an incident-based network redundancy index to analyze the network's ability to provide alternative services under a specific rail disruption. The impacts on service operations are analyzed through the headway changes. From the demand side, we calculate the demand changes at different rail lines, rail stations, bus routes, and bus stops to understand the passenger flow redistribution under incidents. Individual behavior is analyzed using a binary logit model based on inferred passengers' mode choices and socio-demographics inferred from AFC and sale transaction data. Two incidents in the CTA public transit system are used as case studies. The two rail disruption cases have different attributes, one at a location with high network redundancy and the other with low network redundancy.

Results show that the service frequency of the incident line was largely reduced during the incident time. Nearby lines with substitutional functions are also slightly affected. Depending on the incident location, the network's redundancies are different, as well as the passengers' behavior. In the low redundancy scenario, most of the passengers chose to use nearby buses to move, either to their destinations or to the nearby rail lines. In the high redundancy scenario, most of the passengers transferred directly to nearby rail lines.

3.6.2 Policy implications and suggestions

The results of the case study provide useful insights into operations when dealing with incidents. We summarize the main policy implications below.

Planning for incident responses using redundancy index. In Section 3.4.2, we calculate the NRUI for different stations by assuming a one-hour track-block incident. The NRUI can be adapted to different types of incidents, network blockages, and duration. Based on a graph similar to Figure 3-5, transit operators can better plan for future incidents, such as planning alternative services for low-redundancy locations, preparing route recommendation strategies for high-redundancy locations, etc.

Headway management for both the incident line and nearby lines. In Sections 3.5.1 and 3.5.2, we observe that headways increase in both the incident line and nearby lines. The results suggest that transferred passengers from the incident line and unusual operations of the incident line may also affect operations of nearby lines. More comprehensive headway management should be considered during incidents.

Provision of timely customer information. The results indicate that passengers tap into the blocked station during the incident, implying that these passengers are not well informed. Transit agencies should improve their customer information delivery during incidents (especially at fare gates). This can be done through text messaging, Twitter, in-station signs, station staff, and a variety of other methods to keep the passenger informed.

Provision of route recommendations during incidents. During the Blue Line incident, not many passengers use the Green Line, although it is a good alternative (see Section 3.5.2). This suggests that passengers may not act rationally, or they lack knowledge about the available alternatives. Providing route recommendations to passengers during the incident can increase the utilization of alternative services and improve the level of service.

Data-driven methods to design alternative services. The analysis provides a better understanding of how passengers move and the alternatives they may choose, based on which operators can better allocate available buses or trains. For example, most passengers used Bus Route 56 as a substitutional service during the Blue Line incident (Section 3.5.2). CTA may increase the service frequency of these heavily used routes.

Provision of shuttle services to improve the use of alternative routes.

During the Blue Line incident, one of the reasons that the Green Line is not fully utilized may be that it is not directly connected to the Blue Line (Section 3.5.2). Hence, CTA may provide shuttle services to connect the Blue and Green lines to encourage more passengers to follow the recommendation (note that multiple recommendations should be provided to avoid overwhelm of a specific line).

3.7 Appendix

3.7.1 Proof of $\lim_{D_I \to \infty} A_p$

Let k^* be the last trip of path p that can reach the destination during the incident period. Mathematically, $k^* = \arg\min_k \{k = 1, 2, ..., \lfloor D_I/H_p \rfloor \mid D_I - (k-1)H_p \leq L_p\}$. Therefore, we have

$$\min\{D_I - (k-1)H_p, L_p\} = L_p \qquad \forall k \leq k^* \tag{3.12}$$

$$\min\{D_I - (k-1)H_p, L_p\} = D_I - (k-1)H_p \qquad \forall k^* < k \leq \lfloor D_I/H_p \rfloor \tag{3.13}$$

This leads to

$$\begin{aligned}
\lim_{D_I \to \infty} A_p &= \lim_{D_I \to \infty} \frac{1}{D_I} \sum_{k=1}^{\lfloor D_I/H_p \rfloor} \frac{\min\{D_I - (k-1)H_p, L_p\}}{L_p} \cdot C_p \\
&= \lim_{D_I \to \infty} \sum_{k=1}^{k^*} \frac{L_p}{L_p \cdot D_I} \cdot C_p + \lim_{D_I \to \infty} \sum_{k=k^*+1}^{\lfloor D_I/H_p \rfloor} \frac{D_I - (k-1)H_p}{L_p \cdot D_I} \cdot C_p \qquad (3.14) \\
&= \lim_{D_I \to \infty} k^* \frac{1}{D_I} \cdot C_p + \lim_{D_I \to \infty} \sum_{k=k^*+1}^{\lfloor D_I/H_p \rfloor} \frac{D_I - (k-1)H_p}{L_p \cdot D_I} \cdot C_p
\end{aligned}$$

Notice that $k^* \to \lfloor D_I/H_p \rfloor$ as $D_I \to \infty$ because when D_I is large enough, almost all trips can reach the destination. And $\lim_{D_I \to \infty} \lfloor D_I/H_p \rfloor = D_I/H_p$ by definition. Therefore,

$$\lim_{D_I \to \infty} A_p = \lim_{D_I \to \infty} \lfloor D_I/H_p \rfloor \cdot \frac{1}{D_I} \cdot C_p + 0 = \frac{C_p}{H_p} \tag{3.15}$$

Chapter 4

Inferring passenger behavioral responses under disruptions

4.1 Introduction

Urban rail transit plays an important role in urban mobility. However, with aging systems, continuous expansion, and near-capacity operations, service disruptions often occur. Disruptions can range from short-term delays at some stations to shutdowns of entire subway lines over an extended period. These incidents may result in delays and cancellation of thousands of trips as well as economic and opportunity losses [6].

Consequently, there is growing research interest and literature in the area of rail disruption analysis and management. These efforts can be classified into two types: supply-oriented and demand-oriented [135]. The supply-oriented research focuses on analyzing the network vulnerability and improving network resilience from the supply and operation perspectives. The literature in rail transit network vulnerability based on complex network theory is very intensive. It explores the vulnerability of network topology when some nodes or links of the network are failed. Degree, betweenness, centrality measures, and connectivity methods are usually used [136, 137, 138, 139]. From the operations point of view, many studies look at adjusting the timetable [140], managing rolling stock [141], and designing shuttle buses [91] during urban rail disruptions to ensure operational feasibility and improve system efficiency.

Demand-oriented research focuses on understanding and modeling passengers' behavior under rail disruptions. Transit users' behavior can be significantly different in the event of service disruptions and vary depending on the stage of the trip at the time of the disruption [85]. A better understanding of passengers' behavior in the event of disruption is important for operators to recommend alternative routes, adjust the capacity of rail lines, and provide shuttle services [142]. However, nearly all of the previous research investigated passenger behavior using survey-based methods [121, 84, 82]. For example, Lin et al. [85] used a joint revealed and stated preference (SP) survey to estimate transit user mode choice in response to a transit service disruption in the City of Toronto. Rahimi et al. [106] utilized survey data collected in the Chicago Metropolitan Area to analyze how transit users respond to unplanned service disruptions and the factors that affect their behavior. Survey-based methods are usually time-consuming and labor-intensive. Besides, SP surveys require passengers to respond to hypothetical situations, which may not reflect the actual travel choices of passengers [89].

Recently, thanks to the widely adopted automated fare collection (AFC) system, passengers' travel information is recorded in the AFC data, providing opportunities to capture individual choices under rail disruptions using data-driven approaches. However, studies using AFC data to explore the impact of unplanned disruptions on individual responses are limited. Silva et al. [143] proposed a method to analyze large-scale mass transportation systems during unplanned disruptions. They estimated the disruption effects on passenger volumes during incidents using smart card data. van der Hurk [144] developed a model based on smart cards to forecast the route choices of passengers impacted by disruptions under different scenarios. The study shows that operators can help passengers minimize their overall inconvenience by providing individual advice. Sun et al. [89], using AFC data, estimated three groups of passengers (leaving the system, detouring, and continuing to travel) during the rail disruption. Recently, Liu et al. [117] also proposed a data-driven approach to evaluate disruption impacts on system performance and individual responses in urban railway systems using AFC data. They considered four groups of passengers: performing trips, changing travel time, changing stations, and changing modes.

However, there are some limitations in the previous studies. First, the approaches of identifying passenger responses in previous studies are rule-based and deterministic, meaning that they directly map the observed AFC records to a specific response behavior. The rule-based method ignores uncertainty and randomness in passengers' behavior (i.e., the observed AFC records may be due to behavior randomness, rather than the impact of incidents), which may introduce estimation bias. Also, deterministic methods cannot quantify the uncertainty (i.e., variance) in the estimated results. Second, most of the previous studies are based on data from closed AFC systems with both tap-in and tap-out information, which does not apply to many open transit systems where only tap-in information is available (such as the transit systems in Chicago, Boston, and New York). Third, most of the previous studies only considered three or four possible response behaviors. In this chapter, we show that passenger's responses are diverse depending on where they are when the incident happens. There are 19 possible responses identified in this study.

To fill the research gap, this chapter proposes a probabilistic passenger behavior estimation framework under rail disruptions using tap-in-only AFC data. The historical travel trajectories before the incident and the subsequent travel records after the incident are both used for inference and capturing the uncertainty in passengers' behavior. We first identify 19 possible response behaviors that passengers may have based on their decision-making times and locations[1] (i.e, the stage of their trips when an incident happened), including transferring to a bus line, canceling trips, waiting, delaying departure time, etc. A statistical inference model is proposed to estimate the mean and variance of the number of passengers in each of the 19 behavior groups using passengers' AFC data. The urban bus and rail system operated by the Chicago Transit Authority (CTA) is used as a case study. The proposed model is validated with a synthetic data set and applied using an actual data set from CTA. Results show that the proposed model can estimate passengers' travel behavior after the rail

[1] The proposed model is not restricted to the 19 behaviors. The way of recognizing possible responses is general and can be extended to different case studies. See Section 4.2.1 for details.

disruption accurately and outperform the rule-based benchmark model.

The identified 19 behavioral responses can be classified from two aspects. From the behavioral aspect, they can be grouped into 5 aggregated response behaviors including using bus, using rail (changing or not changing route), not using public transit, and not being affected. These five aggregated response behaviors are general and applicable for the incident analysis for any other public transit system. From the methodological aspect, the inference of the 19 behaviors can be classified into four cases based on the information used (historical trips vs. subsequent trips) and the context of the observed transactions (direct incident-related vs. indirect incident-related).

The main contributions of the chapter are as follows:

- Provide a comprehensive framework of passengers' behavior under service disruptions. A total of 19 possible behavior groups for passengers at different stages of their trips are considered, which enables a more detailed modeling framework. The behavior identification is based on when and where passengers are making their decisions during a disruption. The method is general and can be used for other transit systems (the resulting possible behaviors may vary according to the context of the system, i.e., not necessarily 19)

- Propose a probabilistic behavior inference model with a specific formulation for each of the 19 behavior groups. The model enables the estimation of the mean and variance of the number of passengers in each group to capture passenger's behavior uncertainty. To the best of the authors' knowledge, this is the first article providing the estimation for both mean and variance of post-incident behaviors using AFC data.

- Leverage both passengers' historical travel trajectories and their subsequent tap-in records after the incident to facilitate behavior inference. This is contrary to previous studies where only the AFC data on the incident day is used.

The rest of the chapter is organized as follows. Sections 4.2 and 4.3 present the methodology of this study. Section 4.4 discusses the case study for model application and the corresponding results. Section 4.5 concludes the chapter and discusses future research directions.

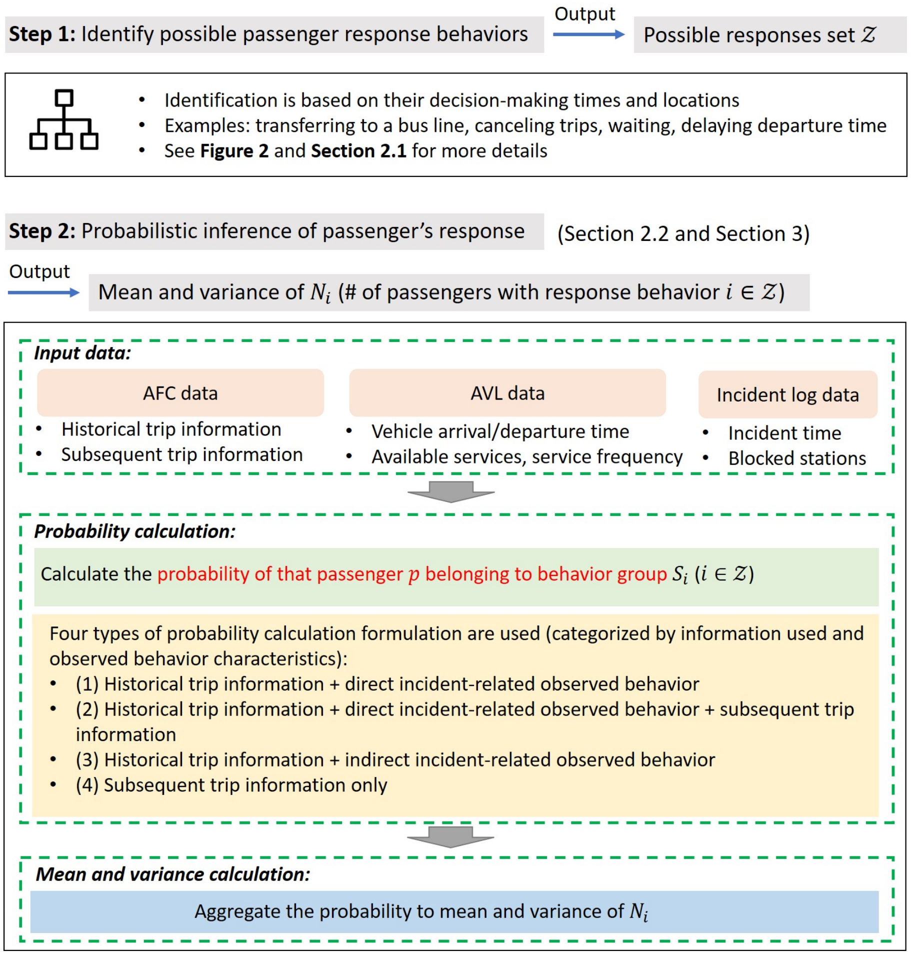
4.2 Model framework

Figure 4-1 shows an overview of the model framework. There are two steps for inferring passenger's responses. At the first step, we aim to identify all possible passenger response behaviors to the incident based on their decision-making times and locations. Details of step 1 are shown in Section 4.2.1. At step 2, based on the results of step 1, we aim to associate each passenger to a specific response behavior by calculating the corresponding probabilities based on the observed passenger AFC records and his/her travel histories. The input data for the inference include AFC, AVL (automated vehicle location), and incident log. There are four different formulations for the probability calculation, which are categorized by the used information and properties of observed AFC records. Then, we aggregate the probabilities to the mean and variance of the number of passengers in the different response behavior groups. Details of the step 2 are shown in Sections 4.2.2 and 4.3.

4.2.1 Passenger behavior under disruptions

A prerequisite for behavior inference is to identify possible options passengers may have during the disruption. According to Sun et al. [89], passenger responses to a service disruption are generally triggered when the delay time is long enough (e.g., greater than 30 minutes). Hence, for a meaningful analysis, this study focuses on substantial unplanned service disruptions (i.e., blockage or shutdown of service as opposed to reduced capacity or frequency) so that there are observable behavior changes.

For an incident beginning at T_1 and ending at T_2, we consider the analysis time period as $[T_s, T_e] = [T_1 - \delta_1, T_2 + \delta_2]$, where δ_1 is set as the maximum travel time in the system because all passengers tapping in before $T_1 - \delta_1$ are not affected. δ_2 is the recovery time for the system after the incident ends, which can be pre-calculated



Step 1: Identify possible passenger response behaviors

Output

Possible responses set $\mathcal{Z}$

- Identification is based on their decision-making times and locations
- Examples: transferring to a bus line, canceling trips, waiting, delaying departure time
- See **Figure 2** and **Section 2.1** for more details

Step 2: Probabilistic inference of passenger's response

(Section 2.2 and Section 3)

Output

Mean and variance of N_i (# of passengers with response behavior $i \in \mathcal{Z}$)

Input data:

AFC data

- Historical trip information
- Subsequent trip information

AVL data

- Vehicle arrival/departure time
- Available services, service frequency

Incident log data

- Incident time
- Blocked stations

Probability calculation:

Calculate the probability of that passenger p belonging to behavior group S_i ($i \in \mathcal{Z}$)

Four types of probability calculation formulation are used (categorized by information used and observed behavior characteristics):

- (1) Historical trip information + direct incident-related observed behavior
- (2) Historical trip information + direct incident-related observed behavior + subsequent trip information
- (3) Historical trip information + indirect incident-related observed behavior
- (4) Subsequent trip information only

Mean and variance calculation:

Aggregate the probability to mean and variance of N_i

Figure 4-1: Framework of the methodology

based on the smart card data [145] (i.e., we assume that after $T_2 + \delta_2$ the system is fully recovered). We only consider passengers who were potentially affected by the incident, defined as passengers who had (or were supposed to have) tap-in records during the analysis period ($[T_s, T_e]$) on the incident day. Passengers who are supposed to tap in are those with historical trips indicating that they may have a rail trip during this period, though we do not observe them on the incident day AFC data. These passengers are considered because they may cancel their trips or use other undetected modes (details can be found in the following sections).

Figure 4-2 summarizes possible passenger behaviors under different cases. A total of 19 possible response behaviors are considered. The general approach to characterize these behaviors is elaborated on below. The approach can be applied to other public

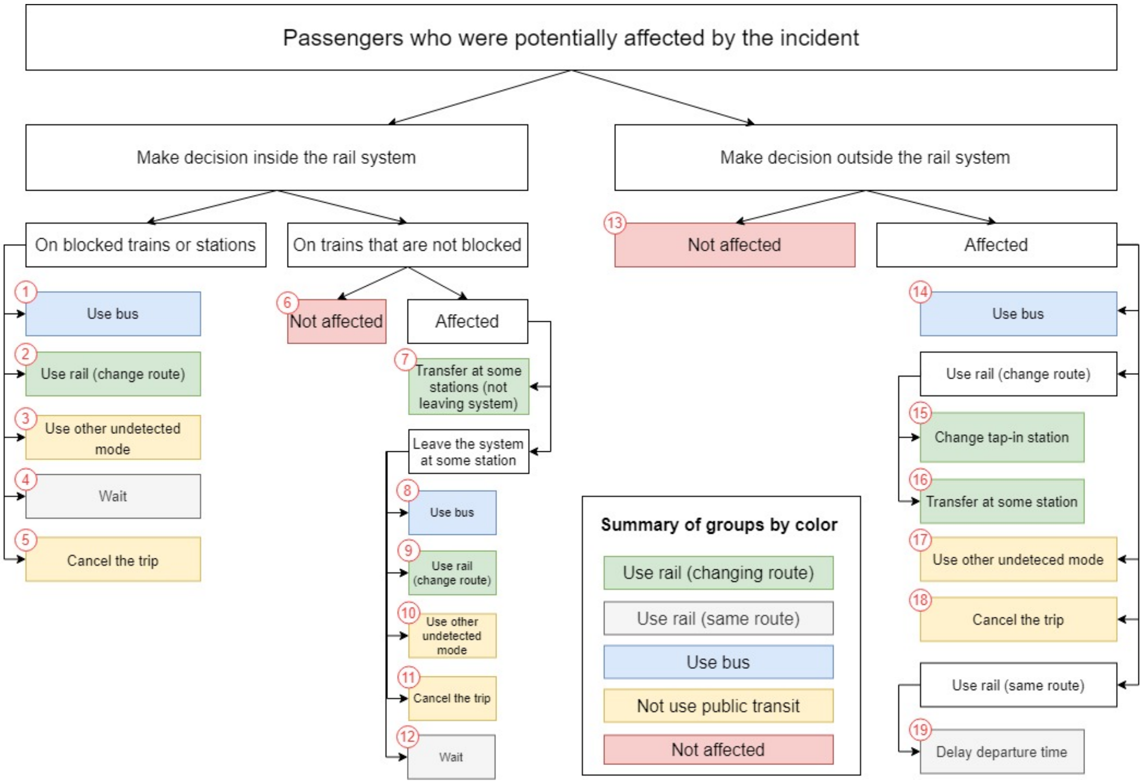


Figure 4-2: Passenger responses to an unplanned rail disruption

transit systems to identify a similar set of possible response behaviors.

Passengers' behavior may vary a lot depending on the stage of their trips at the time of service disruption [146]. Therefore, all potentially affected passengers are first divided into two groups: a) passengers in and b) out of the rail system. The first group of passengers was on a train or inside a station platform when the incident happened, while passengers in the second group have not entered the system yet (e.g., at home).

When the disruption happens, some of the stations in the rail system are blocked (i.e. trains are not allowed to move in these stations) due to the incident. Passengers who are in the blocked stations/trains are forced to leave the system. These passengers have five options: changing to a bus line, changing to another rail line or station, waiting until the system is restored, canceling the trip, or changing to other undetected modes. It is worth noting that if they choose transit services (rail or bus) again, they need to re-tap to use the alternative services. The undetected modes include Transit Network Companies (TNC), walking, bicycling, etc. It is worth noting that using a shuttle bus that was deployed to mitigate the incident impacts can be categorized into "bus" or "undetected mode" depending on whether passengers are required to tap their fare card or not. If passengers are not in the blocked stations, their trains could still move. Hence, they may not be affected by the incident. Or if they were affected, compared with passengers in the blocked stations, they would have one more option: transferring halfway to another line without leaving the system.

For passengers out of the system when the incident happens, if their travel routes on the rail system are not blocked, they are not affected. Otherwise, instead of following the original route, they may choose to use buses, use rails but change the tap-in station, use rail by transferring at a halfway station, use other undetected modes, cancel the trip, or delay their departure time until the system recovers.

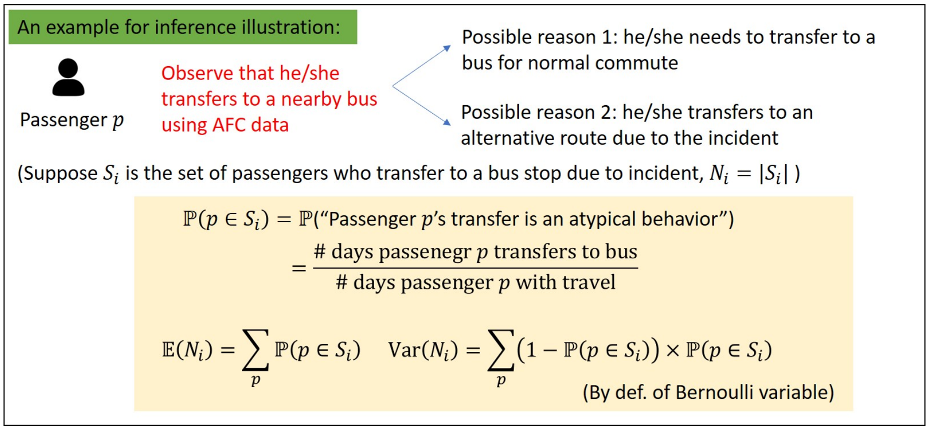
All these behaviors can be summarized into five groups: use rail (changing route), use rail (same route), use buses, not use public transit, and not being affected. Note that these five alternatives are general for different transit systems and can be used to guide the potential behavior identification. To better describe these behaviors, we assign a specific ID to each (i.e., numbers in the red circle in Figure 4-2). These behaviors are inferred separately based on their characteristics in the AFC data.

4.2.2 Probabilistic behavior inference

We propose a probabilistic framework to infer passengers in each behavior group using AFC data. The probabilistic framework facilitates the inference of whether a specific observed behavior for a passenger is due to the incident, or is typical. In this study, we focus on open public transit systems where only tap-in information is available. The AFC data include both bus and rail boarding records.

The key idea of the inference framework is to identify 1) whether an observed AFC data record (e.g., transfer to bus) is atypical or not and 2) whether the atypical behavior is owing to the incident or behavioral randomness. These two questions are answered probabilistically (i.e., obtaining the corresponding probabilities). And the corresponding probabilities are used to calculate the mean and variance of the number of passengers in each behavior group.

Figure 4-3 presents an explanatory example for the probabilistic behavior inference method. Consider a passenger p in the system. We observe that he/she has a transfer record to a nearby bus stop from the incident line. In typical rule-based method [89, 117], this passenger will be directly identified as "transferring to bus due to incident". However, in the probabilistic framework, we consider two possible reasons for this observed record: 1) he/she transfers to a bus for a normal commute. 2) he/she transfers to an alternative route due to the incident. We should only account for the second reason as the impact of incidents. Therefore, we use historical data to calculate the probability that "this transfer is an atypical behavior" (i.e., due to the incident). Then, the mean and variance of the number of passengers with a specific response behavior can be obtained from this probability (by the definition of the Bernoulli random variable).



An example for inference illustration:

Passenger p

Observe that he/she transfers to a nearby bus using AFC data

Possible reason 1: he/she needs to transfer to a bus for normal commute

Possible reason 2: he/she transfers to an alternative route due to the incident

(Suppose S_i is the set of passengers who transfer to a bus stop due to incident, $N_i = |S_i|$)

$$\mathbb{P}(p \in S_i) = \mathbb{P}(\text{"Passenger } p\text{'s transfer is an atypical behavior"}) = \frac{\text{\# days passenegr } p \text{ transfers to bus}}{\text{\# days passenger } p \text{ with travel}}$$

$$\mathbb{E}(N_i) = \sum_p \mathbb{P}(p \in S_i) \quad \text{Var}(N_i) = \sum_p (1 - \mathbb{P}(p \in S_i)) \times \mathbb{P}(p \in S_i)$$

(By def. of Bernoulli variable)

Figure 4-3: Illustration example of the probabilistic behavior inference

Notation

Denote S_i as the set of passengers who have behavior i **in response to the incident**, $i \in \mathcal{Z} = \{1, 2, ..., Z\}$ (behavior IDs are shown in Figure 4-2, for example, "Behavior 1" means offloading from the train when an incident happens and using bus to respond to the incident). Let N_{S_i} be the number of passengers in set S_i. The day when the incident happens is referred to as the **incident day**. A **normal day** is defined as a day without (substantial) incidents in the analysis period and area and with the same day of the week as the incident day. For example, if an incident happens on Friday [8:00 $\sim$ 9:00] at Line X, then a normal day can be all Fridays in the last 2 months where there are no substantial incidents occurring during [8:00-δ_1 $\sim$ 9:00+δ_2] at Line X.

Note that we use the term "no substantial incidents" due to the high frequency of various types of incidents in a public transit system and it may be hard to find an "absolute normal day" without any incidents. The selection of normal days is a trade-off between sample sizes and accuracy. A larger number of normal days can provide more observations to estimate the habitual behaviors of passengers. However, it may also include days with incidents that can introduce bias. Usually, we aim to have normal days with consistent demand and supply characteristics, and are significantly different from the incident day (as shown in the case study, Section 4.4.4).

Suppose that we have collected the AFC data of the incident day and a total of M normal days. Let $\mathcal{P}$ be the set of all potentially affected passengers, which is defined as the set of all passengers with at least one AFC data record in $[T_s, T_e]$ on the incident day or any of the M normal days. Let $\mathcal{P}^H \subseteq \mathcal{P}$ be a subset of passengers with reliable history trips and M_p be the number of normal days that passenger p has trips on ($M_p \leq M$). Then $\mathcal{P}^H = \{p \in \mathcal{P} : M_p \geq M^{\text{R}}\}$, which means passengers with more than M^{R} normal days with travel, where M^{R} is a predetermined threshold to recognize passengers with reliable history trips. In future studies, a more complicated method to define $\mathcal{P}^H$ can be explored considering the travel regularity [127].

Consider a passenger $p \in \mathcal{P}$ with a public transit trip chain { $(o_{p_1}, t_{p_1}, m_{p_1})$,$(o_{p_2}, t_{p_2}, m_{p_2})$,..., $(o_{p_{K_p}}, t_{p_{K_p}}, m_{p_{K_p}})$} within the analysis time period. o_{p_k} is the origin of the k-th trip. t_{p_k} is the start time (transaction time) of the k-th trip. And m_{p_k} is the mode of k-th trip ($m_k \in$ {rail, bus}). It holds that $T_s \leq t_{p_1} < t_{p_2} < ... < t_{p_{K_p}} \leq T_e$. We define $\mathcal{P}^F \subseteq \mathcal{P}$ as the subset of passengers with subsequent trips after the incident on the incident day, that is, $\mathcal{P}^F = \{p \in \mathcal{P} : p$ has trips after T_e on the incident day$\}$. According to previous destination estimation studies for tap-in only systems [129, 130, 131], the destination of the trip $(o_{p_k}, t_{p_k}, m_{p_k})$ can be inferred using information of the next trip $(o_{p_{k+1}}, t_{p_{k+1}}, m_{p_{k+1}})$ (i.e., the trip chain method). The basic idea is to use the next tap-in location to estimate the destination of the current trip. Hence, for $p \in \mathcal{P}^F$, we can obtain the destination of the trip $(o_{pK_p}, t_{pK_p}, m_{pK_p})$. It is worth noting that if the incident happened in the evening, we would extend $\mathcal{P}^F$ to include passengers with trips in the next morning.

As mentioned above, when a disruption happens, some of the stations in the rail system are blocked. The set of all blocked rail stations due to the disruption is denoted as $\mathcal{W}$.

The notation used in this study is summarized in Table 4.1.

Conceptual framework

We first outline the framework of the general inference model. For a specific behavior S_i, we define B_{S_i} as the set of passengers with related **observable** behavior that can be identified from the AFC data. The word "observable" indicates that $\mathbb{1}_{\{p \in B_{S_i}\}}$ is a known constant, where $\mathbb{1}_{\{\cdot\}}$ is an indicator function which returns 1 if the event is true and 0 otherwise. For example, B_{S_i} can be a set of passengers with a bus transfer trip during the incident period, or a set of passengers with a rail tap-in trip during the incident period, etc. The definition of B_{S_i} should satisfy that $S_i \subseteq B_{S_i}$. The goal is to identify S_i from B_{S_i}.

The specification of B_{S_i} depends on to what extent passengers in S_i can be observed in the AFC data. If the behavior of S_i generates many special AFC records, B_{S_i} can be defined in more detail. In this case, $|B_{S_i}|$ is relatively small, which reduces the scope for inferring S_i. On the other hand, if the behavior of S_i does not generate special AFC records, B_{S_i} can only be defined in a general way (e.g., passengers with a rail trip during the incident), bringing challenges in extracting S_i.

According to the context of S_i, there are two types of B_{S_i} regarding their relationship to the incident. For a passenger $p \in B_{S_i}$, historical information can be used to infer whether the behavior that passenger p is showing in B_{S_i} is atypical or not. However, "atypical" may not be enough to conclude whether p is affected by the incident or not. For example, B_{S_i} may be defined as passengers with a bus trip during the incident period. "atypical" only indicates the bus trip is a change of the passenger's

Table 4.1: Notation

Variable	Type	Description
Z	Constant	Total number of behaviors considered.
M	Constant	Total number of normal days considered.
N_{S_i}	Random variable	Number passengers in set S_i.
$\mathcal{P}$	Set	The set of all potentially affected passengers.
$\mathcal{P}^H$	Set	The set of passengers with reliable history trips.
$\mathcal{P}^F$	Set	The set of passengers with future trips after the incident on that day.
S_i	Set	The set of passengers with behavior i (see Figure 4-2).
B_{S_i}	Set	A set of passengers that is defined to infer S_i.
o_{p_k}	Constant	Origin of the k-th trip for passenger p within the analysis period.
t_{p_k}	Constant	Tap-in time of the k-th trip for passenger p within the analysis period.
m_{p_k}	Constant	Travel mode of the k-th trip for passenger p within the analysis period.
K_p	Constant	Total number of public transit trips for passenger p within the analysis period.
M_p	Constant	Total number of normal days passenger p has public transit trips.
T_s	Constant	Start time of the analysis period.
T_e	Constant	End time of the analysis period.
T_1	Constant	Incident start time.
T_2	Constant	Incident end time.
TT_d	Constant	Threshold to identify transfer trips for two consecutive tap-ins.
d_r	Constant	Maximum walking distance for transferring to a rail station.
d_b	Constant	Maximum walking distance for transferring to a bus.
$D(s, s')$	Constant	A function which returns the walking distance between station s and s'.
$\mathbb{1}_{\{\cdot\}}$	Constant or Random variable	Indicator function which returns 1 if the event is true and 0 otherwise.
$\mathcal{W}$	Set	The set of all blocked rail stations during the incident.
$\mathcal{W}_b$	Set	The set of bus stops within walking distance from any of the blocked stations.
$\mathcal{W}_r$	Set	The set of all unblocked rail stations within walking distance from any of the blocked stations.
$\tilde{d}_{p_k}$	Random variable	Inferred original destination for trip k for passenger p.
$\mathcal{D}_{p_k}$	Set	The set of all possible original destinations for trip k for passenger p.
$s_p(T, d)$	Constant	Location of passenger p at time T if his/her destination is d.

habitual behavior. However, the behavioral change on that particular day may be caused by many reasons, not necessarily the incident. To conclude $p \in S_i$, p's behavior needs to satisfy both "atypical" and "change is due to the incident". This type of B_{S_i} is referred to as "indirect incident-related". However, sometimes, if B_{S_i} is specified based on a lot of information related to the incident, we can infer that "atypical" is equivalent to "affected by the incident". For example, if B_{S_i} are passengers with a transfer to bus stops close to the blocked rail stations after the incident, and this behavior is atypical, we can assume this change is due to the incident because B_{S_i} is based on direct incident-related information (i.e., the transfer bus stops are close to the blocked rail stations). This type of B_{S_i} is referred to as "direct incident-related".

Besides historical information, the subsequent trips information after the incident can also be used. As mentioned before, the subsequent tap-in information can be used to infer trip destinations using the trip chain method [129, 130, 131]. Though recent studies also use historical information to infer passenger's destination [147], for the purpose of this study, only subsequent tap-in information is used as the destination estimation part is not the focus of this study. Note that the proposed probabilistic framework is quite general and any destination estimation model can be used as long as the probability of each candidate destination can be obtained (see Section 4.3.2 for details). Although passengers may have multiple path choices in rail systems [148], we assume that all passengers follow the schedule-based shortest path to simplify the formulation [149]. This assumption can be relaxed by summing over all paths with corresponding path choice probabilities in the formulation, instead of only considering a single path. For a passenger p, we obtain his/her original path in the rail system as the shortest path to the inferred destination d. Based on the characteristics of the path (explained below), we define a related event $Y_p(d)$. Since the path is known given d, $\mathbb{1}_{\{Y_p(d)\}}$ is a known constant. S_i can thus be inferred based on $\mathbb{1}_{\{Y_p(d)\}}$ (i.e. the property of the original path). For example, B_{S_i} can be a set of passengers without transfer trips during the incident period. $Y_p(d)$ can be the event that the original path for p is blocked and a transfer is not available. Then if $Y_p(d)$ is true, a passenger $p \in B_{S_i}$ can only use other undetected modes or cancel trips.

In summary, historical trips and subsequent trips after the incident are two types of available information to infer S_i. Therefore, from model formulations perspective, we can characterize the inference model in two dimensions: 1) historical trip information vs. subsequent trip information, 2) indirect incident-related B_{S_i} vs. direct incident-related B_{S_i}. We summarize the formulation in each case as follows:

- **(1) "Historical trip information + direct incident-related B_{S_i}"**: In this case, we have "atypical" = "affected by the incident". Therefore,

  $$\mathbb{E}[\mathbb{1}_{\{p \in S_i\}}] = \mathbb{1}_{\{p \in B_{S_i}\}} \cdot \mathbb{P}(\text{“Behavior atypical”} \mid p \in B_{S_i}) \tag{4.1}$$

  S_1, S_2, S_4, and S_{12} belong to this case. $\mathbb{P}(\text{“Behavior atypical”} \mid p \in B_{S_i})$ is estimated based on the context of S_i using historical trip information. Details of the formulation can be found in Section 4.3.1.

- **(2) "Historical trip information + indirect incident-related B_{S_i}"**: In this case, we need to satisfy both "atypical" and "the change is due to the incident" in order to identify S_i. Therefore,

  $$\mathbb{E}[\mathbb{1}_{\{p \in S_i\}}] = \mathbb{1}_{\{p \in B_{S_i}\}} \cdot \mathbb{P}(\text{“Behavior atypical”,“Change is due to the incident”} \mid p \in B_{S_i}) \tag{4.2}$$

  $S_5, S_{11}, S_{14}, S_{15}, S_{17}, S_{18}$, and S_{19} belong to this case. The joint probability $\mathbb{P}(\cdot, \cdot \mid p \in B_{S_i})$ is not estimated directly. Instead, we show that it can be estimated using the difference of the marginal probabilities between normal and incident days. Details of the formulation can be found in Section 4.3.3.

- **(3) "Subsequent trip information only"**: In this case, the event of path properties (as a function of the inferred destination d) can help to identify S_i.

Hence,

$$\mathbb{E}[\mathbb{1}_{\{p\in S_i\}}] = \sum_d \mathbb{1}_{\{p\in B_{S_i}\}} \cdot \mathbb{1}_{\{Y_p(d)\}} \cdot \mathbb{P}(\text{“Original destination is } d\text{” } \mid p \in B_{S_i}) \tag{4.3}$$

S_3, S_7, S_{10}, and S_{16} belong to this case. Some behavior assumptions are made when two groups are indistinguishable by the above formulation. $\mathbb{P}$(“Original destination is” $d \mid p \in B_{S_i})$ is estimated based on a destination inference model [131] with subsequent trip information. Details of the formulation can be found in Section 4.3.4.

- **(4) “Historical trip information + direct incident-related B_{S_i} + Subsequent trip information”**: This scenario is a combination of historical and future information. Hence, we combine Eq. 4.1 and 4.3:

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\{p\in S_i\}}] = \sum_d \mathbb{1}_{\{p\in B_{S_i}\}} &\cdot \mathbb{P}(\text{“Behavior atypical” } \mid p \in B_{S_i}) \\ &\cdot \mathbb{1}_{\{Y_p(d)\}} \cdot \mathbb{P}(\text{“Original destination is } d\text{” } \mid p \in B_{S_i})\end{aligned} \tag{4.4}$$

$S8$ and $S9$ belong to this case. Details of the formulation can be found in Section 4.3.2.

The above cases and the corresponding formulations are used to infer whether a specific passenger belongs to a certain group. The expected number of passengers in the group is calculated as

$$\mathbb{E}[N_{S_i}] = \sum_{p\in\mathcal{P}} \mathbb{E}[\mathbb{1}_{\{p\in S_i\}}] \tag{4.5}$$

It is worth noting that there are no explicit criteria to assign the inference of S_i to one of the four cases. There is a trade-off between including more information and dealing with sample sparsity. For example, one may argue that both historical and subsequent trip information should be included for all inferences. However, many passengers do not have reliable history trips or future trips (i.e. $p \notin \mathcal{P}^H \cap \mathcal{P}^F$). The inference for those passengers can only be approximated by the results of $p \in \mathcal{P}^H \cap \mathcal{P}^F$ (details in Section 4.3). Hence, simply including more information will lead to higher approximation errors due to sample sparsity, which is the reason that we have four types of formulations and some of them only include either future or history information, but not both. Determining the formulation for an S_i needs empirical knowledge and numeral tests to judge which kinds of information are more critical for the inference.

Uncertainty

In this study, we estimate the variance of the N_{S_i} ($\text{Var}[N_{S_i}]$) to quantify the uncertainty. $\text{Var}[N_{S_i}]$ captures the behavioral randomness of passengers in B_{S_i}. The behavior of a passenger in B_{S_i} is atypical or not (i.e., $\mathbb{1}_{\{\text{“Behavior atypical”} \mid p \in B_{S_i}\}}$) is an indicator random variable. High behavioral randomness indicates high variance of N_{S_i} because we cannot easily conclude whether a passenger's observed behavior in the incident day is typical or not. In this case, $\mathbb{P}(\text{“Behavior atypical”} \mid p \in B_{S_i})$ is close to 0.5 (where $\text{Var}[N_{S_i}]$ reaches the maximum), which implies that the passenger's behavior pattern is hard to estimate from the historical trips.

Besides passengers' inherent travel irregularity, $\text{Var}[N_{S_i}]$ is also determined by the definition of B_{S_i}. If B_{S_i} is specified narrowly, such as a set of passengers with a transfer trip to bus stops near the blocked rail stations after the incident, passengers may seldom have this “complicated” behavior on normal days. If a passenger has this behavior in the incident day, it is highly likely to be atypical (i.e., $\mathbb{P}(\text{“Behavior atypical”} \mid p \in B_{S_i})$ is close to 1). In this case, the $\text{Var}[N_{S_i}]$ is relatively low. However, if B_{S_i} has a very broad definition, such as a set of passengers with a bus trip in the incident period, $\mathbb{P}(\text{“Behavior atypical”} \mid p \in B_{S_i})$ may be close to 0.5 because passengers may use different modes on different normal days and it is difficult to infer having a bus trip is atypical or not on the incident day. In this case, the $\text{Var}[N_{S_i}]$ is relatively high. Since the definition of B_{S_i} is according to S_i, $\text{Var}[N_{S_i}]$ provides the information about whether S_i is easy to be inferred by the AFC data or not (low variance means S_i can be inferred more precisely).

4.3 Model formulation

In this section, we elaborate on the inference formulation for every behavior group. The section is organized by the formulation cases mentioned in Section 4.2.2. However, due to the tedious derivations and some formulation duplication, we only present the formulations for a part of behavior groups. The complete formulations can be found in Section 4.6.

4.3.1 Historical trip information + direct incident-related B_{S_i}: Inferring S_1 and S_2

By definition, passengers in S_1 and S_2 have at least one rail tap-in record before T_1 because they were in the blocked stations/trains when the incident happened. Since passengers who decide to use the public transit system again after alighting need to re-tap in, passengers in S_1 have another bus tap-in record after T_1, and passengers in S_2 have another rail tap-in record after T_1.

As passengers in S_1 and S_2 left the rail system from the blocked stations, the re-tap-in bus/rail stations should be close to the blocked stations and the time difference between two consecutive tap-ins should not be too large. Otherwise, they may be two separate trips instead of a transfer. Let TT_d be the tap-in time difference threshold for transferring. We assume that if $t_{p_k} - t_{p_{k-1}} < TT_d$, trip k is a transfer trip following trip $k-1$[2]. Denote the walking distance threshold for passengers transferring to a bus (resp. rail) as d_b (resp. d_r). Then the set of bus (resp. rail) stops close to the blocked stations is defined as $\mathcal{W}_b = \{s : s \text{ is a bus station and } \exists s' \in \mathcal{W} \text{ s.t. } D(s,s') \leq d_b\}$ (resp. $\mathcal{W}_r = \{s : s \notin \mathcal{W} \text{ is a rail station and } \exists s' \in \mathcal{W} \text{ s.t. } D(s,s') \leq d_r\}$), where $D(s,s')$ returns the walking distance between stations s and s'.

To identify passengers in S_1, we define a passenger set $B_{S_1} = \{p : \exists k \in \{1, ..., K_p -$

[2] This is a typical way for tap-in only public transit systems to determine transfer trips for fare calculation. Future study may include tap-out time estimation model to better define a transfer trip

$1\}$ s.t. $t_{p_k} \leq T_1 < t_{p_{k+1}},\ t_{p_{k+1}} - t_{p_k} < TT_d,\ m_{p_k} = \text{rail},\ m_{p_{k+1}} = \text{bus},\ o_{p_{k+1}} \in \mathcal{W}_b\}$. B_{S_1} represents passengers with a rail tap-in record before the incident and a bus transferring tap-in record after the incident. And the second tap-in station is within the walking distance of the blocked stations. As we described above, passengers in S_1 should also in B_{S_1} ($S_1 \subseteq B_{S_1}$). However, S_1 and B_{S_1} are not necessarily equivalent because passengers in B_{S_1} may transfer to a bus stop as a normal routine, that is, they did not transfer to a bus line in response to the rail disruption. Denote the event that p was affected by the incident as A_p. Then we have

$$\mathbb{E}[N_{S_1}] = \sum_{p\in\mathcal{P}} \mathbb{E}[\mathbb{1}_{\{p\in S_1\}}] = \sum_{p\in\mathcal{P}} \mathbb{E}[\mathbb{1}_{\{p\in B_{S_1}\}} \cdot \mathbb{1}_{\{A_p \mid p\in B_{S_1}\}}] = \sum_{p\in\mathcal{P}} \mathbb{1}_{\{p\in B_{S_1}\}} \cdot \mathbb{P}(A_p \mid p \in B_{S_1}) \tag{4.6}$$

Note that $\mathbb{1}_{\{p\in B_{S_1}\}}$ is a constant because for every p, we observe whether it belongs to B_{S_1} or not using the AFC data from the incident day. $\mathbb{P}(A_p \mid p \in B_{S_1})$ is calculated as

$$\mathbb{P}(A_p \mid p \in B_{S_1}) = 1 - \underbrace{\frac{\#\text{ normal days } p \text{ showing trip records described in } B_{S_1}}{M_p}}_{\text{Prob. that transferring to a bus stop near blocked station is a typical behavior}} \qquad \forall p \in \mathcal{P}^H. \tag{4.7}$$

Eq. 4.7 means that given a passenger with the observed behavior described in B_{S_1} on the incident day, the probability that this behavior is atypical[3] equals to 1 minus the relative frequency that the passenger has the same behavior on normal days. For example, if p transferred to a bus stop in $\mathcal{W}_b$ on every normal day, then transferring to the bus stop in $\mathcal{W}_b$ is highly likely to be a routine, rather than a change in behavior due to the incident (i.e., $\mathbb{P}(A_p \mid p \in B_{S_1}) = \frac{0}{M_p} = 0$). Then, p will not be counted into S_1.

If history information of p is unavailable or very limited (i.e., $p \notin \mathcal{P}^H$), Eq. 4.7

[3] Formulation type 1, "atypical" = "affected by the incident" in this case

may fail to work. In this scenario, we assume

$$\mathbb{P}(A_p \mid p \in B_{S_1}) = \frac{\sum_{p' \in \mathcal{P}^H} \mathbb{P}(A_{p'} \mid p' \in B_{S_1})}{|\mathcal{P}^H \cap B_{S_1}|} \quad \forall p \notin \mathcal{P}^H \tag{4.8}$$

which estimates the corresponding probability of passengers with little historical information using that of passengers with enough historical information. This is a typical way to estimate behavior of passengers without enough information in the AFC data [131], though it may be biased considering different behavior patterns for $p \in \mathcal{P}^H$ and $p \notin \mathcal{P}^H$. There is no better way to address this issue given data limitations.

As $\mathbb{1}_{\{A_p \mid p \in B_{S_1}\}}$ is a Bernoulli random variable with probability $\mathbb{P}(A_p \mid p \in B_{S_1})$, the corresponding variance of N_{S_1} can be calculated as

$$\begin{aligned} \text{Var}[N_{S_1}] &= \sum_{p \in \mathcal{P}} (\mathbb{1}_{\{p \in B_{S_1}\}})^2 \cdot \text{Var}[\mathbb{1}_{\{A_p \mid p \in B_{S_1}\}}] \\ &= \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_1}\}} \cdot [\mathbb{P}(A_p \mid p \in B_{S_1}) - \mathbb{P}(A_p \mid p \in B_{S_1})^2] \end{aligned} \tag{4.9}$$

Similarly, for passengers in S_2, similarly, we can define $B_{S_2} = \{p : \exists k \in \{1, ..., K_p - 1\} \text{ s.t. } t_{p_k} \leq T_1 < t_{p_{k+1}},\ t_{p_{k+1}} - t_{p_k} < TT_d,\ m_{p_k} = \text{rail},\ m_{p_{k+1}} = \text{rail},\ o_{p_{k+1}} \in \mathcal{W}_r\}$. Then we have

$$\mathbb{E}[N_{S_2}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_2}\}} \cdot \mathbb{P}(A_p \mid p \in B_{S_2}) \tag{4.10}$$

where $\mathbb{P}(A_p \mid p \in B_{S_2})$ can be calculated in the same way as Eq. 4.7 and 4.8 by replacing B_{S_1} with B_{S_2}. And the variance of N_{S_2} can be calculated as

$$\text{Var}[N_{S_2}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_2}\}} \cdot [\mathbb{P}(A_p \mid p \in B_{S_2}) - \mathbb{P}(A_p \mid p \in B_{S_2})^2] \tag{4.11}$$

4.3.2 Historical trip information + direct incident-related B_{S_i} + subsequent trip information: Inferring S_8 and S_9

Passengers in groups S_8 and S_9 continued to use the public transit system after the incident. Hence, they have at least one tap-in record before T_1 and at least one tap-in record after T_1. The difference between S_8, S_9 and S_1, S_2 is that passengers in S_8 and S_9 leave the rail system at some upstream station before the blocked stations. To differentiate S_8 and S_9 with other normal transfer passengers, we need to infer their original route and consider whether the route is blocked. If their original routes are not blocked, the transfers are not due to the incident.

We first identify S_8. Consider a passenger $p \in \mathcal{P}^F$. Suppose $\exists k \in \{1, ..., K_p - 1\}$ s.t. $t_{p_k} \leq T_1 < t_{p_{k+1}}, t_{p_{k+1}} - t_{p_k} < TT_d, m_{p_k} =$ rail, $m_{p_{k+1}} =$ bus, and $o_{p_{k+1}} \notin \mathcal{W}_b$, which means p has a rail trip before the incident and a bus transfer trip after the incident, and the boarding stop of bus trip is not close to the blocked stations (otherwise he/she is already considered in the inference of S_1). If p was affected by the incident, the transferring trip $k + 1$ would be an atypical behavior for p.

Denote the event "transferring is atypical for p" as TA_p. Let $(o_{p_{k^*}}, t_{p_{k^*}}, m_{p_{k^*}})$ be the next non-transfer trip of trip $k + 1$. Mathematically, $k^* = \min\{k' > k + 1 : t_{p_{k'}} - t_{p_{k+1}} > TT_d\}$. Given TA_p, if without any incident, the original trip chain for passenger p is $\{..., (o_{p_k}, t_{p_k}, m_{p_k}), (o_{p_{k^*}}, t_{p_{k^*}}, m_{p_{k^*}}), ...\}$, the observed transfer bus trip $(o_{p_{k+1}}, t_{p_{k+1}}, m_{p_{k+1}})$ is caused by the disruption. Our goal is to use trip k^* to infer the original destination of trip k (i.e. the destination under normal condition). This can be done from the destination estimation model using the trip chain method [129, 130, 131]. Let the set of all possible *original* destinations for trip k be $\mathcal{D}_{p_k}$. and $\tilde{d}_{p_k}$ the random variable representing the original destination of trip k. The destination estimation model provides $\mathbb{P}(\tilde{d}_{p_k} = d)$ for any $d \in \mathcal{D}_{p_k}$.

However, trip k^* may not exist for some p because the subsequent trip information may not be available (e.g., $p \notin \mathcal{P}^F$). For $p \notin \mathcal{P}^F$, the destination distribution can be approximated by $p \in \mathcal{P}^F$ [131]:

$$\mathbb{P}(\tilde{d}_{p_k} = d) = \frac{\sum_{p' \in \mathcal{P}^F : o_{p_k} = o_{p'_k}} \mathbb{P}(\tilde{d}_{p'_k} = d)}{|\{p' \in \mathcal{P}^F : o_{p_k} = o_{p'_k}\}|} \quad \forall p \notin \mathcal{P}^F, d \in \mathcal{D}_{p_k}. \tag{4.12}$$

Eq 4.12 means that the probability of $\tilde{d}_{p_k} = d$ for $p \notin \mathcal{P}^F$ is estimated as the average value of $p \in \mathcal{P}^F$ with the same origin.

As we assume that, for a given $\tilde{d}_{p_k}$, passengers follow the shortest path [149], the original route for p from o_{p_k} to $\tilde{d}_{p_k}$ can be obtained. Using automated vehicle location (AVL) data and a transit loading model [150, 4], we can further infer the location of passenger p in the rail system at time T_1 for a given $\tilde{d}_{p_k}$. Suppose that at time T_1, p was in location $s_p(T_1, \tilde{d}_{p_k})$ (which corresponds to a station or some middle point between two stations). Then, if the remaining route segment from $s_p(T_1, \tilde{d}_{p_k})$ to $\tilde{d}_{p_k}$ was blocked, p would be affected by the incident. Let the event that the original route of p is blocked given the original destination is d be $RB_p(d)$.

We define $B_{S_8} = \{p : \exists k \in \{1, ..., K_p - 1\} \text{ s.t. } t_{p_k} \leq T_1 < t_{p_{k+1}},\ t_{p_{k+1}} - t_{p_k} < TT_d,\ m_{p_k} = \text{rail},\ m_{p_{k+1}} = \text{bus}, o_{p_{k+1}} \notin \mathcal{W}_b\}$, which represents passengers with a rail tap-in record before the incident and a bus transferring tap-in record after the incident. Then we have $\mathbb{1}_{\{p \in S_8\}} = \mathbb{1}_{\{p \in B_{S_8}\}} \cdot \mathbb{1}_{\{TA_p\}} \cdot \sum_{d \in \mathcal{D}_{p_k}} \mathbb{1}_{\{RB_p(d)\}} \cdot \mathbb{1}_{\{\tilde{d}_{p_k} = d\}}$. Note that $\mathbb{1}_{\{TA_p\}}$ and $\mathbb{1}_{\{RB_p(d)\}}$ are independent because the former is determined by the historical trips while the later is determined by the subsequent trips after the incident. Therefore, the number of passengers in S_8 can be calculated as:

$$\mathbb{E}[N_{S_8}] = \sum_{p \in \mathcal{P}} \mathbb{E}[\mathbb{1}_{\{p \in S_8\}}] = \sum_{p \in \mathcal{P}} \sum_{d \in \mathcal{D}_{p_k}} \mathbb{1}_{\{p \in B_{S_8}\}} \cdot \mathbb{P}(TA_p \mid p \in B_{S_8}) \cdot \mathbb{1}_{\{RB_p(d)\}} \cdot \mathbb{P}(\tilde{d}_{p_k} = d) \tag{4.13}$$

$\mathbb{1}_{\{RB_p(d)\}}$ is a constant because given the original destination and path, we can conclude whether the path is blocked or not. $\mathbb{P}(TA_p \mid p \in B_{S_8})$ can be calculated in the same way as Eq. 4.7 and 4.8 by replacing B_{S_1} and A_p with B_{S_8} and TA_p, respectively.

The variance of N_{S_8} can be calculated as

$$\begin{aligned}\mathrm{Var}[N_{S_8}] = &\sum_{p\in\mathcal{P}}\sum_{d\in\mathcal{D}_{p_k}} \mathbb{1}_{\{p\in B_{S_8}\}} \cdot \mathbb{1}_{\{RB_p(d)\}} \\ &\cdot [\mathbb{P}(TA_p \mid p \in B_{S_8}) \cdot \mathbb{P}(\tilde{d}_{p_k} = d) - \quad \mathbb{P}(TA_p \mid p \in B_{S_8})^2 \cdot \mathbb{P}(\tilde{d}_{p_k} = d)^2]\end{aligned} \tag{4.14}$$

Similarly, for passengers in S_9, we have $B_{S_9} = \{p : \exists k \in \{1, ..., K_p - 1\}$ s.t. $t_{p_k} \leq T_1 < t_{p_{k+1}},\ t_{p_{k+1}} - t_{p_k} < TT_d,\ m_{p_k} = \text{rail},\ m_{p_{k+1}} = \text{rail}, o_{p_{k+1}} \notin \mathcal{W}_r\}$. Then:

$$\mathbb{E}[N_{S_9}] = \sum_{p\in\mathcal{P}}\sum_{d\in\mathcal{D}_{p_k}} \mathbb{1}_{\{p\in B_{S_9}\}} \cdot \mathbb{P}(TA_p \mid p \in B_{S_9}) \cdot \mathbb{1}_{\{RB_p(d)\}} \cdot \mathbb{P}(\tilde{d}_{p_k} = d) \tag{4.15}$$

$$\begin{aligned}\mathrm{Var}[N_{S_9}] = &\sum_{p\in\mathcal{P}}\sum_{d\in\mathcal{D}_{p_k}} \mathbb{1}_{\{p\in B_{S_8}\}} \cdot \mathbb{1}_{\{RB_p(d)\}} \\ &\cdot [\mathbb{P}(TA_p \mid p \in B_{S_9}) \cdot \mathbb{P}(\tilde{d}_{p_k} = d) - \mathbb{P}(TA_p \mid p \in B_{S_9})^2 \cdot \mathbb{P}(\tilde{d}_{p_k} = d)^2]\end{aligned} \tag{4.16}$$

4.3.3 Historical trip information + indirect incident-related B_{S_i}: Inferring S_5 and S_{11}

S_5 and S_{11} are people who were already in the rail system and decided to cancel their trips because of the rail disruption. The AFC records of passengers in S_5 and S_{11} can be described as $B_{S_{5,11}} = \{p : t_{p_{K_p}} \leq T_1,\ m_{p_{K_p}} = \text{rail}\}$, which means passengers having at least one rail tap-in record before T_1 and no tap-in record between T_1 and T_e.

Consider a passenger $p \in \mathcal{P}^F \cap \mathcal{P}^H$. Let $(o_{p_{k^*}}, t_{p_{k^*}}, m_{p_{k^*}})$ be the next non-transfer trip following trip K (i.e., $k^* = \min\{k' > K : t_{p_{k'}} - t_{p_{K_p}} > TT_d\}$). As k^* is the next non-transfer trip right after K, p had no non-transfer trips within $[t_{p_{K_p}}, t_{p_{k^*}}]$ on the incident day. We use an example to illustrate the AFC records that may help to identify S_5 and S_{11}. Consider a passenger who plans to go to the supermarket on the incident day. He/she was in the system when the incident happened. Suppose that he/she decided to cancel his/her trip and return home. Then he/she would not have the typical returning trip from the supermarket. In this situation, k^* may be some other trips late in the evening or the first trip in the next day. However, in the historical AFC records. the typical trip right after K_p should be the returning trip from the supermarket. Therefore, we can assume that if passenger p has high probability of having trips within $[t_{p_{K_p}}, t_{p_{k^*}}]$ on normal days, he/she is very likely to cancel the trip K_p because the typical following trip for K_p that is supposed to occur in $[t_{p_{K_p}}, t_{p_{k^*}}]$ does not exist on the incident day.

However, it is worth noting that since we only have public transit trip records, passengers who do not cancel trips but use other travel modes to replace both trip K_p and the returning trip may also be identified as “cancel trips”. Consider the example above, if a passenger takes Uber to the supermarket and then takes Uber back. He/she would be identified as “cancel trips”. However, the information in AFC data is not enough to differentiate these two groups of passengers. Hence, in this study, we assume that the incident only changes passengers’ mode choices of trips in the analysis period, which implies that the returning trip travel mode for the passenger will be public transit if he/she usually uses public transit. Note that this assumption can be relaxed if we focus on estimating the number of passengers “not using public transit” in an aggregated framework (see Figure 4-2).

Denote the event that passenger $p \in B_{S_{5,11}}$ canceled trip K_p after the incident as CT_p. Based on the assumption above, we can derive the probability as

$$\begin{aligned}&\mathbb{P}(CT_p \mid p \in B_{S_{5,11}}) \\ &= 1 - \frac{\#\text{ normal days } p \text{ has rail trips in } [T_s, T_1] \text{ with origin } o_{p_{K_p}} \text{ but no trip in } [t_{p_{K_p}}, t_{p_{k^*}}]}{\#\text{ normal days } p \text{ has rail trips in } [T_s, T_1] \text{ with origin } o_{p_{K_p}}} \qquad (4.17)\\ &\quad \forall p \in \mathcal{P}^H \cap \mathcal{P}^F\end{aligned}$$

The second term in Eq. 4.17 represents the conditional probability that there is no trip in $[t_{p_{K_p}}, t_{p_{k^*}}]$ on normal days given that the passenger already has a rail trip in $[T_s, T_1]$ with origin $o_{p_{K_p}}$. The lower is this probability, the higher is the probability that this behavior is atypical (i.e. the passenger actually cancels his/her trip) on the incident day.

For $p \notin \mathcal{P}^H \cap \mathcal{P}^F$, similar to Eq. 4.8, we can approximate the probability by that of passengers in $\mathcal{P}^H \cap \mathcal{P}^F$:

$$\mathbb{P}(CT_p \mid p \in B_{S_{5,11}}) = \frac{\sum_{p' \in \mathcal{P}^H \cap \mathcal{P}^F} \mathbb{P}(CT_{p'} \mid p \in B_{S_{5,11}})}{|\mathcal{P}^H \cap \mathcal{P}^F \cap B_{S_{5,11}}|} \quad \forall p \notin \mathcal{P}^H \cap \mathcal{P}^F \tag{4.18}$$

As mentioned before, passengers may cancel trips due to many reasons, not necessarily because of the incidents. Therefore, we need to consider the event $CT_p \cap A_p$, which represents passengers canceling trips because of the incident. However, directly calculating $\mathbb{P}(CT_p,\ A_p \mid p \in B_{S_{5,11}})$ is difficult. The following equations show an aggregate calculation approach:

$$\begin{aligned}
\mathbb{E}[N_{S_5} + N_{S_{11}}] &= \sum_{p \in \mathcal{P}} \mathbb{E}[\mathbb{1}_{\{p \in S_5 \cup S_{11}\}}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{5,11}}\}} \cdot \mathbb{P}(CT_p,\ A_p \mid p \in B_{S_{5,11}}) \\
&= \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{5,11}}\}} \cdot \mathbb{P}(CT_p \mid p \in B_{S_{5,11}}) \cdot \mathbb{P}(A_p \mid CT_p,\ p \in B_{S_{5,11}}) \\
&= \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{5,11}}\}} \cdot \mathbb{P}(CT_p \mid p \in B_{S_{5,11}})(1 - \mathbb{P}((A_p)^c \mid CT_p,\ p \in B_{S_{5,11}})) \\
&= N_{CT} - \tilde{N}_{CT}
\end{aligned} \tag{4.19}$$

where

$$N_{CT} := \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{5,11}}\}} \cdot \mathbb{P}(CT_p \mid p \in B_{S_{5,11}}) \tag{4.20}$$

$$\tilde{N}_{CT} := \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{5,11}}\}} \cdot \mathbb{P}(CT_p \mid p \in B_{S_{5,11}}) \cdot \mathbb{P}((A_p)^c \mid CT_p,\ p \in B_{S_{5,11}}) \tag{4.21}$$

N_{CT} is the expected number of passengers who canceled trips on the incident day (not necessarily due to the incident) and $\tilde{N}_{CT}$ is the expected number of passengers who canceled trips on the incident day and the reason is not the incident. We can approximate $\tilde{N}_{CT}$ as the number of passengers canceling trips on normal days. Specifically, denote $\tilde{N}_{CT}^{(j)}$ as the number of canceling-trip passengers calculated by applying Eq.

4.20 to the AFC data of j-th normal day. Then we have

$$\mathbb{E}[N_{S_5} + N_{S_{11}}] = N_{CT} - \tilde{N}_{CT} = N_{CT} - \frac{\sum_{j=1}^{M} \tilde{N}_{CT}^{(j)}}{M} \tag{4.22}$$

To calculate the variance of $N_{S_5} + N_{S_{11}}$, we assume

$$\mathbb{P}(A_p \mid CT_p, p \in B_{S_{5,11}}) = \frac{N_{CT} - \tilde{N}_{CT}}{N_{CT}}. \quad \forall p \in B_{S_{5,11}} \tag{4.23}$$

Eq. 4.23 means the probability that p's behavior is atypical given that he/she canceled trips equals the expected number of passengers canceling trips due to the incident divided by the total expected number of passengers canceling trips (not necessary due to the incident). It implies that we are using population statistics to approximate the individual probability. Then, we can calculate the variance as:

$$\begin{aligned}\text{Var}[N_{S_5} + N_{S_{11}}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{5,11}}\}} \cdot [&\mathbb{P}(CT_p \mid p \in B_{S_{5,11}})\mathbb{P}(A_p \mid CT_p, p \in B_{S_{5,11}}) - \\ &\mathbb{P}(CT_p \mid p \in B_{S_{5,11}})^2 \mathbb{P}(A_p \mid CT_p, p \in B_{S_{5,11}})^2]\end{aligned} \tag{4.24}$$

It is worth noting that the number of passengers canceling trips due to the incident is expected to be small. Therefore, Eq. 4.22 may be smaller than zero due to variations in the AFC data. This means that there is no big difference between the number of canceling trips passengers (Eq. 4.20) on the incident day and on normal days, implying the number of passengers who canceled trips due to the incident is negligible. In this situation, we simply let $\mathbb{E}[N_{S_5} + N_{S_{11}}] = 0$ and $\text{Var}[N_{S_5} + N_{S_{11}}] = 0$.

4.3.4 Subsequent trip information only: Inferring S_3, S_7 and S_{10}

Identifying S_3, S_7 and S_{10} is similar to identifying S_8 and S_9. The inference leverages the subsequent trip information to infer the original routes. We consider these three groups together because they have the same AFC records in the incident day (i.e., at least one rail tap-in record before T_1 and no tap-in record between T_1 and T_e.). We define the corresponding set as $B_{S_{3,7,10}} = \{p : t_{pK_p} \leq T_1,\ m_{pK_p} = \text{rail}\}$.

Passengers in S_7 are those who transfer at some upstream stations (not go out) if their original rail route is blocked. For p in $B_{S_{3,7,10}} \cap \mathcal{P}^F$, let k^* be his/her next non-transfer trip after trip K_p. Then, using the same way as in Section 4.3.2, we can infer the destination distribution for trip K_p (i.e. obtain $\mathbb{P}(\tilde{d}_{pK_p} = d)$ for any $d \in \mathcal{D}_{pK_p}$), as well as their locations when the incident happened (i.e. $s_p(T_1, \tilde{d}_{pK_p})$). For a given $\tilde{d}_{pK_p}$, if the original route from $s_p(T_1, \tilde{d}_{pK_p})$ to $\tilde{d}_{pK_p}$ is blocked, as we do not observe another tap-in record in $[T_1,\ T_2]$, p would only have three options: 1) transferring to alternative routes from $s_p(T_1, \tilde{d}_{pK_p})$ to $\tilde{d}_{pK_p}$ without going out of the rail system (S_7), 2) using other undetected modes ($S_3 + S_{10}$), and 3) canceling the trip ($S_5 + S_{11}$). This section focuses on the first two behaviors. It is worth noting that passengers can transfer only if there exist alternative routes from $s_p(T_1, \tilde{d}_{pK_p})$ to $\tilde{d}_{pK_p}$ within the rail system. Given an inferred original destination $d \in \mathcal{D}_{pK_p}$, we denote the event that p's original route is blocked but transfer is available as $RBTA_p(d)$.

We assume that passengers would not cancel trips when alternative routes were available. Then, if $RBTA_p(d)$ was true, p could either use intra-system transferring or use other undetected modes. However, given the data limitations, there is no available information to differentiate these two behaviors. We, thus, assume that the probability of using rail if a transfer is available, given the destination d of passenger p, is $\alpha_{p,d}$, that is, $\mathbb{P}(\mathbb{1}_{\{URTA_p|d\}} = 1) = \alpha_{p,d}$, where $URTA_p|d$ is the event that passenger p will use rail if a transfer is available given the destination is d. $\alpha_{p,d}$ can be estimated using a discrete choice model (DCM) with the utility expressed as a function of the travel cost, travel time of different travel modes (including transfer by rail, TNC, etc.) [132]. Given d, travel cost and travel time of different travel modes can be obtained from the Google Map API, and the parameters in the DCM can be estimated from survey data [106].

Notice that $\mathbb{1}_{\{URTA_p|d\}}$ is independent of $\mathbb{1}_{\{\tilde{d}_{pK}=d\}}$ (i.e., the conditional independence). And based on the above assumptions, we have

$$\mathbb{E}[N_{S_7}] = \sum_{p\in\mathcal{P}} \mathbb{E}[\mathbb{1}_{\{p\in S_7\}}] = \sum_{p\in\mathcal{P}} \sum_{d\in\mathcal{D}_{p_K}} \mathbb{1}_{\{p\in B_{S_{3,7,10}}\}} \cdot \mathbb{1}_{\{RBTA_p(d)\}} \cdot \alpha_{p,d} \cdot \mathbb{P}(\tilde{d}_{p_K} = d) \tag{4.25}$$

And the corresponding variance is

$$\mathrm{Var}[N_{S_7}] = \sum_{p\in\mathcal{P}} \sum_{d\in\mathcal{D}_{p_K}} \mathbb{1}_{\{p\in B_{S_{3,7,10}}\}} \cdot \mathbb{1}_{\{RBTA_p(d)\}} \cdot [\alpha_{p,d} \cdot \mathbb{P}(\tilde{d}_{p_K} = d) - \alpha_{p,d}^2 \cdot \mathbb{P}(\tilde{d}_{p_K} = d)^2] \tag{4.26}$$

Passengers in $B_{S_{3,7,10}}$ whose original routes were blocked and a transfer is *not* available have two options: 1) using other undetected modes or 2) canceling trips. Hence, we can use the total number of transfer-unavailable passengers minus the number of canceling-trip passengers to represent passengers using other undetected modes ($S_3 + S_{10}$). Note that when a transfer is available, passengers with probability $1-\alpha_{p,d}$ may choose other undetected modes, and should also be counted into S_3+S_{10}. Given an inferred original destination $d \in \mathcal{D}_{p_{K_p}}$, denote the event that p's original route is blocked and transfer is not available as $RBTN_p(d)$. Then,

$$\begin{aligned}\mathbb{E}[N_{S_3} + N_{S_{10}}] = &\sum_{p\in\mathcal{P}} \sum_{d\in\mathcal{D}_{p_{K_p}}} \mathbb{1}_{\{p\in B_{S_{3,7,10}}\}} \cdot \mathbb{1}_{\{RBTA_p(d)\}} \cdot (1-\alpha_{p,d}) \cdot \mathbb{P}(\tilde{d}_{p_{K_p}} = d) + \\ &\sum_{p\in\mathcal{P}} \sum_{d\in\mathcal{D}_{p_{K_p}}} \mathbb{1}_{\{p\in B_{S_{3,7,10}}\}} \cdot \mathbb{1}_{\{RBTN_p(d)\}} \cdot \mathbb{P}(\tilde{d}_{p_{K_p}} = d) - \\ &\mathbb{E}[N_{S_5} + N_{S_{11}}]\end{aligned} \tag{4.27}$$

The first term in Eq. 4.27 indicates passengers with available intra-system transfer routes but still choosing other undetected modes. The second term represents the total number of passengers without intra-system transfer routes. And the third term ($\mathbb{E}[N_{S_5} + N_{S_{11}}]$) is the number of passengers canceling trips, which is calculated in Section 4.3.3.

According to Section 4.3.3, $N_{S_5} + N_{S_{11}} = \sum_{p\in\mathcal{P}} \mathbb{1}_{\{p\in B_{S_{5,11}}\}} \cdot \mathbb{1}_{\{CT_p, A_p \mid p\in B_{S_{5,11}}\}}$. And

$\mathbb{1}_{\{URTA_p|d\}} \cdot \mathbb{1}_{\{\tilde{d}_{pK_p}=d\}}$ is independent of $\mathbb{1}_{\{CT_p, A_p \mid p \in B_{S_{5,11}}\}}$ because the choice behavior ($\mathbb{1}_{\{URTA_p|d\}}$) is estimated from survey data, the destination inference ($\mathbb{1}_{\{\tilde{d}_{pK_p}=d\}}$) is based on subsequent trip information, while the estimation of canceling trips ($\mathbb{1}_{\{CT_p, A_p \mid p \in B_{S_{5,11}}\}}$) is based on historical trip information. So, the variance can be calculated as

$$\begin{aligned}&\mathrm{Var}[N_{S_3} + N_{S_{10}}]\\ &= \sum_{p\in\mathcal{P}} \sum_{d\in\mathcal{D}_{p_K}} \mathbb{1}_{\{p\in B_{S_{3,7,10}}\}} \cdot \mathbb{1}_{\{RBTA_p(d)\}} \cdot [(1-\alpha_{p,d}) \cdot \mathbb{P}(\tilde{d}_{p_K} = d) - (1-\alpha_{p,d})^2 \cdot \mathbb{P}(\tilde{d}_{p_K} = d)^2] + \\ &\quad \sum_{p\in\mathcal{P}} \sum_{d\in\mathcal{D}_{p_K}} \mathbb{1}_{\{p\in B_{S_{3,7,10}}\}} \cdot \mathbb{1}_{\{RBTN_p(d)\}} \cdot [\mathbb{P}(\tilde{d}_{p_K} = d) - \mathbb{P}(\tilde{d}_{p_K} = d)^2] \; + \mathrm{Var}[N_{S_5} + N_{S_{11}}]\end{aligned}$$

where $\mathrm{Var}[N_{S_5} + N_{S_{11}}]$ is obtained in Section 4.3.3.

4.4 Case study

4.4.1 Chicago Transit System

We use data from the CTA transit system as the case study because this research focuses on open public transit systems with only tap-in information and CTA is an open system.

CTA is the second-largest transit system in the United States, providing services in Chicago, Illinois, and some of its surrounding suburbs. The transit network consists of the Chicago "L" (rail) and CTA bus services. It operates 24 hours each day and on an average weekday provides 0.84 and 0.81 million rides on buses and trains, respectively [40]. The map of the CTA rail system is shown in Figure 4-4. The rail system consists of eight lines (named by color) and the "Loop". The Loop, located in the Chicago downtown area, is the 2.88 km long circuit of elevated rail that forms the hub of the Chicago rail system. Its eight stations account for around 10% of all weekday boardings on the CTA trains.

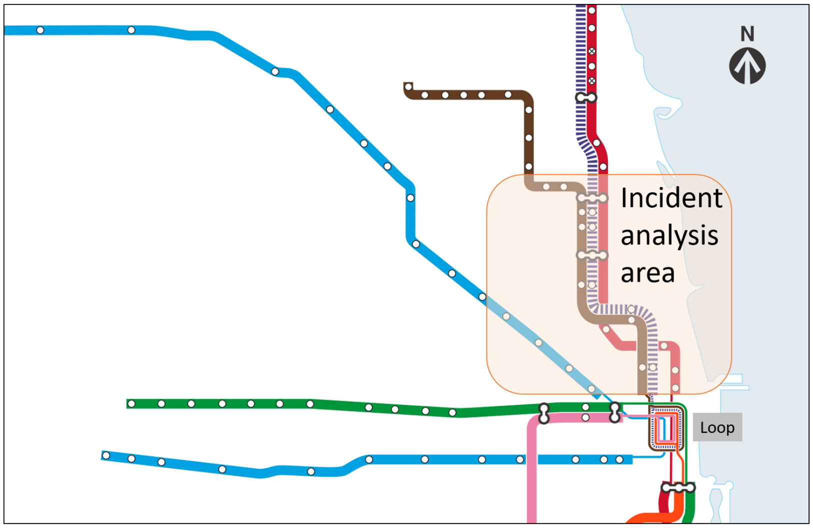


Figure 4-4: CTA rail system map

Two data sources are used in this study: the AFC transaction data and train tracker (or AVL) data. CTA's AFC system is entry-only as passengers only use their fare cards when entering a rail station or boarding a bus. No information about a trip's destination is directly provided. The AVL system provides trains' arrival/departure times at each station.

4.4.2 Disruption background

The rail disruption used in this study happened on September 24, 2019. At 9:09AM, two trains collided at the Sedgwick station on the Brown line (see Figures 4-4 and 4-5). This collision caused an interruption in service with five stations near Sedgwick on both Purple and Brown lines which are paralleled in this area being blocked (Figure 4-5). The disruption lasted for 70 minutes and ended at 10:19 AM when trains returned to normal operations.

The reasons for choosing this incident are as follows: 1) It is a substantial unplanned service disruption that can trigger observable behavior changes. 2) The incident area has enough alternative services (such as nearby rail lines, bus routes, etc.) to cover 19 possible behaviors so that we can illustrate the proposed model's performance.

When the incident happened, passengers who were in blocked stations and trains were cleared out of the system. The station closure sign was placed outside the fare collection gate in blocked stations, reminding passengers about the service suspension. CTA informed passengers about the incident from both the Ventra app (CTA user app to manage and pay fares on CTA) and CTA Tweets right after the disruption. All passengers in the system were informed of train and platform announcements.

During the service interruption, CTA provided bus shuttle services between Fullerton and Merchandise Mart. People who were forced to leave their trains from the blocked stations would re-tap-in if they decided to use CTA normal bus or rail services and were only charged a small transfer fee[4]. However, no tap-in is needed for shuttle bus users. Hence, the shuttle bus is defined as an undetected mode in this study.

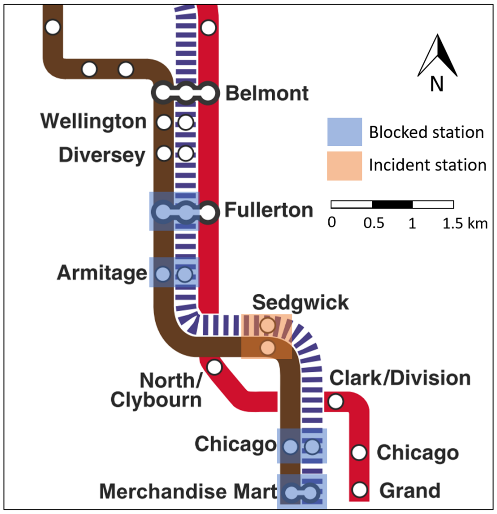


Figure 4-5: Rail disruption case

4.4.3 Parameter settings

Based on the incident information, the incident start time is $T_1 = 9:09$ AM and end time $T_2 = 10:19$ AM. $\delta_1 = \delta_2 = 60$ min is used according to the network scale and the analysis of system recovery time [145]. Therefore, the analysis period is from

[4] Sometimes there is no need to re-tap-in, depending on whether the control center has informed the CTA staff working in rail stations and bus drivers to allow free rides, and whether passengers asked for free rides due to the incident. In this study, we assume all passengers would re-tap-in according to the observation in the AFC data

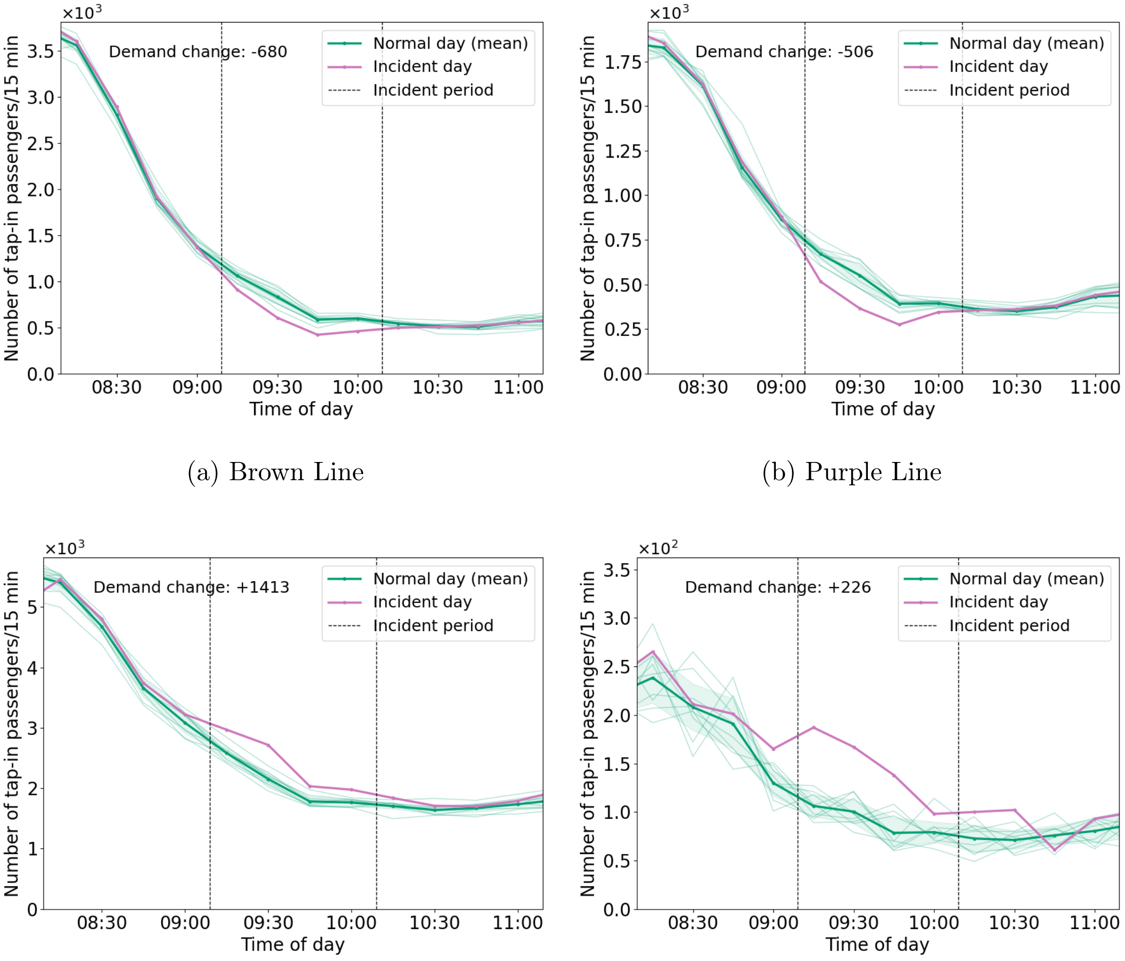
$T_s = 8:09$ AM to $T_e = 11:19$ AM. The normal days are selected as all Fridays (except for the incident day) in September and October, 2019.

The time threshold for transferring $TT_d = 2$ hours is used based on the CTA fare system. The walking distance threshold for bus and rail systems are set as $d_b = 0.7$ km and $d_r = 1.2$ km, respectively. These two numbers are slightly higher than the typical public transit transfer distance [151] so as to capture the increase in wiliness-to-walk during service disruptions.

As discussed before, $\alpha_{p,d}$ and β_p can be calculated based on the passenger's travel time and travel cost for different choices (including canceling trips) using DCM. The parameters in the DCM can be estimated from survey data or extracted from previous survey-based studies [85, 106]. The reason for using $\alpha_{p,d}$ and β_p is that, from AFC data alone, some groups of passengers cannot be identified as they have the same AFC transactions. AFC data only allows estimating $N_{S_3} + N_{S_7} + N_{S_{10}}$ (the number of passengers using intra-system transfers or not using public transit) and $N_{S_{17}} + N_{S_{18}}$ (the number of passengers out of the system when the incident happens and not using public transit) as a whole. Model-based inferences are necessary for differentiating these groups. In this study, as we focus on a data-driven approach, the model-based parameters are set as $\alpha_{p,d} = 0.95$, $\beta_p = 0.9$ for all p and d for simplicity. These values are based on the sample statistics of CTA riders who participated in the survey about travel mode choices during incidents [106].

4.4.4 Descriptive analysis

For a better understanding of the incident, we show the demand patterns of three rail lines (Brown, Purple, and Red) and bus stations around the incident area (i.e., $\mathcal{W}_b$). The line-level demand is calculated as the sum of all station demands in the line.



(a) Brown Line

(b) Purple Line

(c) Red Line

(d) Nearby bus stations

Figure 4-6: Demand comparison for normal days and the incident day. A green thin line represents the demand curve for a single normal day. The green shade areas represent the ±standard deviation. The demand change is calculated as the total number of tap-ins during the incident period (9:09 - 10:19 AM) on the incident day minus that of the normal day average.

Figure 4-6 shows the comparison of the number of tap-in passengers on the incident day and normal days (aggregated by 15-minutes interval). We observe that the normal day demand patterns are relatively consistent compared to the incident day, which enables us to differentiate behavioral discrepancy on the incident day. As expected, the demand on the Brown and Purple Lines (interrupted by the incident) both decreased during the incident (Figures 4-6a and 4-6b). And it gradually returned to normal with the end of the incident. As the Red Line runs adjacent to the Brown and Purple Lines for a large portion (see Figure 4-5) in the incident area and it is not suspended, we see a significant increase in demand during the incident period with a return to normal after the incident is over (Figures 4-6c). In terms of the nearby bus stops, the demand pattern is similar to that of the Red Line.

In terms of the demand change numbers, we see that the demand increase on the Red Line (1,413) is much higher than that in the nearby bus stations, implying that most of the passengers choose the Red Line as the alternative. Note that the total demand decrease in Brown and Purple lines is slightly smaller than the total demand increase due to the fact that some passengers may first tap into the incident lines and then leave. This means that the actual demand decrease is higher than $680 + 506 = 1,186$.

4.4.5 Rule-based benchmark models

We choose the rule-based deterministic method that has been used in previous studies [89, 117] as the benchmark model. The rule-based method directly maps passengers with observed behavior (B_{S_i}) to those who are influenced by the incident (S_i). Note that as this study considers different behavior sets from those of previous studies, we cannot use their rules to classify passengers. For a fair comparison, we use the rule defined in our study (B_{S_i}) as the criterion. Recall that there are four formulations in Section 4.2.2 to infer passengers' responses. For the rule-based model, the number of formulations reduces to two because some cases share the same formulation:

- **"Historical trip information + direct incident-related B_{S_i}"** and **"Historical trip information + indirect incident-related B_{S_i}"**: In the rule-based method, we assume $p \in B_{S_i}$ is equivalent to $p \in S_i$. Therefore, eliminating the probability component in Eqs. 4.1 and 4.2, we have

$$\mathbb{1}_{\{p\in S_i\}} = \mathbb{1}_{\{p\in B_{S_i}\}} \tag{4.28}$$

S_1, S_2, S_4, S_{12}, $S_5, S_{11}, S_{14}, S_{15}, S_{17}, S_{18}$, and S_{19} belong to this case.

- **"Subsequent trip information only"** and **"Historical trip information + direct incident-related B_{S_i} + Subsequent trip information"**: In this case, we first infer a destination d for the passenger. Then, eliminating the probability component in Eqs. 4.3 and 4.4, we have

$$\mathbb{1}_{\{p \in S_i\}} = \mathbb{1}_{\{p \in B_{S_i}\}} \cdot \mathbb{1}_{\{Y_p(d)\}} \tag{4.29}$$

The estimated number of passengers in group S_i is calculated as.

$$\hat{N}_{S_i} = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in S_i\}} \tag{4.30}$$

Since this is a deterministic method, variance information is not available.

4.4.6 Model validation with synthetic data

Synthetic data generation

Since there are no available observations for passengers' actual choices, the model validation is conducted with a simulation-based synthetic data set generated from the actual AFC data. The generation process is as follows. And the illustration diagram is shown in Figure 4-7.

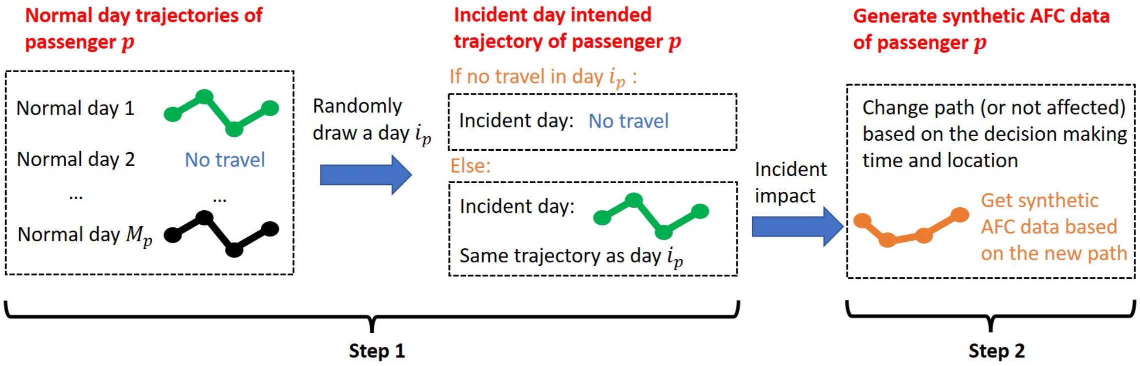

Normal day trajectories of passenger p
Normal day 1
Normal day 2
No travel
...
...
Normal day M_p
Randomly draw a day i_p
Incident day intended trajectory of passenger p
If no travel in day i_p :
Incident day: No travel
Else:
Incident day:
Same trajectory as day i_p
Incident impact
Generate synthetic AFC data of passenger p
Change path (or not affected) based on the decision making time and location
Get synthetic AFC data based on the new path
Step 1
Step 2

Figure 4-7: Diagram for synthetic data generation

Step 1: Sample intended trajectories. For each passenger p who has used the CTA system in any of M normal days (i.e., $M_p \geq 1$), we randomly sample one normal day ID (from 1 to M), denoted i_p. If the passenger does not have an AFC record on the i_p-th normal day, we assume that he/she did not use public transit on the incident day. Otherwise, the AFC records on the i_p-th normal day are treated as his/her intended trajectory. We assume that, on the incident day, passenger p would follow the same travel trajectory as the i_p-th normal day (i.e., tap-in and tap-out records) if there was no incident. For all intended trajectories, the public transit trip destinations are inferred from the destination estimation model [131].

Step 2: Generate synthetic AFC data for the incident. The data from step 1 are the passenger "intended" trajectories under normal conditions. We also need to generate the "actual" AFC records subject to the incident at Sedgwick station (see Section 4.4.2). Specifically, with the intended trajectories of all passengers, we can infer their locations when the incident occurs based on a transit assignment model (see Section 4.3.2). For the purpose of model validation, we assume that passengers' behavior follows the diagram in Figure 4-2. From passengers' locations and intended routes, we can identify all affected and unaffected passengers based on whether their original routes are blocked or not. For all affected passengers, we first enumerate their possible choices based on the stage of the trip they are at when the incident occurs (e.g., at the blocked stations, in the system but not at blocked stations, outside the system, etc.) and availability of different travel modes. Then, each passenger is assigned an available mode based on the choice probabilities. For this application, the choice probabilities are calculated using the behavior model in Rahimi et al. [106] and Lin et al. [85]. If the passenger is assigned with public transit, we find the available nearby bus or alternative rail lines for him/her and calculate his/her tap in time based on the walking distance. The new tap-in record is added to the synthetic data on the incident day. For passengers who decide to wait until the system recovers, we assume they all wait outside the blocked stations and tap in right after T_2. Then new AFC records are added to the synthetic data. If the passenger is out of the system when the incident happens and is assigned with an undetected travel mode or canceling trips, we remove his/her AFC transaction in $[T_s, T_e]$. For passengers in the system deciding to cancel their trips, we remove their subsequent AFC transactions (i.e., returning trips) as assumed in Section 4.3.3. The new AFC records are treated as synthetic data on the incident day (where the incident does happen).

The synthetic AFC data on the incident day and passengers' "true" choices are then used as the ground truth for model validation. Data generation and model estimation processes are replicated 15 times.

Validation criteria

Since the proposed model can output the expected number of passengers in each behavior group (i.e., $\mathbb{E}[N_{S_i}]$) and corresponding variance (i.e., $\text{Var}[N_{S_i}]$), it is worth validating both estimates. The validation of $\mathbb{E}[N_{S_i}]$ is straightforward. As in the synthetic data we have the "true" value of N_{S_i}, a comparison between the "true" N_{S_i} and the estimated $\mathbb{E}(N_{S_i})$ can be conducted (For the benchmark model, the comparison is against $\hat{N}_{S_i}$). Since the data generation and model estimation processes are replicated 15 times, the "true" average of N_{S_i} and estimated $\mathbb{E}[N_{S_i}]$ are compared (Figure 4-8).

To validate $\text{Var}[N_{S_i}]$, we notice that the "true" N_{S_i} in each replication of the synthetic data can be seen as a sample drawn from the underlying behavioral distribution. This distribution is a reflection of passenger's choice probabilities and inferred destination distribution. Therefore, the sample variance of N_{S_i} over the 15 replications can be seen as the "true" $\text{Var}[N_{S_i}]$, which is compared with the estimated $\text{Var}[N_{S_i}]$ (Figure 4-9). Note that since we have 15 estimated $\text{Var}[N_{S_i}]$ from different replications, the average value is used for comparison.

To quantify the estimation errors over all behavior groups, we calculate the root mean square error (RMSE) and mean absolute percentage error (MAPE) as follows:

$$\text{RMSE}(\mathbb{E}[\cdot]) = \sqrt{\frac{\sum_{i=1}^{Z}(\bar{N}_{S_i} - \bar{\mathbb{E}}[N_{S_i}])^2}{Z}}, \tag{4.31}$$

$$\text{MAPE}(\mathbb{E}[\cdot]) = \frac{1}{Z}\sum_{i=1}^{Z}\frac{|\bar{N}_{S_i} - \bar{\mathbb{E}}[N_{S_i}]|}{\bar{N}_{S_i}}, \tag{4.32}$$

where $\bar{N}_{S_i}$ (resp. $\bar{\mathbb{E}}[N_{S_i}]$) is the average value of the "true" N_{S_i} (resp. estimated

$\mathbb{E}[N_{S_i}]$) over the 15 replications. The RMSE and MAPE of $\text{Var}[N_{S_i}]$ are calculated in a similar way.

Results

Model estimation results with synthetic data are shown in Figures 4-8 and 4-9. Note that we exclude the results of N_{S_6} and $N_{S_{13}}$ (number of not affected passengers) in the graph as their values are too large and may distort the comparison. In Figure 4-9, the standard deviations (i.e., $\sqrt{\text{Var}[N_{S_i}]}$) are shown instead of variance for unit consistency.

Figure 4-8 presents the estimated results of $\mathbb{E}[N_{S_i}]$ (probabilistic model) and $\hat{N}_{S_i}$ (rule-based). Results show that the probabilistic model can estimate passenger's response behaviors with an RMSE $=$ 144 and MAPE $=$ 20.5%. It significantly outperforms the rule-based benchmark model (RMSE $=$ 536 and MAPE $=$ 60.3%). The absolute errors of the probabilistic model are relatively large for $\mathbb{E}[N_{S_{16}}]$ and $\mathbb{E}[N_{S_7}]$. This may be due to the fact that there are around 30% passengers without future information for destination inference. Their destination distribution is approximated by the inferred population (Eq. 4.12), leading to estimation errors. In terms of the rule-based model, it has a system error (overestimation) because it does not account for the fact that some observed behaviors are due to behavior randomness, rather than the impact of incidents.

Figure 4-9 presents the estimation results for $\sqrt{\text{Var}[N_{S_i}]}$. Note that the rule-based model cannot output estimated variance, thus is not plotted in the figure. Results show that the probabilistic model can capture the patterns of standard deviation for different behavior groups well. The RMSE is 4.4 and MAPE is 69.8%, which is higher compared to the error of the expected values. This is reasonable because variance is the second moment which in general is harder to estimate than the first moment (i.e., expectation).

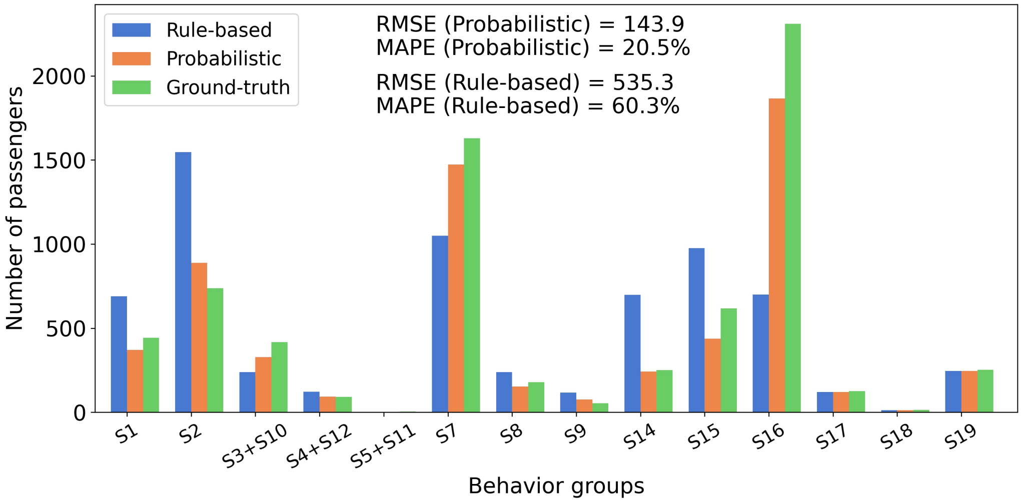


Figure 4-8: Estimation results of expectations with synthetic data

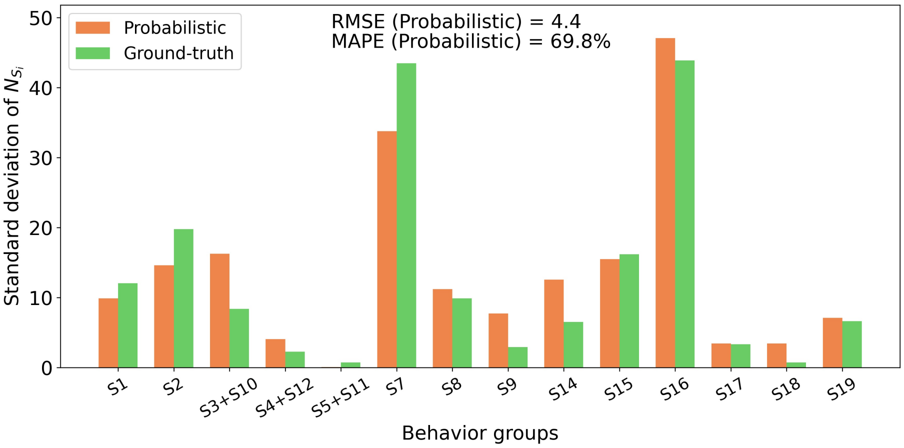


Figure 4-9: Estimation results of standard deviations with synthetic data

4.4.7 Model application with real-world data

Results

In the real-world data, we only implement the probabilistic method. Table 4.2 summarizes the estimation results for the real-world data from the CTA system. Overall, most of the passengers (97.43%) are not affected by the incident. This is reasonable because the incident only affected a small area. 69.51% of all affected passengers choose to use rail by changing routes. This is expected because the Red Line is a good substitution for the blocked Brown and Purple lines. Most of the OD pairs can be connected by the Red line when the Brown and Purple lines do not work. 6.57% of passengers choose to wait or delay their departure times (i.e., using rail without changing routes). 15.72% choose to use buses while 8.09% choose to not use public transit.

Table 4.2: Passenger behavior estimation results

Behavior (Prop.; Impacted Prop.[2])	Group	Mean	Variance (Coeff. of variation[1], %)	Proportion (%)	Proportion (Impacted[2], %)
Use rail changing route	S2	595	157.4 (2.11)	0.25	9.61
(1.79%; 69.51%)	S7	1282	1005.7 (2.47)	0.53	20.71
	S9	56	49.0 (12.5)	0.02	0.9
	S15	831	675.5 (3.13)	0.35	13.43
	S16	1538	2639.4 (3.34)	0.64	24.85
Use rail not changing route	S4+S12	48	11.5 (7.07)	0.02	0.78
(0.17%; 6.57%)	S19	365	295.9 (4.71)	0.15	5.9
Use bus	S1	315	87.8 (2.97)	0.13	5.09
(0.40%, 15.72%)	S8	202	170.5 (6.46)	0.08	3.26
	S14	456	412.9 (4.46)	0.19	7.37
Not use public transit	S3+S10	291	255.8 (5.5)	0.12	4.7
(0.21%, 8.09%)	S17	180	180.2 (7.46)	0.07	2.91
	S5+S11	10	10.4 (32.18)	0	0.16
	S18	20	20.0 (22.39)	0.01	0.32
No impact	S6	63503	1748.1 (0.07)	26.37	N.A.
(97.43%, N.A.)	S13	171085	4223.9 (0.04)	71.06	N.A.

[1]: Coefficient (Coeff.) of variation is calculated as the standard division divided by the mean.

[2]: Impacted proportion (prop.) is the proportion within all affected passengers (excluding S6 and S13).

The variance in Table 4.2 captures the behavioral randomness of S_i and how much information of S_i can be captured in B_{S_i} by AFC data (see Section 4.2.2). Generally, the variances are proportional to the means. The coefficients of variation for N_{S_1} and N_{S_2} are low, meaning these two behaviors are relatively easy to be captured by AFC data. This is reasonable because multiple tap-in records are generated by these behaviors, leading to the direct incident-related B_{S_i}. Canceling trips and using other undetected modes have a relatively high coefficient of variation. This means these two behaviors are hard to be estimated using the AFC data.

Figure 4-10 shows the behavior distribution for passengers **in** the rail system when the incident happened. 46% of those passengers choose the inside rail transfer (i.e. transfer without leaving the rail system). This is reasonable because passengers coming from stations north of the blocked stations (main morning peak demand) have multiple rail transfer stations (such as Belmont and Fullerton) that connect the suspended Brown and Purple lines to the Red line (see Figure 4-5). This allows passengers to conveniently continue to use the rail system without leaving the system. 19% and 23% of passengers choose to leave the system and transfer to a bus line and other rail stations, respectively. Around 10% of passengers choose to use other undetected modes. And only a small proportion of passengers choose to wait (2%) or cancel their trips (0.3%). Overall, the estimated proportions of different behaviors are reasonable.

Figure 4-11 shows the behavior distribution for all affected passengers **out of** the rail system when the incident happened. Similar to the results above, most of those passengers (45%) chose to transfer to another rail line without leaving the system. 25% of them changed tap-in stations and 13% chose to use buses. We also observe that 11% of passengers delayed their departure time and 5% used other undetected modes. Only around 1% of passengers canceled their trips. Compared to the results above, we find there is a decrease in the percentage of passengers using buses and other undetected modes and an increase in using rail. This is reasonable because when passengers are out of the system, they are more flexible in choosing rail routes, thus more likely to keep using rail services.

Analysis of real-world results

Though there is no direct validation for the estimation results using real-world data, we propose two indirect approaches to discuss the reasonableness of the results.

The first is to compare the ridership increase in bus stops and rail stations that are close to the blocked stations (i.e., $\mathcal{W}_b$ and $\mathcal{W}_r$). The ridership increase at these bus stops and rail stations should be similar to (slightly larger than) N_{S_1} and N_{S_2}, respectively. "Slightly larger" is because some ridership increase may be passengers

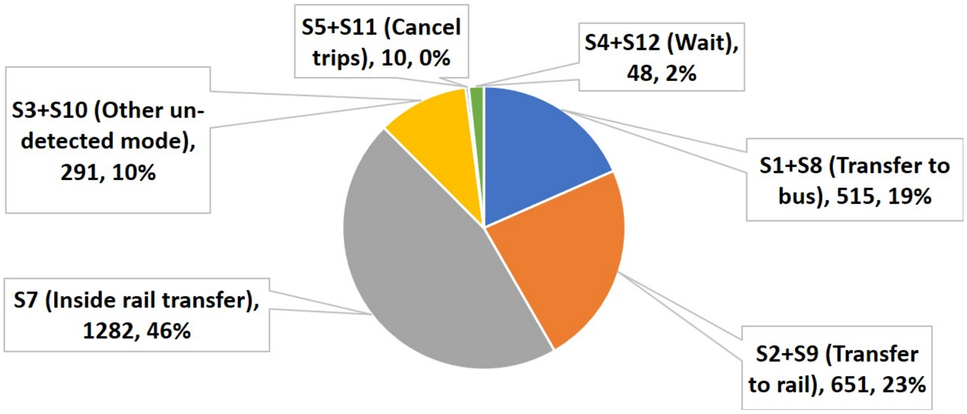


Figure 4-10: Behavior distribution for passengers in the rail system when the incident happened (texts in boxes are behavior description, number of passengers, and proportion, respectively).

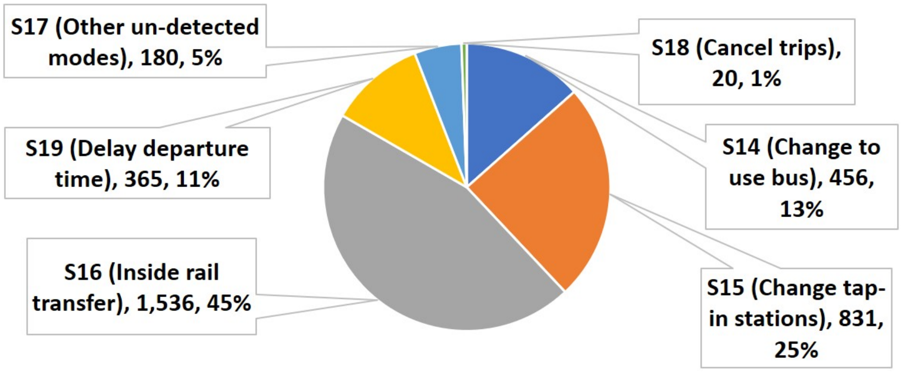


Figure 4-11: Behavior distribution for passengers out of the rail system when the incident happened.

living in the nearby neighborhoods, which do not belong to S_1 and S_2. The ridership increase is calculated as the number of tap-in passengers during the incident period minus the mean on normal days. The ridership increase for nearby bus stops is 401 passengers (slightly larger than the estimated $\mathbb{E}[N_{S_1}] = 315$), and for rail stations 720 passengers (slightly larger than the estimated $\mathbb{E}[N_{S_2}] = 595$), which is as expected.

The second approach is based on the CTA incident logs. CTA incident logs report that "run 505 (Purple line) unloads around 300 customers" and "run 416 (Brown line) unloads around 500 customers". According to the AVL data, these two trains are the only trains that unloaded passengers. Assuming that passengers who entered the blocked stations between T_1 and the time of the last train departure waited at the platforms, there were a total of 437 waiting passengers on the platforms of the blocked stations when the incident happened (based on the AFC and AVL data). According to Figure 4-2, the total number of unloaded and waiting passengers should be equal to the number of passengers at the blocked stations (i.e., $\sum_{i=1}^{5} N_{S_i}$). Hence, the estimated value of $\sum_{i=1}^{5} \mathbb{E}[N_{S_i}]$ should be close to $300 + 500 + 437 = 1,237$ passengers. However, the inference model provides estimates for $\mathbb{E}[N_{S_1}]$ and $\mathbb{E}[N_{S_2}]$, but not $\mathbb{E}[N_{S_3}]$, $\mathbb{E}[N_{S_4}]$, and $\mathbb{E}[N_{S_5}]$ (because $\mathbb{E}[N_{S_3}+N_{S_{10}}]$, $\mathbb{E}[N_{S_4}+N_{S_{12}}]$, and $\mathbb{E}[N_{S_5}+N_{S_{11}}]$ are estimated as a whole). Since $\mathbb{E}[N_{S_4}+N_{S_{12}}]$ and $\mathbb{E}[N_{S_5}+N_{S_{11}}]$ are relatively small, $\sum_{i=1}^{5} \mathbb{E}[N_{S_i}]+\mathbb{E}[N_{S_{10}}]+\mathbb{E}[N_{S_{11}}]+\mathbb{E}[N_{S_{12}}]$ should be slightly greater than 1,237 and $\mathbb{E}[N_{S_1}]+\mathbb{E}[N_{S_2}]$ slightly smaller than 1,237 if the estimates are correct. A simple calculation leads to

$$\mathbb{E}[N_{S_1} + N_{S_2}] = 910 < 1237 < \sum_{i=1}^{5} \mathbb{E}[N_{S_i}] + \mathbb{E}[N_{S_{10}}] + \mathbb{E}[N_{S_{11}}] + \mathbb{E}[N_{S_{12}}] = 1259, \tag{4.33}$$

supporting the validity of the estimation results.

4.5 Conclusion and discussion

This study proposes a probabilistic framework to infer passengers' response behavior to an unplanned rail service disruption using smart card data in a tap-in-only public transit system. We enumerate 19 possible behaviors that passengers may have based on the stages of their trips when the incident happened. A probabilistic model is proposed to estimate the mean and variance of the number of passengers in each of the 19 groups using passengers' historical and subsequent trip information. Based on the information used and the context of the behavior, four cases of formulations are used in the probabilistic model. Data from the CTA public transit system (bus and urban rail) is used for the case study with a rail incident. The model is implemented with both synthetic data (consistent with the CTA AFC data) and real-world data. The main conclusions of this study are as follows:

- The proposed approach can estimate passengers' behavior well and outperform the rule-based benchmark model. Results with synthetic data show that the RMSE and MAPE for the estimated expected number of passengers in each behavior group are 143.9 and 20.5%, respectively. The RMSE and MAPE for the estimated standard deviation are 4.4 and 69.8%, respectively. The estimation results with real-world data are consistent with the incident's context. An indirect model validation using ridership change information and incident log data demonstrates the reasonableness of the results.

- Results with real-world data find that most of the passengers (97.43%) are not affected by the incident. This is reasonable because the incident only affected a small area. The incident we analyzed has high service redundancy with the Red line substituting the blocked Brown and Purple lines. Our model results show that in the high redundancy case, most of the affected passengers (69.51%) choose to use rail by changing routes.

- Based on the results, CTA operators can confirm that the Red line is a good alternative and quantify the impact. To relieve the incident impact, operators can increase service frequency in the Red line. The model indicates that only 8.1% of passengers choose to leave the public transit system. This number can help CTA conduct the service loss analysis due to the incident.

The proposed model has several practical significances. First, The model is data-driven. Compared to the conventional survey-based methods, the proposed approach can effectively estimate passengers' responses without collecting data manually. Second, the output results can help transit operators better understand passengers' choices during a disruption, based on which they can design better operating strategies on the supply side to mitigate the impact of incidents. For example, for heavily used alternative services during the disruption, operators can increase the service frequency or provide shuttle buses with similar routes. Third, based on the results, operators can identify congestion in the network. They can disseminate information (e.g., route recommendation) to passengers, or conduct flow control at the gate level, to avoid overloaded routes.

Future studies can focus on the following directions. 1) Estimate the estimation error (i.e., $\text{Var}[\mathbb{E}[N_{S_i}]]$). The estimation error is another type of uncertainty. It comes from the fact that we are using sample data to estimate a specific probability. For example, as $\mathbb{P}(A_p \mid p \in B_{S_i})$ is estimated from the historical travel trajectories, we actually only obtain the estimated value (i.e., $\hat{\mathbb{P}}(A_p \mid p \in B_{S_i})$). It is a random variable and the corresponding variance $\text{Var}[\hat{\mathbb{P}}(\cdot)]$ reflects the estimation error. The estimation error depends on sample sizes (i.e., amount of normal day data) and passenger travel irregularity. The challenge of estimating $\text{Var}[\hat{\mathbb{P}}(\cdot)]$ is that this value is not available for passengers without historical information. Future studies can explore approximation techniques with reasonable distributional assumptions to calculate estimation errors. 2) Apply the model to different incident cases. According to Mo et al. [145], the incident locations and the redundancy of surrounding public transit alternatives are influential in passenger mode choice behavior. Future studies may analyze more case studies and compare passengers' behavioral responses under different scenarios. 3) Analyze individual-level choices. In this study, we only output the aggregate level mode choice behavior (i.e., N_{S_i}). Future work may explicitly output $\mathbb{P}(p \in S_i)$, and analyze its relationship with passengers' characteristics (such as home location, fare card type, travel frequency, etc.).

These future studies can help improve the proposed method, and make the understanding of passenger responses more accurate. For example, with better quantification of estimation uncertainty, we can develop more robust or stochastic optimization methods for shuttle service design, headway design, path recommendations, flow control, etc.

Though there are extensive data in the AFC and AVL systems, machine learning methods do not fit into this study because of the lack of actual observed responses behaviors (i.e., lack of labels). In the future, if some passenger's actual response behavior can be observed (e.g., from self-report data, probe GPS data, or cell phone data), a supervised learning model may be trained to predict passengers' responses to incidents. The features may include passenger's spatial and temporal travel histories, incident information, and supply information. These features can be embedded with many advanced deep learning methods such as long short-term memory (LSTM) networks, convolutional neural networks (CNN), graph neural networks (GNN), etc

4.6 Appendix: Model formulations (continued)

4.6.1 Historical trip information + direct incident-related B_{S_i} (continued)

Inferring S_4 and S_{12}

S_4 and S_{12} are passengers who waited until the system recovered. Passengers in S_4 left and waited outside the blocked stations. Thus, they have at least one tap-in record before T_1, and another tap-in record at the blocked stations after T_2. We assume that passengers in S_{12} also waited outside the blocked stations (passengers usually take the train up to the blocked stations then start to wait).

We define $B_{S_{4,12}} = \{p : \exists k \in \{1, ..., K_p - 1\} \text{ s.t. } t_{p_k} \leq T_1,\ t_{p_{k+1}} \geq T_2,\ m_{p_k} = m_{p_{k+1}} = \text{rail},\ o_{p_{k+1}} \in \mathcal{W}\}$, which means passengers with a rail tap-in trip before the incident and another rail tap-in trip after the system recovery, with the second tap-in station one of the blocked stations. As passengers who tap-in again at a blocked station may do so not because of the incident but as part of a normal routine, similar to Eq. 4.6,

$$\mathbb{E}[N_{S_4} + N_{S_{12}}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{4,12}}\}} \cdot \mathbb{P}(A_p \mid p \in B_{S_{4,12}}). \tag{4.34}$$

$\mathbb{1}_{\{p \in B_{S_{4,12}}\}}$ is a constant. $\mathbb{P}(A_p \mid p \in B_{S_{4,12}})$ can be calculated the same way as Eq. 4.7 and 4.8 by replacing B_{S_1} with $B_{S_{4,12}}$. The variance of $N_{S_4} + N_{S_{12}}$ is

$$\text{Var}[N_{S_4} + N_{S_{12}}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{4,12}}\}} \cdot [\mathbb{P}(A_p \mid p \in B_{S_{4,12}}) - \mathbb{P}(A_p \mid p \in B_{S_{4,12}})^2] \tag{4.35}$$

4.6.2 Historical trip information + indirect incident-related B_{S_i} (continued)

Inferring S_{14} and S_{15}

Passengers in S_{14} and S_{15} have not entered the rail system when the incident happens. Therefore, they have no rail tap-in records before T_1.

We first consider passengers in S_{14}. Consider a $p \in \mathcal{P}^H$ with $T_1 < t_{p_1} < T_2$ and $m_{p_1} = \text{bus}$, which means the first trip for p during the incident period is bus. Define $B_{S_{14}} = \{p : T_1 < t_{p_1} < T_2, m_{p_1} = \text{bus}\}$. And define the event that p changed from rail to bus on the incident day as CB_p. The probability of CB_p for $p \in B_{S_{14}}$ can be calculated as

$$\mathbb{P}(CB_p \mid p \in B_{S_{14}}) = 1 - \frac{\#\text{ normal days } p\text{'s first trip in } [T_1, T_2] \text{ is bus}}{\#\text{ normal days } p \text{ has trips in } [T_1, T_2]} \quad \forall p \in \mathcal{P}^H \tag{4.36}$$

Eq. 4.36 means the probability of CB_p equals 1 minus the frequency of using a bus on normal days. A high frequency of using a bus on normal days means using a bus is highly likely the typical behavior for p, instead of a change in the behavior. For $p \notin \mathcal{P}^H$, we can approximate the probability by

$$\mathbb{P}(CB_p \mid p \in B_{S_{14}}) = \frac{\sum_{p' \in \mathcal{P}^H \cap \mathcal{P}^F} \mathbb{P}(CB_{p'} \mid p \in B_{S_{14}})}{|\mathcal{P}^H \cap B_{S_{14}}|} \quad \forall p \notin \mathcal{P}^H \tag{4.37}$$

However, passengers may change from rail to bus due to many reasons, not necessarily because of the incident. Similar to Eq. 4.19, we have

$$\begin{aligned}\mathbb{E}[N_{S_{14}}] &= \sum_{p \in \mathcal{P}} \mathbb{E}[\mathbb{1}_{\{p \in S_{14}\}}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{14}}\}} \cdot \mathbb{P}(CB_p, A_p \mid p \in B_{S_{14}}) \\ &= N_{CB} - \tilde{N}_{CB}\end{aligned} \tag{4.38}$$

where $N_{CB} := \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{14}}\}} \cdot \mathbb{P}(CB_p \mid p \in B_{S_{14}})$ is the expected number of passengers who change from rail to bus on the incident day. $\tilde{N}_{CB} := \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{14}}\}} \cdot$

$\mathbb{P}(CB_p \mid p \in B_{S_{14}})\mathbb{P}((A_p)^c \mid CB_p, p \in B_{S_{14}})$ is the expected number of passengers who change from rail to bus but *not* because of the incident. It can be approximated by the number of passengers changing from rail to bus on normal days. Similar to Eq. 4.22,

$$\mathbb{E}[N_{S_{14}}] = N_{CB} - \tilde{N}_{CB} = N_{CB} - \frac{\sum_{j=1}^{M} \tilde{N}_{CB}^{(j)}}{M} \tag{4.39}$$

where $\tilde{N}_{CB}^{(j)}$ is the number of passengers changing from rail to bus on the j-th normal day, calculated with the same method of calculating N_{CB} but using the j-th normal day AFC data. Similar to to Eq. 4.23 and 4.24, we can calculate the variance of $N_{S_{14}}$ as:

$$\begin{aligned}\text{Var}[N_{S_{14}}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{14}}\}} \cdot [&\mathbb{P}(CB_p \mid p \in B_{S_{14}})\mathbb{P}(A_p \mid CB_p, p \in B_{S_{14}}) - \\ &\mathbb{P}(CB_p \mid p \in B_{S_{14}})^2 \mathbb{P}(A_p \mid CB_p, p \in B_{S_{14}})^2]\end{aligned} \tag{4.40}$$

where $\mathbb{P}(A_p \mid CB_p, p \in B_{S_{14}}) = (N_{CB} - \tilde{N}_{CB})/N_{CB}$ for all $p \in B_{S_{14}}$ (similar to Eq. 4.23).

For passengers in S_{15}, we define $B_{S_{15}} = \{p : T_1 < t_{p_1} < T_2, m_{p_1} = \text{rail}\}$, and denote the event that p changes tap-in station to o_{p_1} on the incident day as CS_p. Similar to Eq. 4.36, we have

$$\mathbb{P}(CS_p \mid p \in B_{S_{15}}) = 1 - \frac{\#\text{ normal days that } p\text{'s first rail tap-in station in } [T_1, T_2] \text{ is } o_{p_1}}{\#\text{ normal days that } p \text{ has rail trips in } [T_1, T_2]} \quad \forall p \in \mathcal{P}^H \tag{4.41}$$

Analogue to the estimation of $\mathbb{E}[N_{S_{14}}]$, we have

$$\begin{aligned}\mathbb{E}[N_{S_{15}}] &= \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{15}}\}} \cdot \mathbb{P}(CS_p, A_p \mid p \in B_{S_{15}}) \\ &= N_{CS} - \tilde{N}_{CS}\end{aligned} \tag{4.42}$$

where $N_{CS} := \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{15}}\}} \cdot \mathbb{P}(CS_p \mid p \in B_{S_{15}})$ is the expected number of passengers changing tap-in stations on the incident day. And $\tilde{N}_{CS} := \frac{\sum_{j=1}^{M} \tilde{N}_{CS}^{(j)}}{M}$, where $\tilde{N}_{CS}^{(j)}$ is the number of passengers changing tap-in stations on the j-th normal day, which is calculated with the same method as calculating N_{CS} using the AFC data on the j-th normal day. The variance of $N_{S_{15}}$ can be calculated the same way as in Eq. 4.40 by replacing $B_{S_{14}}$ and CB_p by $B_{S_{15}}$ and CS_p, respectively.

Inferring S_{17} and S_{18}

Passengers who decided to use other undetected modes or cancel trips after the incident (i.e., S_{17} and S_{18}) have no rail tap-in records between T_1 and T_e on the incident day. The inference is based on passengers who were supposed to have tap-in records in this period according to their behavior on normal days. Define $B_{S_{17,18}} = \{p : p \text{ has rail tap-in records within } [T_1, T_e] \text{ on any of the } M_p \text{ normal days, but not on the incident day}\}$. These are potential passengers who might change to other undetected modes or cancel trips on the incident day. Due to the nature of the AFC data, there is no direct way to differentiate these two groups. We assume that the probability of a passenger p using other undetected modes in this situation is β_p, that is, $\mathbb{P}(\mathbb{1}_{\{UMOS_p\}} = 1) = \alpha_{p,d}$, where $UMOS_p$ is the event that passenger p will other undetected modes when he/she is outside the system. The value of β_p can be obtained from previous survey-based studies (similar to α_p). Note that if we focus on the aggregate estimation of the passengers who do not use public transit (i.e., canceling trips + using other undetected modes), the value of β_p is not needed.

Consider a passenger $p \in B_{S_{17,18}}$. As in section 4.3.3, we assume that if p has a high probability of having trips in $[T_1, T_e]$ on normal days, then the disappearance of the trip on the incident day is highly likely an atypical behavior (i.e., canceling the trip or switching to undetected modes). Define the event that p canceled the trip or switched to undetected modes on the incident day as $CTSM_p$. According to the assumption above and Eq. 4.17:

$$\mathbb{P}(CTSM_p \mid p \in B_{S_{17,18}}) = \frac{\#\text{ normal days } p \text{ having rail trips in } [T_1, T_e]}{M_p} \qquad \forall p \in \mathcal{P}^H \tag{4.43}$$

However, as p may cancel the trip or switch to other undetected modes for other reasons, not necessarily due to the incident. We have $\mathbb{1}_{\{p \in S_{17}\}} = \mathbb{1}_{\{p \in B_{S_{17,18}}\}} \cdot \mathbb{1}_{\{CTSM_p \cap A_p \mid p \in B_{S_{17,18}}\}} \cdot \mathbb{1}_{\{UMOS_p\}}$. Since $\mathbb{1}_{\{UMOS_p\}}$ and $\mathbb{1}_{\{CTSM_p \cap A_p \mid p \in B_{S_{17,18}}\}}$ are independent, similar to Eq. 4.19 - 4.21, we have

$$\begin{aligned}\mathbb{E}[N_{S_{17}}] &= \sum_{p \in \mathcal{P}} \mathbb{E}[\mathbb{1}_{\{p \in S_{17}\}}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{17,18}}\}} \cdot \beta_p \cdot \mathbb{P}(CTSM_p, A_p \mid p \in B_{S_{17,18}}) \\ &= N_{CTSM1} - \tilde{N}_{CTSM1}\end{aligned} \tag{4.44}$$

where $N_{CTSM1} := \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{17,18}}\}} \cdot \beta_p \cdot \mathbb{P}(CTSM_p \mid p \in B_{S_{17,18}})$ is the expected number of passengers using other undetected modes on the incident day (not necessarily due to the incident). $\tilde{N}_{CTSM1} := \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{17,18}}\}} \cdot \beta_p \cdot \mathbb{P}(CTSM_p \mid p \in B_{S_{17,18}}) \cdot \mathbb{P}((A_p)^c \mid CTSM_p,\ p \in B_{S_{17,18}})$ is the expected number of passengers using other undetected modes and the reason is not the incident day, which can be approximated by the number of passengers using other undetected modes on normal days: $\tilde{N}_{CTSM1} = \frac{\sum_{j=1}^{M} \tilde{N}_{CTSM1}^{(j)}}{M}$, where $\tilde{N}_{CTSM1}^{(j)}$ is the expected number of passengers using other undetected modes on the j-th normal day, which is calculated with the same method as calculating N_{CTSM1} but with the AFC data on the j-th normal day.

And the variance of N_{17} is

$$\begin{aligned}\text{Var}[N_{S_{17}}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{17,18}}\}} \cdot [&\beta_p \cdot \mathbb{P}(CTSM_p \mid p \in B_{S_{17,18}}) \cdot \mathbb{P}(A_p \mid CTSM_p, p \in B_{S_{17,18}}) - \\ &\beta_p^2 \cdot \mathbb{P}(CTSM_p \mid p \in B_{S_{17,18}})^2 \cdot \mathbb{P}(A_p \mid CTSM_p, p \in B_{S_{17,18}})^2]\end{aligned} \tag{4.45}$$

where

$$\mathbb{P}(A_p \mid CTSM_p, p \in B_{S_{17,18}}) = \frac{N_{CTSM} - \tilde{N}_{CTSM}}{N_{CTSM}}. \quad \forall p \in B_{S_{17,18}} \tag{4.46}$$

$N_{CTSM}, \tilde{N}_{CTSM}$ are calculated the same way as $N_{CTSM1}, \tilde{N}_{CTSM1}$ by replacing β_p to 1.

Similarly, for passengers in S_{18}, $\mathbb{E}[N_{S_{18}}]$ and $\text{Var}[N_{S_{18}}]$ are calculated the same way as $\mathbb{E}[N_{S_{17}}]$ and $\text{Var}[N_{S_{17}}]$, respectively, by replacing β_p to $1 - \beta_p$.

Inferring S_{19}

Passengers in S_{19} are those who continued to use their original routes but delayed their departure times. In this study, we define "delay departure time" as departing $2\sigma_p$ later than μ_p, where μ_p is the mean departure time of p's first rail trip in the analysis period on the normal days, and σ_p is the corresponding standard deviation. μ_p and σ_p can be calculated using the tap-in times of previous rail trips at station o_{p_1} on normal days. We define $B_{S_{19}} = \{p : t_{p_1} \geq T_2,\ m_{p_1} = \text{rail},\ t_{p_1} > \mu_p + 2\sigma_p\}$, which is the set of passengers who delayed their departure times and departed after T_2 (i.e., after system recovery). However, as passengers may delay departure time for different reasons, not necessarily because of the incidents, similar to Eq. 4.19, we have

$$\begin{aligned}\mathbb{E}[N_{S_{19}}] &= \sum_{p\in\mathcal{P}} \mathbb{E}[\mathbb{1}_{\{p\in S_{19}\}}] = \sum_{p\in\mathcal{P}} \mathbb{1}_{\{p\in B_{S_{19}}\}} \cdot \mathbb{P}(A_p \mid p \in B_{S_{19}}) \\ &= \sum_{p\in\mathcal{P}} \mathbb{1}_{\{p\in B_{S_{19}}\}}[1 - \mathbb{P}((A_p)^c \mid p \in B_{S_{19}})] \\ &= N_{DD} - \tilde{N}_{DD}\end{aligned} \tag{4.47}$$

where $N_{DD} := \sum_{p\in\mathcal{P}} \mathbb{1}_{\{p\in B_{S_{19}}\}}$ is the expected number of passengers who delayed departure time on the incident day. And $\tilde{N}_{DD} := \sum_{p\in\mathcal{P}} \mathbb{1}_{\{p\in B_{S_{19}}\}}\mathbb{P}((A_p)^c \mid p \in B_{S_{16}})$ is the expected number of passengers who delayed departure time but *not* because of the incident, which can be approximated by the number of passengers delaying departure time on normal days. Therefore, similar to Eq. 4.22, we have

$$\mathbb{E}[N_{S_{16}}] = N_{DD} - \tilde{N}_{DD} = N_{DD} - \frac{\sum_{j=1}^{M} \tilde{N}_{DD}^{(j)}}{M} \tag{4.48}$$

where $\tilde{N}_{DD}^{(j)}$ is the number of passengers delaying departure time on j-th normal day, calculated with the same method of calculating N_{DD} but using the j-th normal day AFC data. Similar to Eq. 4.23 and 4.24, we can calculate the variance of $N_{S_{19}}$ as:

$$\text{Var}[N_{S_{19}}] = \sum_{p\in\mathcal{P}} \mathbb{1}_{\{p\in B_{S_{19}}\}}[\mathbb{P}(A_p \mid p \in B_{S_{19}}) - \mathbb{P}(A_p \mid p \in B_{S_{19}})^2] \tag{4.49}$$

where $\mathbb{P}(A_p \mid p \in B_{S_{16}}) = (N_{DD} - \tilde{N}_{DD})/N_{DD}$ as per Eq. 4.23.

4.6.3 Subsequent trip information only (continued)

Inferring S_{16}

Passengers in S_{16} are those who did not change tap-in stations, but chose to transfer halfway to avoid the blocked stations. We assume that passengers who make decisions after the incident are informed of the service interruption. Hence, if they decided to still use rail between T_1 and T_2, the possible situations for them are 1) changing tap-in station (S_{15}), 2) choosing alternative routes by transferring (S_{16}), and 3) not affected. Let $B_{S_{16}} = \{p : T_1 < t_{p_1} < T_2,\ m_{p_1} = \text{rail}\}$, which means passengers with a rail trip during the incident time. We notice that the third possibility can be excluded if we find that a passenger's original path is blocked. Therefore, for all passengers in $\mathcal{P}^F$, we first infer their destinations based on the next non-transfer trip after $(t_{p_1}, o_{p_1}, m_{p_1})$ (see Section 4.3.2). Given an inferred destination $d \in \mathcal{D}_{p_1}$, denote the event that p's original path is blocked but a transfer option is available as $RBTA_p(d)$. From the above analysis, we know that all passengers in $B_{S_{16}}$ and with $\mathbb{1}_{\{RBTA_p(d)\}} = 1$ can only be in S_{15} and S_{16}. Define $N_{B_{S_{16}}\cap RBTA} := \sum_{p\in\mathcal{P}} \sum_{d\in\mathcal{D}_{p_1}} \mathbb{1}_{\{p\in B_{S_{16}}\}} \cdot \mathbb{1}_{\{RBTA_p(d)\}} \cdot \mathbb{1}_{\{\tilde{d}_{p_1}=d\}}$, which is the number of passengers with a rail trip during the incident and the original route blocked. Therefore, the mean of $N_{S_{16}}$ can be calculated as:

$$\mathbb{E}[N_{S_{16}}] = \mathbb{E}[N_{B_{S_{16}} \cap RBTA} - N_{S_{15}}] = \sum_{p \in \mathcal{P}} \sum_{d \in \mathcal{D}_{p_1}} \mathbb{1}_{\{p \in B_{S_{16}}\}} \cdot \mathbb{1}_{\{RBTA_p(d)\}} \cdot \mathbb{P}(\tilde{d}_{p_1} = d) - \mathbb{E}[N_{S_{15}}] \tag{4.50}$$

where $\mathbb{E}[N_{S_{15}}]$ is estimated as in Section 4.6.2. To calculate $\mathrm{Var}[N_{S_{16}}]$, we notice that the covariance between $N_{B_{S_{16}} \cap RBTA}$ and $N_{S_{15}}$ is zero:

$$\begin{aligned}
&\mathrm{Cov}[N_{15}, N_{B_{S_{16}} \cap RBTA}] \\
&= \mathrm{Cov}\left[\sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{15}}\}} \cdot \mathbb{1}_{\{CS_p, A_p \mid p \in B_{S_{15}}\}}, \sum_{p \in \mathcal{P}} \sum_{d \in \mathcal{D}_{p_1}} \mathbb{1}_{\{p \in B_{S_{16}}\}} \cdot \mathbb{1}_{\{RBTA_p(d)\}} \cdot \mathbb{1}_{\{\tilde{d}_{p_1} = d\}}\right] \\
&= 0
\end{aligned} \tag{4.51}$$

This is based on the observation that $\mathrm{Cov}[\mathbb{1}_{\{CS_p, A_p \mid p \in B_{S_{15}}\}}, \mathbb{1}_{\{\tilde{d}_{p'_1} = d\}}] = 0$ for all $p, p' \in \mathcal{P}$ (even if $p = p'$, this still holds because the derivation of destination relies on future information while the derivation of atypical behavior relies on historical information). Hence, the variance of $N_{S_{16}}$ can be estimated as.

$$\mathrm{Var}[N_{S_{16}}] = \sum_{p \in \mathcal{P}} \sum_{d \in \mathcal{D}_{p_1}} \mathbb{1}_{\{p \in B_{S_{16}}\}} \cdot \mathbb{1}_{\{RBTA_p(d)\}} \cdot [\mathbb{P}(\tilde{d}_{p_1} = d) - \mathbb{P}(\tilde{d}_{p_1} = d)^2] + \mathrm{Var}[N_{S_{15}}] \tag{4.52}$$

4.6.4 Other

Passengers in S_6 and S_{13} are those who are not affected by the incident. They are inferred based on the results of other groups, which do not belong to any formulation cases and thus are described separately in this section.

Inferring S_6 and S_{13}

Passengers in S_6 are those who were not affected by the incident even though they were in the rail system while the incident happened. According to the diagram in

Figure 4-2, we can infer N_6 as all passengers in the rail system subtracting other subgroups of passengers given the mutually exclusive definition. Define $B_{S_6} = \{p : \exists k \in \{1, ..., K_p\} \text{ s.t. } t_{p_k} < T_1,\ m_{p_k} = \text{rail}\}$, which means all passengers who might be in the rail system when the incident happened. Therefore, we have

$$\mathbb{E}[N_{S_6}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_6}\}} - \sum_{i=1}^{5} \mathbb{E}[N_{S_i}] - \sum_{i=7}^{12} \mathbb{E}[N_{S_i}] \tag{4.53}$$

Note that $\mathbb{E}[N_{S_3} + N_{S_{10}}]$, $\mathbb{E}[N_{S_4} + N_{S_{12}}]$, and $\mathbb{E}[N_{S_5} + N_{S_{11}}]$ are calculated as a whole (see Sections 4.3.4, 4.3.3, and 4.6.1).

The calculation of variance needs to consider the possible correlation among N_{S_i}. First of all, B_{S_1}, B_{S_2}, B_{S_8}, B_{S_9}, and $B_{S_{4,12}}$ do not intersect with other B_{S_i}'s , which implies N_{S_1}, N_{S_2}, N_{S_8}, N_{S_9}, and $N_{S_4} + N_{S_{12}}$ are independent and they are also independent of other N_{S_i}'s (because the behavior of different passengers is assumed to be independent). As shown in Section 4.3.4, the inference of $N_{S_5} + N_{S_{11}}$ uses the historical trip while the inference of $N_{S_3} + N_{S_{10}}$ and N_{S_7} relies on the information of subsequent trips (after the incident). Hence, $N_{S_5} + N_{S_{11}}$ is independent of $N_{S_3} + N_{S_{10}}$ and N_{S_7}. Then, the variance of N_{S_6} can be calculated as

$$\begin{aligned} \text{Var}[N_{S_6}] = &\sum_{i=1}^{2} \text{Var}[N_{S_i}] + \sum_{i=8}^{9} \text{Var}[N_{S_i}] + \text{Var}[N_{S_4} + N_{S_{12}}] \\ &+ \text{Var}[N_{S_5} + N_{S_{11}}] + \text{Var}[N_{S_3} + N_{S_{10}} + N_{S_7}] \end{aligned} \tag{4.54}$$

Note that the variance of $N_{S_3} + N_{S_{10}} + N_{S_7}$ can be calculated as a whole according to Section 4.3.4:

$$\begin{aligned} &\text{Var}[N_{S_3} + N_{S_{10}} + N_{S_7}] \\ &= \sum_{p \in \mathcal{P}} \sum_{d \in \mathcal{D}_{p_K}} \mathbb{1}_{\{p \in B_{S_{3,7,10}}\}} \cdot [\mathbb{1}_{\{RBTA_p(d)\}} + \mathbb{1}_{\{RBTN_p(d)\}}] \cdot [\mathbb{P}(\tilde{d}_{p_K} = d) - \mathbb{P}(\tilde{d}_{p_K} = d)^2] \\ &\quad + \text{Var}[N_{S_5} + N_{S_{11}}] \end{aligned} \tag{4.55}$$

$N_{S_{13}}$ can be inferred in a similar way as the total number of potentially affected passengers outside the system minus the number of passengers in other groups. It is worth noting that the potentially affected passengers include those who do not have tap-in records on the incident day (e.g., $B_{S_{17,18}}$). Define $B_{S_{13}} = \{p : t_{p_1} \geq T_1\} \cup B_{S_{17,18}}$ as the set of passengers outside the system who were potentially affected. Then,

$$\mathbb{E}[N_{S_{13}}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{13}}\}} - \sum_{i=14}^{19} \mathbb{E}[N_{S_i}] \tag{4.56}$$

The variance of $N_{S_{13}}$ also needs to consider the correlations. Notice that $B_{S_{14}}$ and $B_{S_{19}}$ do not intersect with other B_{S_i}'s. So, $N_{S_{14}}$ and $N_{S_{19}}$ are independent of other N_{S_i}'s. According to 4.6.2, the variance of $N_{S_{17}} + N_{S_{18}}$ can be estimated as a whole:

$$\begin{aligned}\text{Var}[N_{S_{17}} + N_{S_{18}}] = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p \in B_{S_{17,18}}\}} \cdot [&\mathbb{P}(CTSM_p \mid p \in B_{S_{17,18}}) \cdot \mathbb{P}(A_p \mid CTSM_p, p \in B_{S_{17,18}}) - \\ &\mathbb{P}(CTSM_p \mid p \in B_{S_{17,18}})^2 \cdot \mathbb{P}(A_p \mid CTSM_p, p \in B_{S_{17,18}})^2]\end{aligned} \tag{4.57}$$

And from 4.6.3, $N_{S_{15}}$ and $N_{S_{16}}$ are independent, and independent of $N_{S_{17}} + N_{S_{18}}$ since $B_{S_{17,18}}$ does not intersect with $B_{S_{15}}$ or $B_{S_{16}}$. Therefore, The variance of $N_{S_{13}}$ can be estimated as:

$$\text{Var}[N_{S_{13}}] = \sum_{i=14}^{16} \text{Var}[N_{S_i}] + \text{Var}[N_{S_{17}} + N_{S_{18}}] + \text{Var}[N_{S_{19}}] \tag{4.58}$$

Chapter 5

Station-based path recommendations during public transit disruptions under demand uncertainty

5.1 Introduction

5.1.1 Background

Public transit (PT) systems play an important role in urban mobility. However, with aging systems, continuous expansion, and near-capacity operations, service disruptions often occur. These incidents may result in delays, cancellation of trips, and economic losses [6].

This study considers significant service disruptions in public transit systems where the service (or line/route) is interrupted for a relatively long period of time (e.g., 1 hour). During a disruption, affected passengers need to find an alternative path or use other travel modes (such as transfer to another bus route). However, due to a lack of knowledge of the system state (especially during incident time), the alternative routes chosen by passengers may not be optimal or even cause more congestion [145]. For example, during a rail disruption, most of the passengers may choose bus routes that are parallel to the interrupted rail line as an alternative. However, given the limited capacity of buses, the parallel bus line may be over-saturated and passengers have to wait for a long time to board due to being denied boarding (or left behind).

5.1.2 Objectives and Challenges

One of the strategies to better guide passengers is to provide path recommendations so that the passenger flows are re-distributed in a better way and the system travel times are reduced. This can be seen as solving an **optimal passenger flow distribution (or assignment) problem** over a public transit network. However, there are several challenges to this problem.

- First, as the objective is to reduce the system travel time, an analytical formulation to calculate passengers' travel times is needed. However, a passenger's waiting times at the boarding and transfer stations are not only determined by other waiting passengers, but also those who already boarded the same line as they reduce the vehicle's capacity [152]. This complicated interaction makes it difficult to have an analytical formulation for passenger's travel time when the left behind is not negligible (which is usually the case during service disruptions). More details on this challenge are elaborated in Section 5.2.3.

- Second, there are many uncertainties in the system, such as the number of passengers using the PT system during incidents (i.e., demand uncertainty), incident duration, and whether passengers would follow the recommendations or not (i.e., behavior uncertainty). Previous studies have not considered uncertainties in modeling an optimal passenger flow problem.

This study aims to propose a path recommendation model to reduce the crowding during public transit disruptions, also taking into account uncertainties due to inaccurate demand estimates. Different from previous recommendation systems that focus on maximizing individual preferences, this study targets a system objective by minimizing the total travel time of all passengers (including those who are not in the incident line/area). To address the aforementioned first challenge, we propose a simulation-based linearization to convert the total system travel time to a linear function of path flows using a first-order approximation, which leads to a tractable optimization problem. For the second challenge, this study focuses on the demand uncertainty (i.e., how many passengers will use the PT system during a service disruption) and models it within the robust optimization (RO) framework. The proposed approach is applied in a case study using data from the Chicago Transit Authority (CTA) system during a real-world urban rail disruption.

The main contributions of this chapter are as follows:

- To tackle the non-analytical system travel time calculation, we propose a simulation-based linearization to convert the total system travel time to a linear function of path flows using first-order approximation. Importantly, we utilize the physical interaction between passengers and vehicles in a public transit system to efficiently calculate the gradient (i.e., marginal change of travel time) without running the simulation multiple times (as opposed to traditional black-box optimization).

- We use RO to model the demand uncertainty which protects the model against inaccurate demand estimation. Specifically, we derive the closed-form robust counterpart with respect to the intersection of one ellipsoidal and three polyhedral uncertainty sets. These uncertainties capture the demand variations and the potential demand reduction during an incident. We also provide a feasible way of combining historical and survey data to quantify the uncertainty parameters.

The remainder of this chapter is organized as follows. The literature review is presented in Section 5.2. In Section 5.3, we describe the problem and discuss the solution methods. Section 5.4 discusses model extensions and generalizability. We apply the proposed framework on the CTA system as a case study in Section 5.5. The model results are analyzed in Section 5.6. Finally, we conclude the chapter and summarize the main findings in Section 5.7.

5.2 Literature review

5.2.1 Path recommendations during incidents

Most previous studies on path recommendations under incidents were conducted at a single OD level. That is, the main objective is to find available routes or the shortest path given an OD pair when the network is interrupted by incidents. For example, Bruglieri et al. [153] designed a trip planner to find the fastest path in the public transit network during service disruptions based on real-time mobility information. Böhmová et al. [154] developed a routing algorithm in urban public transportation to find reliable journeys that are robust against system delays. Roelofsen et al. [155] provided a framework for generating and assessing alternative routes in case of disruptions in urban public transport systems. To the best of the authors' knowledge, none of the previous studies have considered path recommendations at the system level, that is, providing path recommendations for passengers of different OD pairs and with different departure times so that the system travel time is reduced.

5.2.2 Passenger evacuation under emergencies

Providing path recommendations during disruptions is related to the topic of passenger evacuation under emergencies. The objective of evacuation is usually to minimize the total evacuation time. In general, these papers can be categorized into micro-level and macro-level based on how passenger flows are modeled and the spatial scope of the study area.

The micro-level studies usually use an agent-based simulation model to evaluate different evacuation strategies within some infrastructure. For example, Wang et al. [156] simulated the passenger evacuation under a fire emergency in Metro stations. Chen et al. [157] developed four modeling approaches including a queuing model and an agent-based simulation to calculate the evacuation time under different emergency situations and evacuation plans. Hassannayebi et al. [158] used an agent-based and discrete-event simulation model to assess the service level performance and crowdedness in a metro station under various disruption scenarios (e.g., train failure in the tunnel and fire at the station gallery). Zhou et al. [16] proposed a hybrid bi-level model to optimize the number and initial locations of leaders who guide passenger's evacuation in urban rail transit stations during an evacuation.

The macro-level studies consider a larger study area (e.g., city-level) and aim to evacuate passengers from the incident area through various transportation modes. For example, Abdelgawad and Abdulhai [159] developed an evacuation model to determine the routing and scheduling of subway and bus transit services used to alleviate congestion pressure during the evacuation of busy urban areas. Wang et al. [160] proposed an optimal bus bridging design method under operational disruptions on a single metro line. Tan et al. [161] proposes an evacuation model with urban bus networks as alternatives in the case of common metro service disruptions by jointly designing the bus lines and frequencies.

The macro-level passenger evacuation is similar to the setup of this study, but with the following major differences. First, in our study, the service disruption is not as severe as an emergency situation. The service will recover after a period of time and passengers are allowed to wait at a station. They do not necessarily need to cancel trips or follow evacuation plans as required in evacuation studies. Second, in this study, we assume that the service adjustment is known. The focus is on providing information to the passengers to better utilize the existing resources/capacities of the system.

5.2.3 Travel time calculation in public transit networks

Passengers' travel time has two components: in-vehicle time and waiting time. In-vehicle time is not affected by passenger flows once passengers are onboard, thus is easy to model (e.g., modeled as a constant). However, the waiting time is more complicated to calculate if the system is congested with left behind due to capacity constraints.

Passengers' travel time is usually modeled in the context of transit assignment, using two major approaches: frequency-based (static) and schedule-based (dynamic).

In the frequency-based transit assignment approach, the waiting time is either assumed to be inversely proportional to the (effective) service frequency [162, 163, 164], or modeled as a congestion function (e.g., BRP) of previously boarded flows and new arrival flows with exogenously-calibrated parameters [152]. The former method does not consider the left behind, and the latter only outputs a generalized waiting cost (rather than the waiting time as the vehicle capacity is not explicitly modeled) and requires a dedicated calibration process. Therefore, the frequency-based transit assignment model is not suitable for this study because congestion and left behind are not negligible during disruptions.

In terms of the schedule-based models [165, 166, 167, 168], the waiting time can only be obtained after a dynamic network loading (or simulation) process. For example, Schmöcker et al. [168] used the fail-to-board probability to model the left behind. This probability is updated after each network loading and can be used to calculate the waiting time. However, in this way, the waiting time is still a constant within each iteration. There is no direct way to formulate waiting time as a function of path flows.

Since formulating travel time as a function of path flows remains a challenge, the optimal passenger flow distribution in transit networks has no closed-form formulation. This study proposes a simulation-based first-order approximation to solve the original problem iteratively. With the proposed tractable linear programming model, the uncertainties can also be incorporated.

5.2.4 Robust optimization (RO)

RO is a common approach to handle data uncertainty in optimization problems. RO generally needs to first specify a scope some uncertain parameters. The scope is referred to as the "uncertainty set". The optimization problem is conducted over the worst-case realizations within the specified uncertainty set. This method is suitable for applications where there are uncertainty related to the model input parameters and when uncertainties can lead to significant penalties or infeasibility in practice. Since the solutions are optimal under the worst-case scenario, we treat the outputs of RO as a robust solution.

The solution method for RO problems involves generating a deterministic equivalent formulation, called the robust counterpart. Computational tractability of the robust counterpart has been a major practical difficulty [169]. A variety of uncertainty sets have been identified for which the robust counterpart is reasonably tractable [170].

The studies on RO has grown substantially over the past decades. Seminal papers include [171], [172] and [173]. Comprehensive surveys on the early literature can be found in Ben-Tal et al. [169] and Bertsimas et al. [170]. The development of the RO methodology has allowed researchers to tackle problems with data uncertainty in a range of fields. Examples include renewable energy network design [174], supply chain operations [175], health care logistics [176], and ride-hailing [177].

However, to the best of the authors' knowledge, no existing papers have incorporated RO techniques into path recommendations during service disruptions. This research gap is important to address given the potentially inaccurate estimates of demand in public transit networks during an incident.

5.3 Methodology

5.3.1 Problem description

Consider a service disruption in an urban rail system starting at time T_s and ending at T_e. During the disruption, some stations in the incident line (or the whole line) are blocked. Passengers in the blocked trains are usually offloaded to the nearest platforms. To respond to the incident, some changes in the operations are made, such as dispatching shuttle buses, rerouting existing services, short-turning in the incident line, headway adjustment, etc. Assume that we have all information about the operating changes. These changes define a new PT service network and alternative path sets. Our objective is to design an origin-destination (OD) based recommendation system. That is, when the incident happens, passengers can use their phones, websites, or electrical boards at stations to access the recommendation system. They input their **origin station, destination station, and departure time** to get a recommended path. The recommendation aims to minimize the system travel time, that is, the sum of all passengers' travel times, including passengers at nearby lines or bus routes without incidents (note that these passengers may experience additional crowding due to transfer passengers from the incident line).

Let $\mathcal{K}$ be the predetermined set of all OD pairs that may need path recommendation. $\mathcal{K}$ is defined based on whether an OD are affected by the incident or not. Note that as path recommendations start at T_s, the origin for passengers who are already in the system (e.g., offloaded passengers from the blocked vehicles) is their current location (as opposed to their initial origin such as the boarding station). We aim to provide recommendations for passengers whose OD pair is in $\mathcal{K}$ and departure time in the range from T_s to some time after T_e, since the congestion may last longer than T_e and passengers departing after T_e may also need guidance. The period of recommendation starts at time point (h_0) and consists of time intervals ($h_1, ..., h_H$) of equal length τ (e.g., 10 minutes). Specifically, h_0 represent the time point at T_s. Recommendations at T_s focus on passengers who are already in the system (and their departure times are T_s). And h_t ($t \geq 1$) represents the time interval $(T_s + (t-1)\tau, T_s + t\tau]$. Recommendations at h_t ($t \geq 1$) focus on passengers who were not in the system when the incident happened and their departure times are in $(T_s + (t-1)\tau, T_s + t\tau]$. Let the set of all recommendation times be $\mathcal{H} := \{h_0, h_1, ..., h_H\}$.

Given the operations during the incident, we obtain a feasible path set R_k for each OD pair k. Note that R_k includes all feasible services that are provided by the PT operator. A path $r \in R_k$ may be waiting for the system to recover (i.e., using the incident line), or transfer to nearby bus lines, using shuttle services, etc. We do not consider non-PT modes, such as Uber or driving for the following reasons: 1) The study aims to design a path recommendation system used by PT operators to provide path alternative recommendations to all PT users. Considering non-PT modes needs the supply information of all other travel modes and even consider non-PT users (such as the impact of traffic congestion on drivers), which is beyond the scope of this study. Future research may consider a multi-modal path recommendation system. 2)

Passengers using non-PT modes can be simply treated as demand reduction for the PT system. So their impact on the PT system is still captured.

Let d_{hk} be the number of passengers using the PT system with OD pair $k \in \mathcal{K}$ and departure time $h \in \mathcal{H}$. It can be treated as the normal demand minus the number of passengers leaving the PT system. As we do not have full information about future demand and number of passengers leaving the system, d_{hk} is an uncertainty variable which will be discussed in Section 5.3.4. Let f_{hkr} be the number of passengers departing at time interval h using OD pair k and path $r \in R_k$. By definition:

$$\sum_{r \in R_k} f_{hkr} = d_{hk} \quad \forall h \in \mathcal{H}, k \in \mathcal{K} \tag{5.1}$$

Let p_{hkr} be the corresponding path share of f_{hkr} (i.e., $p_{hkr} = f_{hkr}/d_{hk}$ and $\sum_{r \in R_k} p_{hkr} = 1$). For convenience of description, we define $\mathcal{F} := \{(h,k,r) : \forall h \in \mathcal{H}, \forall k \in \mathcal{K}, r \in R_k\}$ as the set of all path indices. Then the optimal flow problem can be formulated as:

$$\begin{aligned}
\min_{\boldsymbol{f},\boldsymbol{p}} \quad & Z(\boldsymbol{f}) = \text{Sum of all passengers' travel time} && \text{(5.2a)}\\
\text{s.t.} \quad & \sum_{r \in R_k} p_{hkr} = 1 \qquad \forall\, h \in \mathcal{H}, k \in \mathcal{K}, && \text{(5.2b)}\\
& f_{hkr} = d_{hk} \cdot p_{hkr} \quad \forall\, (h,k,r) \in \mathcal{F}, && \text{(5.2c)}\\
& f_{hkr} \geq 0 \qquad \forall\, (h,k,r) \in \mathcal{F}, && \text{(5.2d)}\\
& 0 \leq p_{hkr} \leq 1 \qquad \forall\, (h,k,r) \in \mathcal{F} && \text{(5.2e)}
\end{aligned}$$

where $\boldsymbol{f} := (f_{hkr})_{h,k,r \in \mathcal{F}}$ and $\boldsymbol{p} := (p_{hkr})_{h,k,r \in \mathcal{F}}$. $Z(\boldsymbol{f})$ is the system travel time which has no analytical expression. It can only be obtained after each network loading or simulation process (see Section 5.2.3). Note that using both $\boldsymbol{f}$ and $\boldsymbol{p}$ in the optimization problem is redundant, but it is useful for explaining the methodology.

If there is no uncertainty in the system, the optimal path shares (p^*_{hkr}) obtained from the solution of Eq. 5.2 are the recommendation proportions. That is, for all passengers with OD pair k and departure time h, the system will recommend them to use path r with probability p^*_{hkr}. However, Eq. 5.2 is a conceptual formulation, it cannot be solved directly because $Z(\boldsymbol{f})$ has no analytical expression. Moreover, given the uncertainties in demand, the final recommended path shares may not be p^*_{hkr}. In the following section, we elaborate on how to solve the robust "optimal flow problem" with demand uncertainties.

5.3.2 Event-based public transit simulator

Simulator design

Before introducing the solution procedure for Eq. 5.2, we first describe an event-based public transit simulator used in this study [4]. It can be used to evaluate $Z(\boldsymbol{f})$ and facilitate simulation-based linearization.

Figure 5-1 summarizes the main structure of the simulator. The inputs for the simulator are time-dependent OD demand (or smart card data), path shares, network structure, and train movement data (or timetable). Three objects are defined: trains, queues, and passengers. Trains are characterized by routes, train ID, current locations, and capacities. Passengers are queued based on their arrival times. Three different types of passengers are represented: left-behind passengers who were denied boarding from previous trains, new tap-in passengers from outside the system, and new transfer passengers from other lines. The left-behind passengers are usually at the head of the queue.

An event-based modeling framework is used to load the passengers onto the network. Two types of events are considered: train arrivals and train departures. The events are sorted by time and processed sequentially until all events are successfully completed during the analysis period. Train event lists (arrivals and departures) are generated according to the actual train movement data or timetable. Each event contains a train ID, occurrence time, and location (platform). Passengers are assigned to a path based on the corresponding input path shares. Note that in this study, a "path" is defined with specific boarding and transfer stations and lines. We assume passengers following a path will only board vehicles belonging to the specific line, even though there are multiple lines that serve a trip segment. Hence, there is no "common line" problem [152] in this study because "common lines" will be treated as different paths.

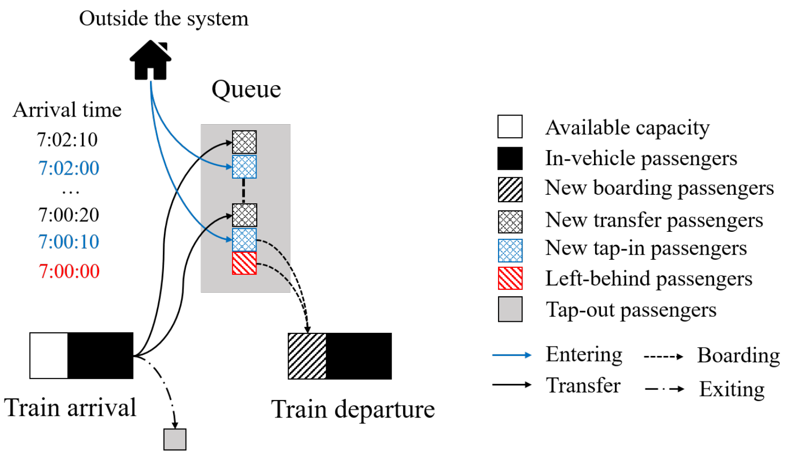


Figure 5-1: Structure of the network loading model (adapted from Mo et al. [4])

For an arrival event, the train offloads passengers who reach their destination or need to transfer at the station and updates its state (e.g. train load and in-vehicle passengers). For passengers who reach their destinations, their tap-out times are calculated by adding their egress time. For those who transfer at the station, their arrival times at the next platform are calculated based on the transfer time. The transfer passengers are added to the waiting queue in order of their arrival times at the next platform.

For departure events, the queue on the platform is updated by the new tap-in passengers, that is, passengers who arrive at the platform after the last train departed are added into the queue based on their arrival times. Passengers board the train according to a First-Come-First-Serve (FCFS) discipline until the train reaches its capacity. Passengers who cannot board are left behind and wait in the queue for the next train. The states of the train and the waiting queue are updated accordingly.

The simulator can record every passenger's trajectory during the whole travel process, including tap-in time, platform arrival time, boarding time, alighting time, tap-out time, etc. This information can be used for the simulation-based linearization of the objective function $Z(\boldsymbol{f})$.

Simulating service disruptions

Given a service disruption, the event list is modified to incorporate the incident's impact on the supply side. Specifically, all incidents' impacts can be reflected by changes in vehicles' arrival and departure times. For example, the blockage of a rail line can be represented with some vehicles in the line having long dwell times at the corresponding stations during the incident period. The dispatching of shuttle buses can be seen as adding a new set of events (vehicle arrivals and departures) associated with the new bridging route. The headway adjustment of existing routes can also be captured by the new vehicle arrival and departure times. In this way, the event-based simulator can conveniently model service disruptions without changing the framework.

From the passenger side, when an incident happens, all passengers in blocked trains are offloaded to the nearest platform. Depending on the input path choices (i.e., recommendation strategies) $\boldsymbol{p}$, offloading passengers are re-assigned to a new alternative path and join the queues at the corresponding boarding station. After reassigning the offloading passengers, the simulator continues to run from the incident time to the end of the simulation period (note that passengers who have not entered the system when the incident occurs will have a new path choice depending on the input $\boldsymbol{p}$).

5.3.3 Simulation-based linearization of the objective function

In this section, we propose a simulation-based linearization for the non-analytical $Z(\boldsymbol{f})$ based on a first-order approximation. $Z(\boldsymbol{f})$ can be approximated as:

$$\hat{Z}(\boldsymbol{f}) = Z(\tilde{\boldsymbol{f}}) + (\boldsymbol{f} - \tilde{\boldsymbol{f}})^T \frac{\partial Z(\boldsymbol{f})}{\partial \boldsymbol{f}}|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}} \tag{5.3}$$

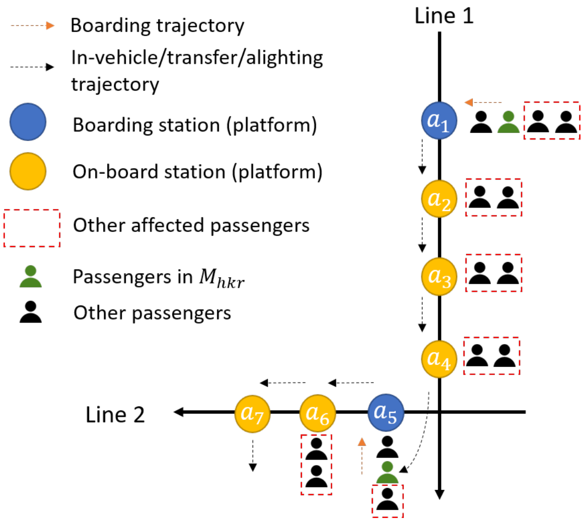
where $\hat{Z}(\boldsymbol{f})$ is the first-order approximation of $Z(\boldsymbol{f})$. $\tilde{\boldsymbol{f}}$ is a reference flow for the first-order approximation. $Z(\tilde{\boldsymbol{f}})$ is the system travel time estimated by simulation with $\tilde{\boldsymbol{f}}$ as input. $\frac{\partial Z(\boldsymbol{f})}{\partial \boldsymbol{f}} = (\frac{\partial Z(\boldsymbol{f})}{\partial f_{hkr}})_{h,k,r\in\mathcal{F}}$ is the gradient vector of $Z(\boldsymbol{f})$. As $\tilde{\boldsymbol{f}}$ and

$Z(\tilde{\boldsymbol{f}})$ are pre-determined, the only unknown part is $\frac{\partial Z(\boldsymbol{f})}{\partial \boldsymbol{f}}|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}}$. Notice that $\frac{\partial Z(\boldsymbol{f})}{\partial f_{hkr}}|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}}$ represents the change of system travel time caused by one unit of flow change in f_{hkr}. It can be approximated as:

$$\frac{\partial Z(\boldsymbol{f})}{\partial f_{hkr}}|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}} \approx \frac{Z(\tilde{\boldsymbol{f}} + \boldsymbol{e}_{hkr}) - Z(\tilde{\boldsymbol{f}})}{1} \tag{5.4}$$

where $\boldsymbol{e}_{hkr}$ represents a vector with only the (h, k, r)-th element being 1 and others zero. Eq. 5.4 represents the numerical approximation of the gradient. Now we only need to calculate $Z(\tilde{\boldsymbol{f}} + \boldsymbol{e}_{hkr}) - Z(\tilde{\boldsymbol{f}})$. A naive method to do that is to run a simulation with $\tilde{\boldsymbol{f}} + \boldsymbol{e}_{hkr}$ as input. However, as running the simulation is time-consuming, this method is not efficient. Note that since we already run a simulation with $\tilde{\boldsymbol{f}}$ as input, it is possible to directly calculate the marginal change due to the additional unit of flow (i.e., calculate the additional travel time increase to the system if one additional flow is added to $\tilde{f}_{hkr}$).

Consider an example journey of $\tilde{f}_{hkr}$ in Figure 5-2. Let M_{hkr} be the set of passengers composing the flow of $\tilde{f}_{hkr}$ (i.e., the green passengers in Figure 5-2). These passengers have origin station a_1 and destination station a_7, and the path includes a transfer from station a_4 to station a_5. Let the average travel time of $\tilde{f}_{hkr}$ be $T^{\text{A}}_{hkr}(\tilde{\boldsymbol{f}})$. Suppose that one more passenger is added to $\tilde{f}_{hkr}$.



Boarding trajectory
In-vehicle/transfer/alighting trajectory
Boarding station (platform)
On-board station (platform)
Other affected passengers
Passengers in M_{hkr}
Other passengers
Line 1
Line 2
a_1
a_2
a_3
a_4
a_5
a_6
a_7

Figure 5-2: Explanation for the impact of adding additional one unit flow to the system

First of all, the system travel time is increased by $T^{\mathrm{A}}_{hkr}(\tilde{\boldsymbol{f}})$ due to the increase in the flow amount. Note that considering the marginal calculation, we ignore the impact of the added passenger on the increase in $T^{\mathrm{A}}_{hkr}(\tilde{\boldsymbol{f}})$. Besides, all passengers in the red-dashed square may experience higher travel times. Passengers at station a_1 and a_5 who queue behind the green passenger may have additional waiting time if the train that M_{hkr} used is full after departure (under the simulation results of $\tilde{\boldsymbol{f}}$), because the increase of the flow by one in $\tilde{f}_{hkr}$ will occupy one available capacity for these waiting passengers, and one of them will have to board the next train (i.e., wait for one more headway). Mathematically, let V^b_{hkr} be the set of vehicles that the M_{hkr} passengers board at station b. Adding an additional passenger to M_{hkr} means one more passenger board one of the vehicles in V^b_{hkr}. Let $\mathbb{1}_{\{\mathrm{Full}^b_v\}}$ be an indicator of whether vehicle v is full or not after its departure from station b. Then the total increase in system travel time for passengers queuing behind M_{hkr} is:

$$T^{\mathrm{Q}}_{hkr}(\tilde{\boldsymbol{f}}) = \sum_{b \in B_{hkr}} \sum_{v \in V^b_{hkr}} \frac{\mathbb{1}_{\{\mathrm{Full}^b_v\}} \cdot W^b_v}{|V^b_{hkr}|} \tag{5.5}$$

where B_{hkr} is the set of all boarding stations for M_{hkr} passengers (in this example, a_1 and a_5). W^b_v is the headway of vehicle v at station b. The sum over all vehicles is because we do not specify the exact vehicle that the additional passenger will board, and thus take the average over all vehicles. In this example, since there are two boarding stations for M_{hkr} (a_1, a_5), $T^{\mathrm{Q}}_{hkr}(\tilde{\boldsymbol{f}})$ is approximately two headways if the vehicles are full.

For passengers waiting at stations where M_{hkr} are already on-board (referred to as on-board stations, e.g., station a_2), adding one flow to $\tilde{f}_{hkr}$ reduces the available capacity when the vehicle arrives at these on-board stations. The queuing passengers at the on-board stations may not be able to board due to the reduction of capacity. Specifically, if a vehicle is full when it departs from an onboard station under flow pattern $\tilde{\boldsymbol{f}}$, adding one passenger to $\tilde{f}_{hkr}$ makes one passenger waiting at the on-board station unable to board his/her original boarded vehicle. And the system travel time is increased by one headway for each of these onboard stations. Mathematically, let

O^v_{hkr} be the set of all on-board stations for M_{hkr} and vehicle $v \in V^b_{hkr}$. For example, for vehicles in Line 1, O^v_{hkr} will be a_2, a_3, and a_4. Then the travel time increase for passengers waiting at on-board stations is:

$$T^{\mathrm{O}}_{hkr}(\tilde{\boldsymbol{f}}) = \sum_{b \in B_{hkr}} \sum_{v \in V^b_{hkr}} \frac{1}{|V^b_{hkr}|} \sum_{a \in O^v_{hkr}} \mathbb{1}_{\{\mathrm{Full}^a_v\}} \cdot W^a_v \tag{5.6}$$

Therefore, in this way, depending on whether the vehicle is full or not under flow pattern $\tilde{\boldsymbol{f}}$, the increase in system travel time due to adding one passenger to $\tilde{f}_{hkr}$ can be calculated without running the simulation again. These increases come from three parts: 1) the average travel time of M_{hkr} due to increasing in flow amount (i.e., $T^{\mathrm{A}}_{hkr}(\tilde{\boldsymbol{f}})$), 2) the additional waiting time for passengers queuing behind M_{hkr} (i.e., $T^{\mathrm{Q}}_{hkr}(\tilde{\boldsymbol{f}})$), and 3) the additional waiting time for passengers queuing at M_{hkr}'s on-board stations (i.e., $T^{\mathrm{O}}_{hkr}(\tilde{\boldsymbol{f}})$). Specifically, we have

$$Z(\tilde{\boldsymbol{f}} + \boldsymbol{e}_{hkr}) - Z(\tilde{\boldsymbol{f}}) = T^{\mathrm{A}}_{hkr}(\tilde{\boldsymbol{f}}) + T^{\mathrm{Q}}_{hkr}(\tilde{\boldsymbol{f}}) + T^{\mathrm{O}}_{hkr}(\tilde{\boldsymbol{f}}) \tag{5.7}$$

Consequently, $\frac{\partial Z(\boldsymbol{f})}{\partial \boldsymbol{f}}|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}}$ can be obtained from Eq. 5.4. Define $\boldsymbol{\beta}(\tilde{\boldsymbol{f}}) := \frac{\partial Z(\boldsymbol{f})}{\partial \boldsymbol{f}}|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}}$. Then the objective function becomes:

$$\hat{Z}(\boldsymbol{f}) = Z(\tilde{\boldsymbol{f}}) + \boldsymbol{\beta}(\tilde{\boldsymbol{f}})^T(\boldsymbol{f} - \tilde{\boldsymbol{f}}) \tag{5.8}$$

where $\boldsymbol{\beta}(\tilde{\boldsymbol{f}}) = (\beta_{hkr})_{h,k,r \in \mathcal{F}}$ and $\beta_{hkr} = \frac{\partial Z(\boldsymbol{f})}{\partial f_{hkr}}|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}}$. Eq. 5.8 is a linear function of $\boldsymbol{f}$, which supports for addressing uncertainties in the optimization problem.

5.3.4 Demand uncertainty

The uncertainty of d_{hk} comes from two different parts. The first is the inherent demand variations across different days, and the second is the uncertainty in how many passengers leave the PT system during the incident. In this section, these two uncertainties are considered as a whole by introducing an ellipsoidal uncertainty set and three polyhedral uncertainty sets.

From constraint 5.2c, we can substitute $f_{hkr} = d_{hk} \cdot p_{hkr}$ to the objective function and rewrite Eq. 5.8 as:

$$\hat{Z}(\boldsymbol{f}) = \hat{Z}(\boldsymbol{p}) = Z(\tilde{\boldsymbol{f}}) + \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr} \cdot (d_{hk} \cdot p_{hkr} - \tilde{f}_{hkr}) \tag{5.9}$$

Note that β_{hkr} is a function of $\tilde{\boldsymbol{f}}$, for simplicity we ignore $\tilde{\boldsymbol{f}}$ in the derivation process.

To model the uncertainty of d_{hk}, we introduce an auxiliary decision variable t and rewrite the optimal flow problem as:

$$\begin{aligned} \min_{\boldsymbol{p},t} \quad & t && \text{(5.10a)} \\ \text{s.t.} \quad & t \geq Z(\tilde{\boldsymbol{f}}) + \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr} \cdot (d_{hk} \cdot p_{hkr} - \tilde{f}_{hkr}), && \text{(5.10b)} \\ & \text{Constraints (5.2b) and (5.2e)} && \text{(5.10c)} \end{aligned}$$

Constraint 5.10b can be rewritten as

$$\sum_{h,k} \sum_{r\in R_k} \beta_{hkr} \cdot d_{hk} \cdot p_{hkr} \leq t - Z(\tilde{\boldsymbol{f}}) + \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr} \tilde{f}_{hkr} \tag{5.11}$$

Eq. 5.11 can be written in a matrix form as:

$$\boldsymbol{a}^T \boldsymbol{p} \leq b \tag{5.12}$$

where $\boldsymbol{a} \in \mathbb{R}^{|\mathcal{F}|}$ with the entry $a_{hkr} = \beta_{hkr} d_{hk}$, $\forall\ (h,k,r) \in \mathcal{F}$. And $b = t - Z(\tilde{\boldsymbol{f}}) + \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr} \tilde{f}_{hkr}$. Define $\boldsymbol{d} = (d_{hk})_{h\in\mathcal{H},k\in\mathcal{K}}$.

Proposition 13. *If $\boldsymbol{d}$ is normally distributed with $\boldsymbol{d} \sim \mathcal{N}(\bar{\boldsymbol{d}}, \boldsymbol{\Sigma})$, then in a RO problem where constraint 5.12 is guaranteed to be satisfied with probability of at least $1 - \varepsilon$ (i.e., $\mathbb{P}[\boldsymbol{a}^T \boldsymbol{p} \leq b] \geq 1 - \varepsilon$), the robust constraint can be formulated as:*

$$(\boldsymbol{A}\bar{\boldsymbol{d}} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{z})^T \boldsymbol{p} \leq b, \quad \forall \boldsymbol{z} \in \mathcal{Z}_E \tag{5.13}$$

where $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{F}|\times HK}$ with entry $A_{hkr,h'k'} = \beta_{hkr}$ if $h = h'$ and $k = k'$, otherwise

$A_{hkr,h'k'} = 0$. $\boldsymbol{D}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}$ (i.e., $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{D}^T$). $\boldsymbol{z}$ are the perturbation variables (i.e., $\boldsymbol{d} = \bar{\boldsymbol{d}} + \boldsymbol{D}\boldsymbol{z}$) and $\mathcal{Z}_E = \left\{\boldsymbol{z} \in \mathbb{R}^{HK} : \|z\|_2 \leq \rho_{1-\varepsilon}\right\}$ (i.e., the ellipsoidal uncertainty set). $\rho_{1-\varepsilon}$ is the $(1-\varepsilon)$-percentile of a standard normal distribution.

Proof.

Step 1: We first prove that $\mathbb{P}[\boldsymbol{a}^T\boldsymbol{p} \leq b] \geq 1-\varepsilon$ is equivalent to $(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \rho_{1-\varepsilon}\left\|(\boldsymbol{A}\boldsymbol{D})^T\boldsymbol{p}\right\|_2 \leq b$.

Since $\boldsymbol{d}$ is normally distributed, we have $\boldsymbol{a} = \boldsymbol{A}\boldsymbol{d}$ is normally distributed with $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{A}\bar{\boldsymbol{d}}, \boldsymbol{A}\Sigma\boldsymbol{A}^T)$. Similarly, $\boldsymbol{a}^T\boldsymbol{p} \in \mathbb{R}$ is also normally distributed with

$$\boldsymbol{a}^T\boldsymbol{p} \sim \mathcal{N}((\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p}, \boldsymbol{p}^T\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{p}) \tag{5.14}$$

If we want constraint 5.12 to hold with probability at least $1-\varepsilon$, it suffices to have:

$$(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \rho_{1-\varepsilon}\sqrt{\boldsymbol{p}^T\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{p}} \leq b \tag{5.15}$$

Substituting $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{D}^T$ into Eq. 5.15 completes the proof of Step 1.

Step 2: We need to show that the robust counterpart of Eq. 5.13 is $(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \rho_{1-\varepsilon}\left\|(\boldsymbol{A}\boldsymbol{D})^T\boldsymbol{p}\right\|_2 \leq b$.

Eq. 5.13 is equivalent to:

$$(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \max_{\boldsymbol{z}\in\mathcal{Z}_{\mathrm{E}}}(\boldsymbol{A}\boldsymbol{D}\boldsymbol{z})^T\boldsymbol{p} \leq b. \tag{5.16}$$

Let $\delta(\boldsymbol{z} \mid \mathcal{Z}_{\mathrm{E}})$ be the indicator function on set $\mathcal{Z}_{\mathrm{E}}$:

$$\delta(\boldsymbol{z} \mid \mathcal{Z}_{\mathrm{E}}) = \begin{cases} 1, & \text{if } \boldsymbol{z} \in \mathcal{Z}_{\mathrm{E}} \\ 0, & \text{otherwise} \end{cases} \tag{5.17}$$

Then the convex conjugate of $\delta(\boldsymbol{z} \mid \mathcal{Z}_{\mathrm{E}})$ (also known as the **support function**) can be derived as [178]:

$$\delta^*(\boldsymbol{y} \mid \mathcal{Z}_{\mathrm{E}}) = \sup_{\boldsymbol{z}\in\mathbb{R}^{HK}} \{\boldsymbol{y}^T\boldsymbol{z} - \delta(\boldsymbol{z} \mid \mathcal{Z}_{\mathrm{E}})\} = \sup_{\boldsymbol{z}\in\mathcal{Z}_{\mathrm{E}}} \boldsymbol{y}^T\boldsymbol{z} = \rho_{1-\varepsilon} \left\|\boldsymbol{y}\right\|_2 \tag{5.18}$$

Therefore, Eq. 5.16 can be rewritten with the convex conjugate:

$$(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \delta^*((\boldsymbol{A}\boldsymbol{D})^T\boldsymbol{p} \mid \mathcal{Z}) = (\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \rho_{1-\varepsilon} \left\|(\boldsymbol{A}\boldsymbol{D})^T\boldsymbol{p}\right\|_2 \leq b \tag{5.19}$$

which finishes the proof of Step 2. Combining Steps 1 and 2 finishes the proof of the whole proposition. □

We observe that the ellipsoidal demand uncertainty performs like a regularization. It prevents $\boldsymbol{p}$ from being large in directions with considerable uncertainty in the demand.

Remark 6. In the RO, the ellipsoidal uncertainty set can be used no matter what distribution $\boldsymbol{d}$ follows. If $\boldsymbol{d}$ is normally distributed, the parameter $\rho_{1-\varepsilon}$ can be interpreted as the probability that constraint 5.12 holds. The use of the multivariate normality assumption in Proposition 13 is for explaining the physical meaning of ellipsoidal uncertainty set and facilitating the choice of hyperparameters (i.e., $\rho_{1-\varepsilon}$ and $\boldsymbol{D}$). Moreover, in the case study, we partially validate the multivariate normality assumption of $\boldsymbol{d}$ using smart card data. The Mardia's Skewness Test [179] shows that $\boldsymbol{d}$ has no significant skewness.

Eq. 5.13 (i.e., the ellipsoidal uncertainty set) captures the correlation between demands at different time intervals and OD pairs. However, it does not impose any upper or lower bounds on d_{hk}. In reality, the demand level for a specific OD pair and time interval is usually bounded, which can be expressed as:

$$d_{hk}^{\mathrm{L}} \leq d_{hk} \leq d_{hk}^{\mathrm{U}} \tag{5.20}$$

where d_{hk}^{L} and d_{hk}^{U} are the corresponding lower and upper bounds for d_{hk}, respectively. Their values can be obtained from historical demand data. Eq. 5.20 can be rewritten in a vector form as $\boldsymbol{d}^{\mathrm{L}} \leq \boldsymbol{d} \leq \boldsymbol{d}^{\mathrm{U}}$, where $\boldsymbol{d}^{\mathrm{U}} = (d_{hk}^{\mathrm{U}})_{h\in\mathcal{H},k\in\mathcal{K}}$ and $\boldsymbol{d}^{\mathrm{L}} = (d_{hk}^{\mathrm{L}})_{h\in\mathcal{H},k\in\mathcal{K}}$. Since we have $\boldsymbol{d} = \bar{\boldsymbol{d}} + \boldsymbol{D}\boldsymbol{z}$, a simple manipulation leads to

$$\boldsymbol{d}^{\mathrm{L}} - \bar{\boldsymbol{d}} \leq \boldsymbol{D}\boldsymbol{z} \leq \boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}} \tag{5.21}$$

We can rewrite it as a "polyhedral uncertainty set": $\mathcal{Z}_{\mathrm{P1}} = \left\{\boldsymbol{z} \in \mathbb{R}^{HK} : \boldsymbol{d}^{\mathrm{L}} - \bar{\boldsymbol{d}} \leq \boldsymbol{D}\boldsymbol{z} \leq \boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}}\right\}$.

Eq. 5.20 ensures the boundaries for each individual demand. Another similar constraint for the demand uncertainty is that: within a given time interval, the total demand across all OD pairs should also be bounded. This constraint can avoid some extreme scenarios that Eq. 5.20 cannot capture (e.g., all d_{hk} are at the lower or upper bounds). Mathematically:

$$d_h^{\mathrm{L}} \leq \sum_{k\in\mathcal{K}} d_{hk} \leq d_h^{\mathrm{U}} \tag{5.22}$$

where d_h^{L} and d_h^{U} are the lower and upper bounds for the total demand in time interval h, which can be obtained from the historical demand. Define $\boldsymbol{S} \in \mathbb{R}^{H\times HK}$, where the element $S_{h,h'k} = 1$ if $h = h'$, otherwise $S_{h,h'k} = 0$. Then Eq. 5.22 can be rewritten in a matrix form:

$$\boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}} - \boldsymbol{S}\bar{\boldsymbol{d}} \leq \boldsymbol{S}\boldsymbol{D}\boldsymbol{z} \leq \boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}} - \boldsymbol{S}\bar{\boldsymbol{d}} \tag{5.23}$$

where $\boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}} = (d_h^{\mathrm{U}})_{h\in\mathcal{H}}$ and $\boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}} = (d_h^{\mathrm{L}})_{h\in\mathcal{H}}$. And Eq. 5.23 can also be represented as a polyhedral uncertainty set: $\mathcal{Z}_{\mathrm{P2}} = \left\{\boldsymbol{z} \in \mathbb{R}^{HK} : \boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}} - \boldsymbol{S}\bar{\boldsymbol{d}} \leq \boldsymbol{S}\boldsymbol{D}\boldsymbol{z} \leq \boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}} - \boldsymbol{S}\bar{\boldsymbol{d}}\right\}$.

As the RO aims to optimize under the "worst case" scenario and our objective function is the system travel time, intuitively, the worst-case scenario will be the largest demand in the uncertainty set. This may make the worst-case demand unrealistic since the extremely large demand seldom happens. What we expect in the RO is that the model can capture some critical OD pairs where the high demand in these OD pairs can make the system more congested (as opposed to high demand in all OD pairs). In order to let the RO capture critical OD pairs, we add an additional constraint on the total demand:

$$\sum_{h\in\mathcal{H},k\in\mathcal{K}} d_{hk} \leq \Gamma \cdot \sum_{h\in\mathcal{H},k\in\mathcal{K}} \bar{d}_{hk} \tag{5.24}$$

where $\Gamma > 0$ is a predetermined constant. $\Gamma = 1$ means we assume the total demand in the worst case scenario is the same as the nominal one, but the spatial and temporal distributions are different. The worst case scenario will have more demand on critical OD pairs but less demand on others. The value of Γ can be determined based on the highest total demand observed over a time period.

Similarly, Eq. 5.24 can be written in a matrix form:

$$\mathbf{1}^T(\bar{\boldsymbol{d}} + \boldsymbol{D}\boldsymbol{z}) \leq \Gamma \cdot \mathbf{1}^T\bar{\boldsymbol{d}} \tag{5.25}$$

where $\mathbf{1} \in \mathbb{R}^{HK}$ is a vector with all elements one. And we define another polyhedral uncertainty set: $\mathcal{Z}_{\text{P3}} = \big\{\boldsymbol{z} \in \mathbb{R}^{HK} : \mathbf{1}^T(\bar{\boldsymbol{d}} + \boldsymbol{D}\boldsymbol{z}) \leq \Gamma \cdot \mathbf{1}^T\bar{\boldsymbol{d}}\big\}$.

Therefore, the final robust constraint for Eq. 5.12 is

$$(\boldsymbol{A}\bar{\boldsymbol{d}} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{z})^T\boldsymbol{p} \leq b, \quad \forall \boldsymbol{z} \in \mathcal{Z}_{\text{E}} \cap \mathcal{Z}_{\text{P}} \cap \mathcal{Z}_{\text{P2}} \cap \mathcal{Z}_{\text{P3}} \tag{5.26}$$

To derive the robust counterpart of the constraint, we first introduce the following lemma.

Lemma 2. *For a constraint $\bar{\boldsymbol{a}}^T\boldsymbol{x} + \delta^*(\boldsymbol{P}^T\boldsymbol{x} \mid \mathcal{Z}) \leq b$, let $\mathcal{Z}_1, ..., \mathcal{Z}_k$ be closed convex sets, such that $\bigcap_i ri(\mathcal{Z}_i) \neq \emptyset$*[1]*, and let $\mathcal{Z} = \cap_{i=1}^k \mathcal{Z}_i$. Then,*

$$\delta^*(\boldsymbol{y} \mid \mathcal{Z}) = \min_{\boldsymbol{y}_1,\dots,\boldsymbol{y}_k} \{\sum_{i=1}^k \delta^*(\boldsymbol{y}_i \mid \mathcal{Z}_i) \mid \sum_{i=1}^k \boldsymbol{y}_i = \boldsymbol{y}\},$$

and the constraint becomes

$$\begin{cases} \bar{\boldsymbol{a}}^T\boldsymbol{x} + \sum_{i=1}^k \delta^*(\boldsymbol{y}_i \mid \mathcal{Z}_i) \leq b \\ \sum_{i=1}^k \boldsymbol{y}_i = \boldsymbol{P}^T\boldsymbol{x} \end{cases}$$

[1] $ri(\mathcal{Z}_i)$ indicates the relative interior of the set $\mathcal{Z}_i$.

where $\delta^*(\cdot \mid \cdot)$ is the support function (i.e., convex conjugate of the indicator function).

The proof of Lemma 2 can be found in Ben-Tal et al. [180]. From Proposition 13, we have $\delta^*(\boldsymbol{y} \mid \mathcal{Z}_\text{E}) = \rho_{1-\varepsilon} \left\| \boldsymbol{y} \right\|_2$. For the polyhedral uncertainty set, consider a general form $\mathcal{Z}_\text{P} = \{\boldsymbol{z} : \boldsymbol{H}\boldsymbol{z} \leq \boldsymbol{c}\}$. And the support function for $\mathcal{Z}_\text{P}$ is

$$\delta^*(\boldsymbol{y} \mid \mathcal{Z}_\text{P}) = \max_{\boldsymbol{z}} \{\boldsymbol{y}^T\boldsymbol{z} \mid \boldsymbol{H}\boldsymbol{z} \leq \boldsymbol{c}\} = \min_{\boldsymbol{u}} \{\boldsymbol{c}^T\boldsymbol{u} \mid \boldsymbol{H}^T\boldsymbol{u} = \boldsymbol{y}, \boldsymbol{u} \geq 0\} \tag{5.27}$$

where the second equality follows from linear programming duality. Eq. 5.27 can be used to derive the support function for $\mathcal{Z}_\text{P1}$, $\mathcal{Z}_\text{P2}$, and $\mathcal{Z}_\text{P3}$. For example, consider the robust counterpart for Eq. 5.24, we have

$$\delta^*(\boldsymbol{y}_6 \mid \mathcal{Z}_\text{P3}) = \min_{u_3} \{(\Gamma - 1) \cdot (\boldsymbol{1}^T\bar{\boldsymbol{d}}) \cdot u_3 \mid (\boldsymbol{1}^T\boldsymbol{D})^T u_3 = \boldsymbol{y}_6, u_3 \geq 0\} \tag{5.28}$$

where $\boldsymbol{y}_6 \in \mathbb{R}^{HK}$ and $u_3 \in \mathbb{R}$ are decision variables in the RO model. Note that the subscripts for $\boldsymbol{y}$ and u (i.e., 6 and 3) are used for the consistency in Eq. 5.29.

Based on Lemma 2, the robust counterpart for Eq. 5.26 is

$$(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \rho_{1-\varepsilon} \left\| \boldsymbol{y}_1 \right\|_2 + (\boldsymbol{d}^\text{U} - \bar{\boldsymbol{d}})^T\boldsymbol{u}_1 + (\bar{\boldsymbol{d}} - \boldsymbol{d}^\text{L})^T\boldsymbol{u}_2 + (\boldsymbol{d}_\mathcal{H}^\text{U} - \boldsymbol{S}\bar{\boldsymbol{d}})^T\boldsymbol{v}_1 + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}_\mathcal{H}^\text{L})^T\boldsymbol{v}_2$$

$$+ (\Gamma - 1) \cdot (\boldsymbol{1}^T\bar{\boldsymbol{d}}) \cdot u_3 \leq b \tag{5.29a}$$

$$\boldsymbol{D}^T\boldsymbol{u}_1 = \boldsymbol{y}_2 \tag{5.29b}$$

$$-\boldsymbol{D}^T\boldsymbol{u}_2 = \boldsymbol{y}_3 \tag{5.29c}$$

$$(\boldsymbol{S}\boldsymbol{D})^T\boldsymbol{v}_1 = \boldsymbol{y}_4 \tag{5.29d}$$

$$-(\boldsymbol{S}\boldsymbol{D})^T\boldsymbol{v}_2 = \boldsymbol{y}_5 \tag{5.29e}$$

$$(\boldsymbol{1}^T\boldsymbol{D})^T u_3 = \boldsymbol{y}_6 \tag{5.29f}$$

$$\sum_{i=1}^{6} \boldsymbol{y}_i = (\boldsymbol{A}\boldsymbol{D})^T\boldsymbol{p} \tag{5.29g}$$

$$\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{v}_1, \boldsymbol{v}_2, u_3 \geq 0 \tag{5.29h}$$

Hence, the RO problem can be formulated as

$$\min_{\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y},t} \quad t \tag{5.30a}$$

$$\begin{aligned}\text{s.t.} \quad & \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr} \cdot d_{hk} \cdot p_{hkr} + \rho_{1-\varepsilon} \left\| \boldsymbol{y}_1 \right\|_2 + (\boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}})^T \boldsymbol{u}_1 + (\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}})^T \boldsymbol{u}_2 + (\boldsymbol{d}^{\mathrm{U}}_{\mathcal{H}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T \boldsymbol{v}_1 \\ & + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}}_{\mathcal{H}})^T \boldsymbol{v}_2 + (\Gamma - 1) \cdot (\mathbf{1}^T \bar{\boldsymbol{d}}) \cdot u_3 + Z(\tilde{\boldsymbol{f}}) - \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr} \tilde{f}_{hkr} \leq t\end{aligned} \tag{5.30b}$$

$$\text{Constraints } (5.29b) - (5.29h) \tag{5.30c}$$

$$\text{Constraints } (5.2b) \text{ and } (5.2e) \tag{5.30d}$$

By eliminating t and inserting constraint 5.30b in the objective function it becomes

$$\begin{aligned}\hat{Z}(\boldsymbol{p}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{y})^{\mathrm{RC}} = & \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr} \cdot (d_{hk} \cdot p_{hkr} - \tilde{f}_{hkr}) + \rho_{1-\varepsilon} \left\| \boldsymbol{y}_1 \right\|_2 + (\boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}})^T \boldsymbol{u}_1 + (\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}})^T \boldsymbol{u}_2 \\ & + (\boldsymbol{d}^{\mathrm{U}}_{\mathcal{H}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T \boldsymbol{v}_1 + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}}_{\mathcal{H}})^T \boldsymbol{v}_2 + (\Gamma - 1) \cdot (\mathbf{1}^T \bar{\boldsymbol{d}}) \cdot u_3 + Z(\tilde{\boldsymbol{f}})\end{aligned} \tag{5.31}$$

which yields a second-order cone programming (SOCP).

5.3.5 Solution procedure

After incorporating the demand uncertainty, the final robust counterpart (RC) of the optimal flow problem can be formulated as:

$$
\begin{aligned}
[RC(\tilde{\boldsymbol{f}})] \quad \min_{\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y}} \quad & \hat{Z}(\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y})^{\mathrm{RC}} = \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr}(\tilde{\boldsymbol{f}}) \cdot (d_{hk} \cdot p_{hkr} - \tilde{f}_{hkr}) + \rho_{1-\varepsilon} \left\| \boldsymbol{y}_1 \right\|_2 + (\boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}})^T \boldsymbol{u}_1 \\
& + (\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}})^T \boldsymbol{u}_2 + (\boldsymbol{d}^{\mathrm{U}}_{\mathcal{H}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T \boldsymbol{v}_1 + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}}_{\mathcal{H}})^T \boldsymbol{v}_2 + (\Gamma - 1) \cdot (\boldsymbol{1}^T \bar{\boldsymbol{d}}) \cdot u_3 + Z(\tilde{\boldsymbol{f}}) && (5.32a) \\
\text{s.t.} \quad & \text{Constraints } (5.29b) - (5.29h) && (5.32b) \\
& \sum_{r\in R_k} p_{hkr} = 1 \quad \forall h \in \mathcal{H}, k \in \mathcal{K} && (5.32c) \\
& 0 \leq p_{hkr} \leq 1 \quad \forall (h,k,r) \in \mathcal{F} && (5.32d)
\end{aligned}
$$

This SOCP can be efficiently solved by inner interior point methods that are embedded in many existing solvers.

However, due to the first-order approximation of $Z(\boldsymbol{f})$, $\beta_{hkr}(\tilde{\boldsymbol{f}})$ needs to be updated once a new flow pattern is obtained. Hence, after obtaining $\boldsymbol{p}^*$ from the RC problem, the simulation should be run again to update $\beta_{hkr}(\tilde{\boldsymbol{f}})$. Before that, the corresponding worst-case demand (WD), which will be used as the new $\tilde{\boldsymbol{f}}$, is needed. It can be obtained by solving the worst case $\boldsymbol{z} \in \mathcal{Z}_{\mathrm{E}} \cap \mathcal{Z}_{\mathrm{P1}} \cap \mathcal{Z}_{\mathrm{P2}} \cap \mathcal{Z}_{\mathrm{P3}}$:

$$
\begin{aligned}
[WD(\boldsymbol{p}^*)] \quad \max_{\boldsymbol{z}} \quad & (\boldsymbol{A}\boldsymbol{D}\boldsymbol{z})^T \boldsymbol{p}^* && (5.33a) \\
\text{s.t.} \quad & \left\| \boldsymbol{z} \right\|_2 \leq \rho_{1-\varepsilon} && (5.33b) \\
& \boldsymbol{d}^{\mathrm{L}} - \bar{\boldsymbol{d}} \leq \boldsymbol{D}\boldsymbol{z} \leq \boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}} && (5.33c) \\
& \boldsymbol{d}^{\mathrm{L}}_{\mathcal{H}} - \boldsymbol{S}\bar{\boldsymbol{d}} \leq \boldsymbol{S}\boldsymbol{D}\boldsymbol{z} \leq \boldsymbol{d}^{\mathrm{U}}_{\mathcal{H}} - \boldsymbol{S}\bar{\boldsymbol{d}} && (5.33d) \\
& \boldsymbol{1}^T (\bar{\boldsymbol{d}} + \boldsymbol{D}\boldsymbol{z}) \leq \Gamma \cdot \boldsymbol{1}^T \bar{\boldsymbol{d}} && (5.33e)
\end{aligned}
$$

If the solution for Eq. 5.33 is $\boldsymbol{z}^*$, the worse case demand is $\boldsymbol{d}^* = \bar{\boldsymbol{d}} + \boldsymbol{D}\boldsymbol{z}^*$. Next, we can update $\boldsymbol{\beta}(\tilde{\boldsymbol{f}})$ and $Z(\tilde{\boldsymbol{f}})$ as

$$Z(\tilde{\boldsymbol{f}}), \boldsymbol{\beta}(\tilde{\boldsymbol{f}}) = \textsc{Sim-FOA}(\boldsymbol{d}^*, \boldsymbol{p}^*) \tag{5.34}$$

where $\tilde{\boldsymbol{f}}$ in Eq. 5.34 indicates $\tilde{f}_{hkr} = d^*_{hk} \cdot p^*_{hkr}$. And $\textsc{Sim-FOA}(\cdot)$ is a pseudo function of simulation plus first-order approximation as described in Section 5.3.3.

The RC, WD, and $\textsc{Sim-FOA}(\cdot)$ problems need to be solved iteratively. This can be treated as a fixed-point problem. A conventional way to solve a fixed-point problem is the method of successive average (MSA). In the typical system optimal **traffic** assignment problem, the optimal flow pattern is reached when for every OD pair, the marginal costs of all paths for this OD pair are the same. This implies that, ideally, when the flow distribution is optimal, we should have $\beta_{hkr}(\tilde{\boldsymbol{f}}) = \beta_{hkr'}(\tilde{\boldsymbol{f}})$ for all $r, r' \in R_k \setminus R_k^{\text{NoFlow}}$, where $R_k^{\text{NoFlow}} = \{r \in R_k \mid f_{hkr} = 0\}$ is the path set with zero flows. This implies that at the system optimal assignment, the marginal cost (travel time) of every non-zero flow path is the same (i.e., one cannot decrease the system travel time by switching passengers from one path to another).

However, in our study, this cannot be set as the convergence criterion because, in the dynamic **transit** assignment context, the cost function is not continuous due to left behind. Adding one more passenger to a path may lead to the system travel time increased by one or more headways. The following example illustrates that $\beta_{hkr}(\tilde{\boldsymbol{f}})$ can be arbitrarily large, which may cause the criterion of $\beta_{hkr}(\tilde{\boldsymbol{f}}) = \beta_{hkr'}(\tilde{\boldsymbol{f}})$ never being satisfied.

Example 1. *Consider a single direction bus line with N stations (Figure 5-3) and a fixed headway W. Assume every bus has a capacity of 1. There is one passenger waiting at each station except for the first station (i.e., there are $N-1$ waiting passengers). Now assume that one more passenger is added to station 1. Since the capacity of buses is 1, the newly added passenger will force all waiting passengers to be left behind one more time. Hence, the total added system travel time is $(N-1) \times W$. In this scenario, the $\beta_{hkr}(\tilde{\boldsymbol{f}})$ associated with the added passenger can be arbitrarily large depending on the number of stations N.*

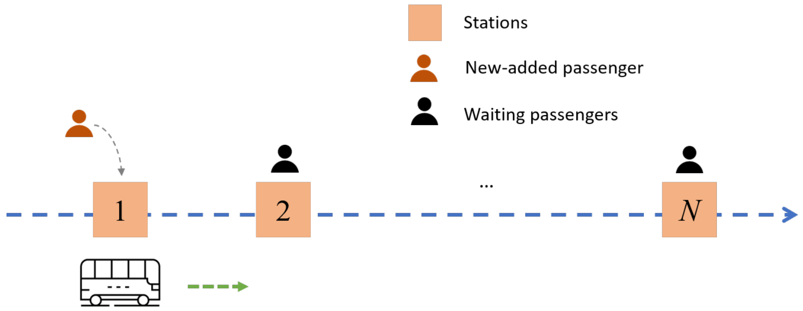


Figure 5-3: Example for arbitrarily large $\beta_{hkr}(\tilde{\boldsymbol{f}})$

Therefore, in this study, we define the convergence criteria based on the value of system travel time (i.e., when the value of the system travel time is relatively stable within a range). Specifically, it is assumed that the MSA algorithm has converged if

$$\left| Z(\tilde{\boldsymbol{f}})^{(n)} - \frac{1}{N^{\text{Cvg}}} \sum_{n'=n-N^{\text{Cvg}}}^{n-1} Z(\tilde{\boldsymbol{f}})^{(n')} \right| \leq \epsilon \tag{5.35}$$

where $Z(\tilde{\boldsymbol{f}})^{(n)}$ is the system travel time at the n-th iteration and ϵ is a predetermined threshold. Eq. 5.35 means that when the current system travel time is close to its average value of the last N^{Cvg} iterations, the algorithm terminates. Taking the average of the last N^{Cvg} iterations can mitigate the impact of fluctuations caused by the discontinuity of the system travel time.

The whole solution algorithm is described in Algorithm 4. Line 6 indicates the MSA step. Lines 10 and 11 mean that we will use the path shares with the smallest system travel time over the last $N^{\text{Cvg}} + 1$ iterations.

Let $\boldsymbol{p}^*$ be the optimal path shares by from Algorithm 4. To realize the optimal path shares in the real world, the following system design can be used:

- Transit operators deploy the recommendation system to smartphone apps, websites, and electrical screens at stations.

- Passengers, when using the system, input their origins, destinations, and departure times.

Algorithm 4 Solution procedure of the robust optimal flow problem

```
Initialize $\boldsymbol{p}^{(0)}$ (e.g., uniform path shares), $\boldsymbol{d}^{(0)}$ (e.g., nominal demand) and specify $N^{\text{Cvg}}, \epsilon$.
Set iteration counter $n = 0$.
do
    $Z(\tilde{\boldsymbol{f}})^{(n)}, \boldsymbol{\beta}(\tilde{\boldsymbol{f}})^{(n)} = \text{Sim-FOA}(\boldsymbol{d}^{(n)}, \boldsymbol{p}^{(n)})$
    Solve the RC problem (Eq. 5.32) with $Z(\tilde{\boldsymbol{f}})^{(n)}$ and $\boldsymbol{\beta}(\tilde{\boldsymbol{f}})^{(n)}$ as inputs, and return $\hat{\boldsymbol{p}}^{(n+1)}$
    $\boldsymbol{p}^{(n+1)} = \frac{1}{n+1}\hat{\boldsymbol{p}}^{(n+1)} + (1 - \frac{1}{n+1})\boldsymbol{p}^{(n)}$
    Solve the WD problem (Eq. 5.33) with $\boldsymbol{p}^{(n+1)}$ as input and return $\boldsymbol{d}^{(n+1)}$
    $n = n + 1$
while $n \leq N^{\text{Cvg}}$ or $\left| Z(\tilde{\boldsymbol{f}})^{(n)} - \frac{1}{N^{\text{Cvg}}} \sum_{n'=n-N^{\text{Cvg}}}^{n-1} Z(\tilde{\boldsymbol{f}})^{(n')} \right| > \epsilon$
$n^* = \arg\min_{n'=n-N^{\text{Cvg}},\dots,n} Z(\tilde{\boldsymbol{f}})^{(n')}$
return $\boldsymbol{p}^{(n^*)}$
```

- For a passenger input OD pair k and departure time h, the system will return a single recommended path r to them with probability p^*_{hkr}.

In this way, we expect the final path flows are close to the system optimal path flows if passengers follow the recommendation.

5.4 Model extensions

5.4.1 Solving the model in a rolling horizon

The model discussed in the previous section is a one-shot solution for path recommendation, which means the model will be run at the beginning of an incident (h_0) and output the recommendations for the whole period of interest $[h_0, h_H]$. In application, the model would be implemented in a rolling horizon framework.

Specifically, at time interval $\tilde{h}$, we first update the demand and supply information, including new demand estimates, new demand uncertainty sets, new available path sets, new service routes and frequencies, new incident duration estimates, etc. Based on the formulation above (i.e., let $h_0 = \tilde{h}$)), we solve the model to obtain recommendations for time $[\tilde{h}, h_H]$. But we only implement the recommendation strategies for the current time $\tilde{h}$ (i.e., $p^*_{\tilde{h}kr}$). In this way, the new information obtained with the evolution of the incident and system operations can be used to improve model performance (this is known as adaptive RO).

5.4.2 Incident duration uncertainty

In this study, we assume operators have a reasonable estimate of incident duration. However, it is possible that we can only obtain a distribution of incident duration. In this section, we show that our formulation can be easily extended to capture the incident duration uncertainty with stochastic optimization (SO) techniques[2].

Let the set of all possible incident scenarios be Ω. For example, Ω may include incidents with duration of 30, 40, or 50 minutes. For each scenario $\xi \in \Omega$, we denote $\beta_{hkr}(\tilde{\boldsymbol{f}};\xi)$ and $Z(\tilde{\boldsymbol{f}};\xi)$ as the approximated gradient and current system travel time under flow $\tilde{\boldsymbol{f}}$ and incident scenario ξ. Hence, the objective function for the RO problem becomes:

$$\begin{aligned}\mathbb{E}[\hat{Z}(\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y})^{\mathrm{RC}}] = \sum_{\xi\in\Omega}\mathbb{P}(\xi)\left[Z(\tilde{\boldsymbol{f}};\xi) + \sum_{(h,k,r)\in\mathcal{F}}\beta_{hkr}(\tilde{\boldsymbol{f}};\xi)\cdot(d_{hk}\cdot p_{hkr} - \tilde{f}_{hkr})\right] + \rho_{1-\varepsilon}\left\|\boldsymbol{y}_1\right\|_2 \\ + (\boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}})^T\boldsymbol{u}_1 + (\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}})^T\boldsymbol{u}_2 + (\boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T\boldsymbol{v}_1 + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}})^T\boldsymbol{v}_2 \\ + (\Gamma - 1)\cdot(\boldsymbol{1}^T\bar{\boldsymbol{d}})\cdot u_3 \qquad (5.36)\end{aligned}$$

where $\mathbb{P}(\xi)$ is the probability of scenario ξ being realized. The expectation above is taking over different incident scenarios. Define $Z(\tilde{\boldsymbol{f}};\Omega) := \sum_{\xi\in\Omega}\mathbb{P}(\xi)Z(\tilde{\boldsymbol{f}};\xi)$ and $\beta_{hkr}(\tilde{\boldsymbol{f}};\Omega) := \sum_{\xi\in\Omega}\mathbb{P}(\xi)\beta_{hkr}(\tilde{\boldsymbol{f}};\xi)$, substituting them into the objective function

$$\begin{aligned}\mathbb{E}[\hat{Z}(\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y})^{\mathrm{RC}}] = \sum_{(h,k,r)\in\mathcal{F}}\beta_{hkr}(\tilde{\boldsymbol{f}};\Omega)\cdot(d_{hk}\cdot p_{hkr} - \tilde{f}_{hkr}) + \rho_{1-\varepsilon}\left\|\boldsymbol{y}_1\right\|_2 + (\boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}})^T\boldsymbol{u}_1 \\ +(\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}})^T\boldsymbol{u}_2 + (\boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T\boldsymbol{v}_1 + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}})^T\boldsymbol{v}_2 + (\Gamma - 1)\cdot(\boldsymbol{1}^T\bar{\boldsymbol{d}})\cdot u_3 + Z(\tilde{\boldsymbol{f}};\Omega) \\ (5.37)\end{aligned}$$

[2] The reason for using SO, instead of RO, to capture incident duration uncertainty is that the worst-case scenario for the incident duration is always the largest one, which makes the problem trivial and may not reflect reality.

As the constraints in the RO problem are not related to incident scenarios (i.e., $\beta_{hkr}(\tilde{\boldsymbol{f}})$ and $Z(\tilde{\boldsymbol{f}})$ are not included in the constraint part), this implies that incorporating the incident duration uncertainty with SO only requires a change in the objective function.

5.5 Case study design

In the case study, we consider an actual incident in the Blue line of the Chicago Transit Authority (CTA) urban rail system (Figure 5-4). The incident starts at 8:14 AM and ends at 9:13 AM on Feb 1st, 2019 due to infrastructure issues between Harlem and Jefferson Park stations. The entire Blue Line was suspended. During the disruption, the Loop (Chicago CBD area) is the destination for most passengers. Usually, there are four paths leading to the Loop: 1) using Blue Line (i.e., waiting for the system to recover), 2) using the parallel bus lines, 3) using the North-South (NS) bus lines to transfer to the Green Line, and 4) using the West-East (WE) bus lines to transfer to the Brown Line. Based on the service structure, we can construct the route sets R_k for each OD pair k.

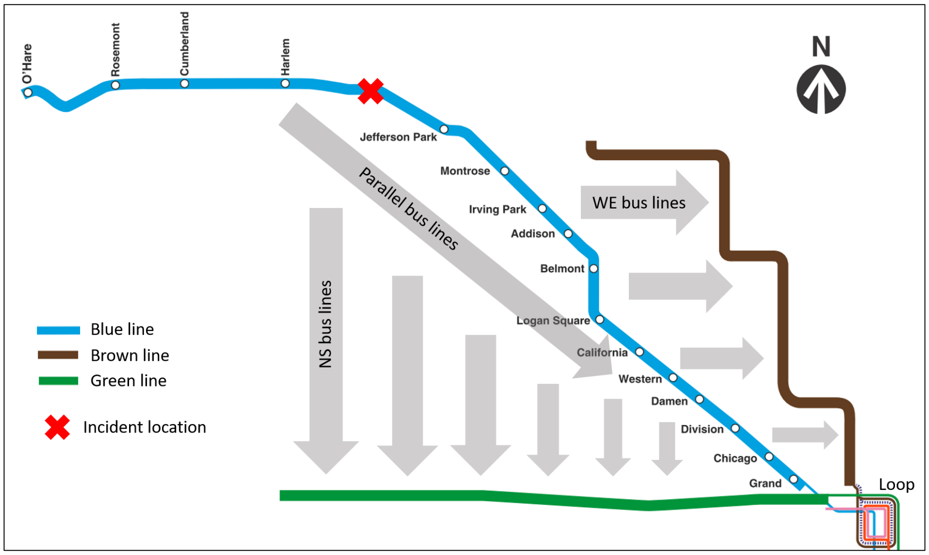


Figure 5-4: Case study diagram

5.5.1 Parameter setting

$\mathcal{K}$ is the set as all OD pairs with origins at the Blue Line and destinations at the Loop. The time interval is set to $\tau = 10$ mins. The time period with recommendation is set as $h_H = 10$, corresponding to 9:44 - 9:54 AM (i.e., 50 minutes after the end of the incident). In this study, we assume that the incident duration is known or can be reasonably estimated. The factor of total demand level, Γ, is set to 1.1, which is the 90% percentile of the total demand distribution.

5.5.2 Quantification of uncertainty sets

The demand uncertainty is determined by the nominal demand $\bar{\boldsymbol{d}}$, covariance matrix $\boldsymbol{\Sigma}$ (which can be used to get $\boldsymbol{D}$), and upper and lower bounds for demand (i.e., $\boldsymbol{d}^{\mathrm{U}}$, $\boldsymbol{d}^{\mathrm{L}}$, $\boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}}$, $\boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}}$). These can be estimated from historical demand. However, as the demand on the incident day is smaller than usual given that some passengers may leave the system, we cannot directly use normal day smart card data as historical demand. One possible solution is to use data from previous days with similar incidents. Nevertheless, this is usually unavailable due to the lack of enough similar incidents. Hence, in this study, we first use survey results and historical smart card data to generate "synthetic historical demand" samples, and then estimate the uncertainty set from the samples.

There are two sources of demand uncertainty: 1) the inherent demand variations across different days and 2) the uncertainty of how many passengers left the PT system during the incident. The first part can be captured by historical smart card data (without incidents). The second part is approximated by the survey results. According to previous survey-based studies, the proportion of the passengers leaving the PT system during incidents is around 10%$\sim$30% [146, 106]. Then, the "synthetic historical demand" is generated as follows:

- Collect smart card data from a recent workday and calculate the demand vector without passengers leaving the system for each (h, k) (the demand for $h = 0$, i.e., offloading passengers, can be obtained using the simulation model).

- For each (h, k), we randomly draw a proportion of leaving passengers from a uniform distribution $\mathcal{U}(10\%, 30\%)^3$. The demand after removing the leaving passengers is the incident period demand vector.

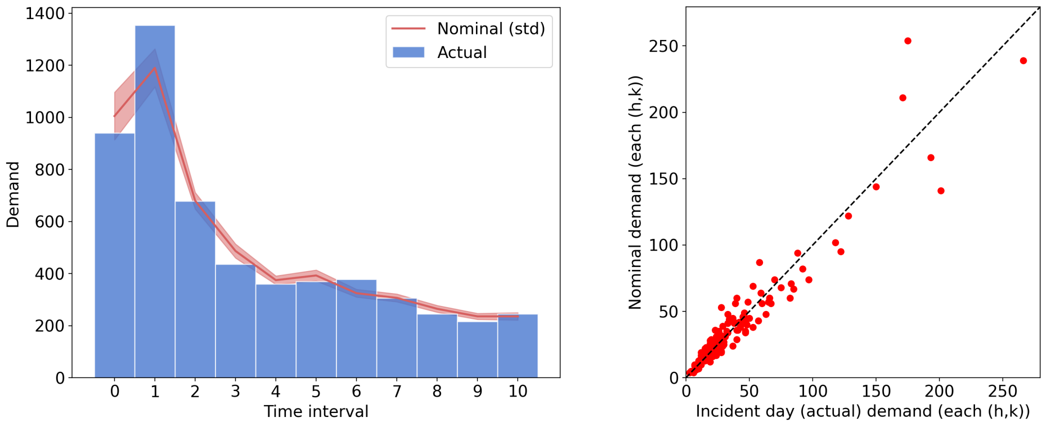
We collected a total of 16 weekdays from Jan 2019 (the previous month of the incident day) and generated 16 sample demand vectors. The mean value is used as the nominal demand $\bar{\boldsymbol{d}}$ and the co-variance matrix $\boldsymbol{\Sigma}$ is estimated from these samples. The upper and lower bounds for demand (i.e., $\boldsymbol{d}^{\mathrm{U}}$, $\boldsymbol{d}^{\mathrm{L}}$, $\boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}}$, $\boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}}$) are set as the samples' maximum and minimum values, respectively.

The hyperparameter $\rho_{1-\varepsilon}$ for the ellipsoidal uncertainty set are chosen from these values: {0, 0.25, 0.52, 0.84, 1.28, 1.64, 2.33}, which corresponds to the {50, 60, 70, 80, 90, 95, 99} percentiles of the standard normal distribution. Note that $\rho_{1-\varepsilon} = 0$ represents the case of no uncertainty (i.e., nominal model).

5.5.3 Data description

The nominal and actual (incident day) demand comparison is shown in Figure 5-5. The total nominal demand is 5,499, similar to the total actual demand (5,531), implying that introducing a proportion of leaving passengers (i.e., 10% - 30%) can capture the demand reduction on the incident day. We also observe that the aggregate nominal demand for each time interval is similar to that of the incident day. The major differences happen at the first two time intervals ($h = 0, 1$). However, looking at the demand for each (h, k) (Figure 5-5b), the differences are more prominent. The discrepancy between nominal and actual demands indicates the potential for the RO approach to perform better.

[3] We use uniform distribution because we have no distributional information of the leaving passenger proportions



(a) Total demand for each time interval h

(b) Demand comparison for each (h, k)

Figure 5-5: Demand patterns

Table 5.1 shows the results of the Mardia test of multivariate normality [179] for demand samples. The Mardia test is used to check whether the sample's multivariate skewness and kurtosis are consistent with a multivariate normal distribution. If both are satisfied, we can assume the samples are multivariate normally distributed. We observe that, in Table 5.1, the synthetic historical demands have consistent skewness but inconsistent kurtosis with the multivariate normal distribution, suggesting that they are not multivariate normally distributed. However, as skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean, the Mardia Skewness testing shows that the demand distribution is symmetric. Hence, it is still reasonable to use the ellipsoidal uncertainty set to describe a symmetric distributed random variable. Moreover, as mentioned in Remark 6, the distribution of a variable does not affect the definition of the uncertainty set (it only affects the calculation of probability guarantees).

5.5.4 Benchmark models

The following approaches are used to obtain benchmark path shares.

Uniform path shares. The uniform path shares are defined as $p_{hkr} = \frac{1}{|R_k|}$ $\forall\ r \in R_k$. This is a naive model corresponding to the intuition of "distributing passengers

Table 5.1: Mardia test of multivariate normality

Test	p-value	Test	p-value
Mardia Skewness	1.00	Mardia Kurtosis	0.00

Note: The null hypothesis is that the samples are multivariate normally distributed. A small p-value indicates we are more likely to reject the null hypothesis.

to different paths” when no information is available.

Capacity-based path shares. The capacity-based path shares aim to assign passengers to different paths according to the path capacity. Specifically, for a path r in OD pair k and time h, we calculate the path capacity as the total available capacity of all vehicles passing through the first boarding station of the path (denoted as C_{hkr}). The capacity-based path shares are defined as

$$p_{hkr} = \frac{C_{hkr}}{\sum_{r \in R_k} C_{hkr}} \quad \forall\, r \in R_k, h \in \mathcal{H}, k \in \mathcal{K}, \tag{5.38}$$

For example, for a path consisting of an NS bus route and the Green Line, C_{hkr} is calculated as the total available capacity of all buses at the boarding station of the NS bus route during time interval h. The available capacity can be obtained from the simulation model using historical demand. The available capacity for the Blue Line (i.e., incident line) depends on the revised schedules during the incident (i.e., the service suspension is considered). When no trains operate on the Blue Line, the corresponding C_{hkr} will be zero.

Status-quo path shares. The status-quo path shares are the inferred path choices of passengers on the incident day. During the incident period, the demand on the WE, NS, and parallel bus lines experience an increase. The difference from the average demand on normal days can be seen as the number of passengers choosing the corresponding path. Hence, by identifying the demand increase for all nearby bus stops, we can get the number of passengers using the parallel bus, NS+Green, and WE+Brown paths for each OD pair k and time interval h. However, the number of waiting passengers in the Blue Line cannot be directly inferred because the CTA system does not record the tap-out information. Hence, we approximate the proportion of waiting passengers based on survey results [96]. Rahimi et al. [96] used a survival model to analyze the waiting time tolerance of CTA riders during a service disruption. The model results provide the proportion of waiting passengers given different system recovery times. Therefore, the status-quo path shares are inferred as follows:

- Step 1: Given the current time interval h and the incident end time T_e, the remaining time until the end of the incident is $T_e - h$. Therefore, if passengers choose to wait, their waiting time will also be $T_e - h$. Based on the hazard model in Rahimi et al. [96], we can obtain the proportion of waiting passengers given the waiting time, denoted as $p_{\text{wait}}(T_e - h)$.

- Step 2: For each OD pair k and time interval h, the number of passengers using the parallel bus, NS+Green, and WE+Brown paths can be calculated based on demand increase compared to the normal demand. Let the demand increase for path r of OD pair k at time h be DI_{hkr}, where $r \in R_k \setminus \{r_{\text{wait}}\}$, r_{wait} represents the path of waiting for the Blue Line.

- Step 3: The status quo path shares are calculated as follows:

$$p_{hkr_{\text{wait}}} = p_{\text{wait}}(T_e - h) \quad \forall\, h \in \mathcal{H}, k \in \mathcal{K}, \tag{5.39}$$

$$p_{hkr} = (1 - p_{hkr_{\text{wait}}}) \cdot \frac{DI_{hkr}}{\sum_{r \in R_k \setminus \{r_{\text{wait}}\}} DI_{hkr}} \quad \forall\, r \in R_k \setminus \{r_{\text{wait}}\},\ h \in \mathcal{H}, k \in \mathcal{K} \tag{5.40}$$

5.6 Results

In this section, we demonstrate the model's performance in two steps. In the first step, results of the optimization model without uncertainty (i.e., the nominal model with $\rho_{1-\epsilon} = 0$) are compared with the three benchmark path shares. In the second step, we compare the results from the robust model with the results from the nominal model in order to assess the value of considering uncertainties in generating path recommendations.

5.6.1 Model convergence

Figure 5-6 shows the convergence of the nominal ($\rho_{1-\epsilon} = 0$) and robust (with $\rho_{1-\epsilon} = 0.84$) models. The simulation-based linearization and MSA successfully decrease the system travel time. The model converges within 35 iterations. Note that the optimal cost for the robust model is higher than the nominal model. This is expected since the robust model assumes the worst-case demand (by definition with higher system travel time). The performance of the corresponding path recommendations will be evaluated based on the actual demand (discussed in the next section).

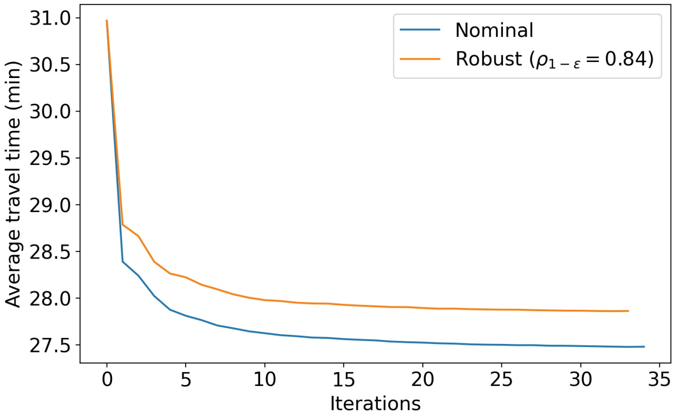


Figure 5-6: Convergence of optimization models

5.6.2 Model evaluation

The optimization model only utilizes information about the nominal demand and the associated uncertainty set. The actual demand is unknown when running the model (otherwise there are no uncertainties). After obtaining the path shares (either from optimization or the benchmark models), the recommendation strategies are evaluated based on the actual incident day demand using the simulation model. We assume passengers would follow the path recommendation. The simulation model can output the travel times of every passenger in the system, and can be used to compare the performance for the path shares obtained from the various approaches. Performance is measured in terms of average travel time and average waiting time.

5.6.3 Nominal vs. Benchmark models

Table 5.2 compares the results for different path shares, The result of no incident scenario is also shown for comparison. The average travel times are calculated over all passengers (a total of 27,007 passengers) and the passengers who originally planned to use the Blue Line (i.e., passengers who are provided with recommendations, a total of 5,531 passengers, a subset of the 27,007 passengers). Results show that the optimization-based path shares outperform all benchmark models. For all passengers in the system, the average travel time is reduced by 9.1% compared to the status quo. And for the incident line passengers, the reduction is even higher (20.6%).

Recommendations based on the uniform path shares result in worse performance than the status quo scenario. This implies that current passengers' choices are not random and show some rationality. The capacity-based path shares can also reduce the system travel time significantly (by 6.9%). However, as the capacity-based path recommendations do not capture the spatial and temporal changes in available capacity due to passenger flow re-distribution, they are worse than the optimization-based results.

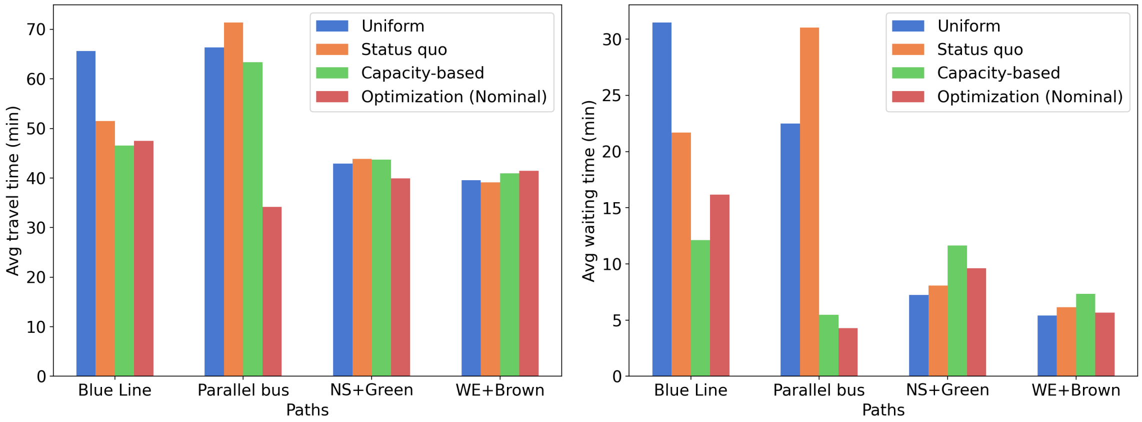
Compared to the no incident scenario, we find that the influence of incidents is significant. Path recommendations can only alleviate the impact of service disruption but are far from eliminating. Even with the optimization-based path recommendations, we still have more than two times of travel time for incident-line passengers compared to the no incident situation.

Table 5.2: Average travel time comparison

Scenarios	All passengers (# 27,003)		Incident-line passengers (# 5,531)	
	Avg travel time (min)	% change[1]	Avg travel time (min)	% change[1]
No incident	21.81	-	18.95	-
Uniform	31.02	+1.7%	54.64	+6.4%
Status quo	30.49	0%	51.34	0%
Capacity-based	28.36	-6.9%	43.23	-15.8%
Optimization (nominal)	27.71	**-9.1%**	40.75	**-20.6%**

[1]: changes compared to the status quo scenario

Figure 5-7 shows the average travel time and waiting time for different paths for all incident line passengers. We observe that the optimization-based path recommendations have more consistent travel time across the four types of paths, implying a better utilization of the system's capacity. However, for other recommendation strategies, passengers using parallel buses have significantly longer travel times than those using other alternatives. Figure 5-7 also shows that the average waiting time for the status quo scenario is around 30 minutes, which means most passengers chose to use the parallel bus during the incident, causing severe congestion. However, with the optimization-based path shares, the average waiting time for the parallel bus is less than 5 minutes (around a headway).



(a) Average travel time for different paths

(b) Average waiting time for different paths

Figure 5-7: Comparison of average travel time and waiting time of different paths for incident line passengers

The objective of this study is to minimize the system travel time. However, under the optimal path shares, some passengers' travel time may be increased compared to the status quo. Figure 5-8 shows the distribution of changes in individual travel time (optimization-based minus the status quo) for all passengers whose path choice under the recommendation scenario is different than their choice in the status quo scenario. Most passengers experience lower travel times. However, some passengers become worse off after following the path recommendations. This is a typical drawback of system optimal (first-best) assignment [181]. Future studies may explore a Pareto-improving (second-best) path recommendation that ensures no individual becomes worse-off. In reality, when implementing the recommendations, some paths that lead to extremely worse travel time compared to the status quo can be dropped from the solution.

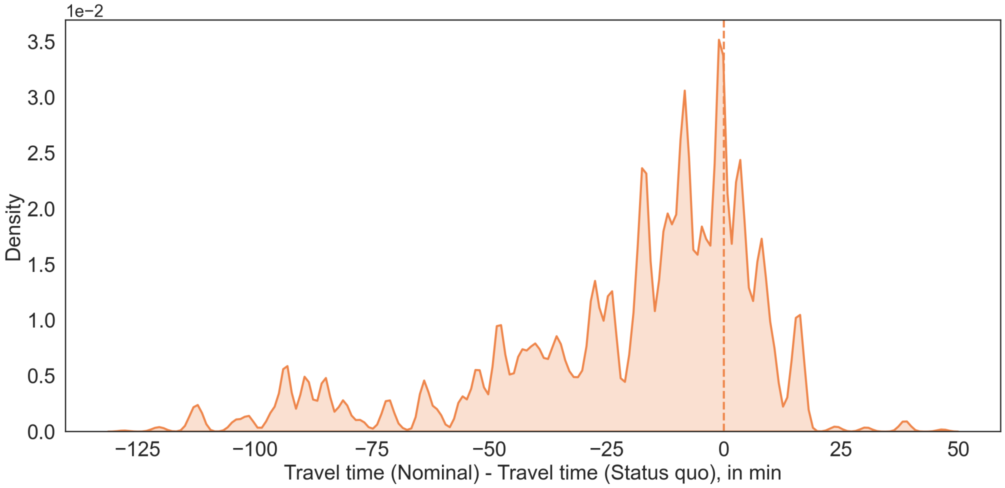


Figure 5-8: Distribution of the change in individual travel time (not including passengers without changes as they will distort the distribution with too much density concentrated at zero)

5.6.4 Robust models vs. Nominal model

Model comparison under actual demand

Figure 5-9 compares the results, in terms of travel time, of the RO approach with different values of $\rho_{1-\epsilon}$ under the actual demand. For all values of the robust model except for $\rho_{1-\epsilon} = 2.33$, the RO approach shows better performance than the nominal model. This implies that considering the demand uncertainty in determining the recommendation can further improve the effectiveness of path recommendation strategies. The best value is $\rho_{1-\epsilon} = 0.84$, where the travel time for the incident line passengers is reduced by 2.91% compared to the nominal model. Note that the percentage decreases are relatively small because some passengers' travel times are not changed. If we only look at incident-line passengers with travel time changes, the average travel times are 47.6 min and 37.9 min for the nominal and RO ($\rho_{1-\epsilon} = 0.84$) scenarios, respectively, where the travel time reductions are 20.4%.

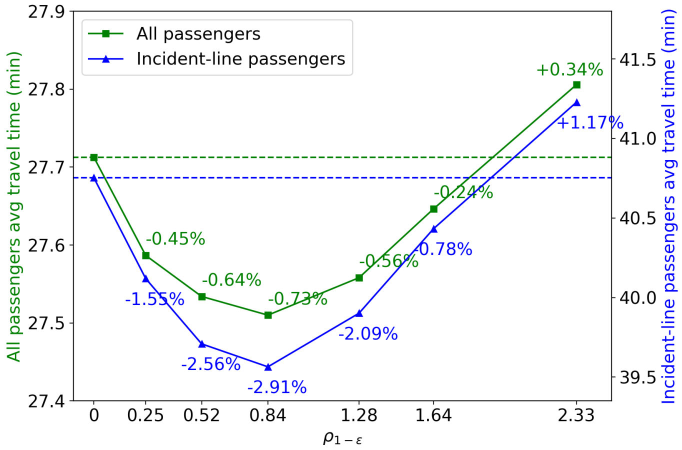


Figure 5-9: Performance of RO. The percentage changes are compared to the nominal scenario

Note that using $\rho_{1-\epsilon} = 2.33$ results in the largest uncertainty set compared to other values. This reflects a very conservative scenario where the agency prefers to plan against a very high realization of demand. In this case, the worst-case demand patterns may deviate from the actual demand too much, thus performing worse than the nominal model. Figure 5-10 illustrates the worst-case demand for different values of $\rho_{1-\epsilon}$. The worst-case demands for the $\rho_{1-\epsilon} = 0.52, 0.84, 1.28$ scenarios are closer to the actual demand, while $\rho_{1-\epsilon} = 2.33$ overestimates the demands, especially for the earliest periods ($h = 0, 1$) (which are the most critical periods). These results are consistent with the travel time performance in Figure 5-9.

Model comparison under random demand

To further validate the model's performance, we test the performance of the solution obtained from the RO approach on the 16 demand samples generated in Section 5.5.2. These demand samples represent different possible realizations of the incident day demand. Figure 5-11 shows the compassion of the random demand samples versus the actual and nominal demands. Notice that the random demand samples include both high and low demand scenarios, which can better validate the performance of the RO approach under different demand patterns.

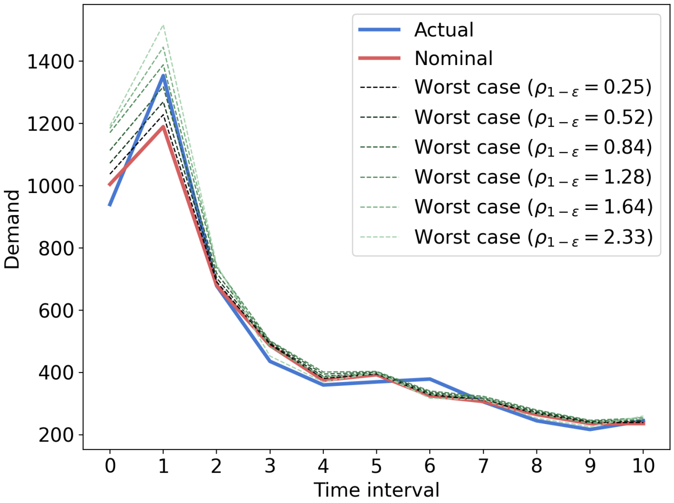


Figure 5-10: Worst-case demand patterns

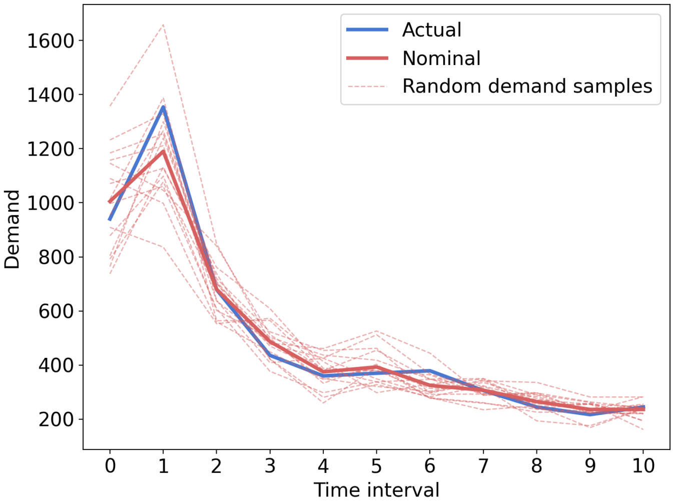


Figure 5-11: Random demand patterns for experiments

Table 5.3 compares the results of average travel time for different RO models. The numbers in the table are the mean values of the 16 experiments. The performances are similar to the results under the actual demand. The RO approach shows better performance than the nominal model for all values of $\rho_{1-\epsilon}$ the robust model except for $\rho_{1-\epsilon} = 2.33$. The reasons may be that the RO approach focuses more on critical OD pairs and time intervals where the path recommendations for them are considered more important for system performance.

Table 5.3: Average travel time comparison for RO models

Models	All passengers		Incident-line passengers	
	Avg travel time (min)	% change[1]	Avg travel time (min)	% change[1]
Nominal ($\rho_{1-\epsilon} = 0$)	27.79	-	41.08	-
$\rho_{1-\epsilon} = 0.25$	27.70	-0.32%	40.57	-1.23%
$\rho_{1-\epsilon} = 0.52$	27.65	-0.48%	40.24	-2.05%
$\rho_{1-\epsilon} = 0.84$	27.64	-0.54%	40.13	-2.31%
$\rho_{1-\epsilon} = 1.28$	27.68	-0.39%	40.41	-1.62%
$\rho_{1-\epsilon} = 1.64$	27.74	-0.17%	40.83	-0.60%
$\rho_{1-\epsilon} = 2.33$	27.86	+0.27%	40.47	+0.96%

[1]: changes compared to the nominal model

5.7 Conclusion and discussion

In this chapter, we propose a path recommendation model to mitigate the congestion during public transit disruptions. Passengers with different ODs and departure times are recommended alternative paths to use such that the total system travel time is minimized. To tackle the non-analytical formulation of travel times due to left behind, we propose a simulation-based first-order approximation to transform the original problem into a linear programm and solve the new problem iteratively with MSA. Uncertainties in demand are modeled using RO techniques to protect the path recommendation strategies against inaccurate estimates. A real-world rail disruption scenario in the CTA system is used as a case study. Results show that even without considering uncertainty, the nominal model can reduce the system travel time by 9.1% (compared to the status quo), and outperforms the benchmark capacity-based path recommendation. The average travel time of passengers in the incident line is reduced more (-20.6% compared to the status quo). After incorporating the demand uncertainty, the robust model further reduces the system travel time. The best robust model with $\rho_{1-\epsilon} = 0.84$ decreases the average travel time of incident-line passengers by 2.91% compared to the nominal model.

The performance improvement by incorporating demand uncertainty is not very significant. The reason may be that demand variations at the incident situation have a limited impact on the optimal path shares. Notice that the demand during an incident is already very high for the system (due to the reduced supply level). Hence, the path recommendation patterns under nominal and worst-case demand may be similar. However, the methodology presented in this study provides a general way to deal with PT demand uncertainty. It can be used for other operations control, optimization, planning, or recommendation applications.

Though we discussed potential model extensions with rolling horizon and incident duration uncertainty, we did not implement these extensions in the case study as the focus has been on the methodology for solving the problem. Incorporating real-time information as an adaptive RO would generally increase model performance [170]. This presents an interesting future research direction. Other future research directions include the following. 1) Current demand uncertainty sets need to be quantified with a budget factor $\rho_{1-\varepsilon}$. The choice of budget factor usually relies on numerical testing [182, 177]. Future studies may also develop data-driven uncertainty quantification methods to automate the hyperparameter tuning task. 2) As shown in Figure 5-8, the system optimal path recommendation may result in worse-off travel time for some passengers, causing equity and fairness issues. Future studies may consider incorporating Pareto-improving constraints to ensure that all passengers are better-off if following our recommendation. 3) In this study, we assume that passengers follow the recommendation. Non-compliance, however, if present, may lead to the actual path flows deviating from the optimal ones. Future research may focus on approaches for path recommendations that capture behavior uncertainty. 4) Finally, this study presents an OD-based (aggregated) path recommendation regime. Passengers with the same OD and departure time are treated homogeneously. In reality, different passengers may have different preferences on path choices. And these preferences can affect their compliance with recommendations. Future studies can develop an individualized path recommendation system considering heterogeneous passenger preferences.

Chapter 6

Individual-based path recommendation under public transit service disruptions considering behavior uncertainty and equity

6.1 Introduction

6.1.1 Background and challenges

With aging systems and near-capacity operations, service disruptions often occur in urban public transit (PT) systems. These incidents may result in passengers delays, cancellation of trips, and economic losses [6].

During a significant disruption where the service is interrupted for a relatively long period of time (e.g., 1 hour), affected passengers usually need to find an alternative path or use other travel modes (such as transfer to another bus route). However, due to a lack of knowledge of the system (especially during incidents), the routes chosen by passengers may not be optimal or even cause more congestion [43]. For example, during a rail disruption, most of the passengers may choose bus routes that are parallel to the interrupted rail line as an alternative. However, given limited bus capacity, parallel bus lines may become oversaturated and passengers have to wait for a long time to board due to being denied boarding (or left behind).

One of the strategies to better guide passengers is to provide path recommendations so that passenger flows are re-distributed in a better way and the system travel times are minimized. This can be seen as solving an optimal passenger flow distribution (or assignment) problem over a public transit network. However, different from the typical flow redistribution problem, there are several unique characteristics and challenges for the path recommendation problem under PT service disruptions.

- Passengers may have different preferences on different alternative paths. This heterogeneity suggests that we cannot treat a group of passengers simply as flows. Individualization is needed in the path recommendation design.

- Passengers may not follow the recommendation. When providing a specific path recommendation to a passenger, their actual path choice is uncertain (though the recommendation may change their preferences). This behavior uncertainty brings challenges in the recommendation system design and has not been considered in the path recommendation literature. In the context of individualization, the behavior uncertainty is also individual-specific, which requires a more granular modeling approach.

- From the operator's point of view, the objective of path recommendations is to reduce system congestion by better utilizing available capacity. However, under system optimal flow patterns, passengers with the same origin, destination (OD), and departure time may end up with very different travel times due to being recommended different paths where some paths are shorter and some are longer [183], resulting in equity issues. Therefore, we do not want the path recommendations to result in passengers having large differences in travel times.

6.1.2 Organization and contributions

To tackle these challenges, this study proposes an individual-based path recommendation model to reduce the system congestion during public transit disruptions, considering behavior uncertainty and passenger travel time equity. We first formulate an optimal flow problem as a linear program based on the model of Bertsimas et al. [184], which solves the optimal path flows for each OD pair and time interval that minimize the system travel time. Then, we add the recommendation decision variables, $x_{p,r}$ (binary variable indicating whether path r is recommended passenger p) and associated constraints to capture the behavior uncertainty. The behavior uncertainty is modeled with a conditional path choice probability distribution for each passenger given their received path recommendation. We introduce two new concepts: ϵ-feasible flows and Γ-concentrated flows, to connect the optimal flow problem with the conditional path choice probabilities. The individual path recommendation problem with behavior uncertainty is a mixed-integer program. We solve it efficiently with Benders decomposition. Finally, we use a post-adjustment heuristic to address the equity requirement. The proposed approach is implemented in the Chicago Transit Authority (CTA) system with a real-world urban rail disruption as the case study.

The main contributions of this chapter are as follows:

- To the best of the authors' knowledge, this is the first article dealing with individual path recommendations under public transit service disruptions considering behavior uncertainty and equity. Previous studies only considered uncertainty in demand [45] or incident duration [161]. And for the objective function, they either focus on minimizing travel time or maximizing individual preferences [155]. Equity has not been considered in the literature.

- To model behavior uncertainty, this chapter proposes a framework with prior path utility and posterior path choice distribution given recommendations. We use two new concepts: ϵ-feasibility and Γ-concentration, to control the mean and variance of path flows due to behavior uncertainty and transform these two requirements to linear constraints in the optimization model using Chebyshev's inequality.

- Benders decomposition (BD) is used to solve the mixed-integer individual path recommendation problem efficiently. Under BD, the master problem becomes a small-scale integer program and the sub-problem reduces to a linear program.

- This chapter mathematically defines the equity requirement in the individual path recommendations, and proposes a post-adjustment heuristic method to solve it. We also propose an integrated mixed-integer programming formulation with both behavior uncertainty and equity requirement, discuss the difficulty in solving the corresponding problem, and highlight the importance of the post-adjustment heuristic.

The remainder of the chapter is organized as follows. Literature review is discussed in Section 6.2. In Section 6.3, we describe the problem conceptually and analytically. Section 6.4 develops the solution methods, including the optimal flow problem formulation, modeling of the behavior uncertainty, Benders decomposition, and the post-adjustment heuristic for equity. Section 6.5 discusses model extensions and generalizability. In Section 6.6, we apply the proposed model on the CTA system as a case study. The model results are analyzed in Section 6.7. Finally, we conclude the chapter and summarize the main findings in Section 6.8.

6.2 Literature review

6.2.1 Individualized recommendations system

Individualized recommendations design is a popular topic in the field of computer science and operations research, with many real-world implications such as Ads ranking [185, 186], mobile news recommendations [187], travel recommendations [188], etc. Most of these recommendation systems focus on individual preference maximization, which, in return, can increase indicators of interest such as click-through rate (CTR) and conversion rate. However, in the context of path recommendations under disruptions, though respecting passenger's preferences is important, the ultimate goal is to minimize the system travel time and mitigate the impact of disruptions, which is different from typical recommendation design literature. Another difference is that, the typical recommendation systems are usually designed with machine learning algorithms trained with the real-world user and system interaction data because they have to learn users' preferences based on their interaction histories. However, in this study, the system travel time can be evaluated using a network loading model. This implies that, instead of using machine learning models, we can use an optimization formulation to determine the individualized path recommendations that minimizes system travel time.

In summary, different from the typical individualized recommendation system literature, this study focuses on system-level objectives instead of individual-level preferences. It leverages an optimization model to design the recommendation, rather than machine learning models.

6.2.2 Path recommendations during disruptions

Most previous studies on path recommendation under incidents are like designing a "trip planner". That is, the main objective is to find available routes or the shortest path given an OD pair when the network is interrupted by incidents. For example, Bruglieri et al. [153] designed a trip planner to find the fastest path in the public transit network during service disruptions based on real-time mobility information. Böhmová et al. [154] developed a routing algorithm in urban public transportation to find reliable journeys that are robust for system delays. Roelofsen et al. [155] provided a framework for generating and assessing alternative routes in case of disruptions in urban public transport systems. To the best of the authors' knowledge, none of the previous studies have considered path recommendations aiming to minimize the system-wide travel time, given equity constraints.

Providing path recommendations during disruptions is similar to the topic of passenger evacuation under emergencies. The objective of evacuation is usually to minimize the total evacuation time. For example, Abdelgawad and Abdulhai [159] developed an evacuation model with routing and scheduling of subway and bus transit to alleviate congestion during the evacuation of busy urban areas. Wang et al. [160] proposed an optimal bus bridging design method under operational disruptions on a single metro line. Tan et al. [161] propose an evacuation model with urban bus networks as alternatives in the case of common metro service disruptions by jointly designing the bus lines and frequencies.

However, although these passenger evacuation papers focus on minimizing the system travel time, there are several differences from this study. First, in our study, the service disruption is not as severe as the emergency situation. We assume the service will recover after a period of time and passengers are allowed to wait. They do not necessarily need to cancel trips or follow the evacuation plan as assumed in previous evacuation studies. Second, in this article, we do not adjust the operations on the supply side. Instead, we focus on providing information to the passengers to better utilize the existing resources/capacities of the system. Third, as mentioned before, this study considers passenger heterogeneity and focuses on individual-level path recommendations, while previous evacuation papers simply model passengers as flows. Besides, we also assume that passengers may not follow the recommendation (i.e., behavior uncertainty) and incorporate equity into consideration, which has not been considered in any evacuation paper before.

6.2.3 Behavior uncertainty

Behavior uncertainty is a well-known challenge in transportation modeling [189]. Typically, passenger's behavior is modeled using various econometrics approaches [132, 190, 191] or machine learning models [192, 193]. These models output the probability distribution for the passenger's possible behavior. At the aggregate level, there are numerous studies using the predicted demand for different transportation applications taking demand uncertainty into consideration, such as ride-sharing [177], transit route planning [194], and supply chain management [195].

However, at the individual level, the number of studies is limited. The main reason is that, individual-level decision-making is usually discrete, it is challenging to use typical robust optimization to address discrete uncertain variables [196]. In terms of stochastic optimization, the number of possible scenarios increases exponentially with the number of individuals in the system. Some studies use simulation to incorporate individual-level behavior uncertainty. For example, Horne et al. [197] use a discrete choice model to simulate how different hybrid energy-economy policies can motivate users' responses. However, to incorporate behavior uncertainty in an optimization model (such as the individual path recommendation model in this study), new modeling techniques are needed.

Another difference in this study compared to previous literature is that the behavior uncertainty (i.e., passenger's response to the recommendation) makes the decision variables (i.e., passenger flow) random variables. Typical robust optimization or stochastic optimization usually assumes the parameters of constraints are random variables, but not the decision variable.

6.2.4 Equity in travel times

Equity has been an important topic in the transportation field [198, 199, 200, 201] and the design of recommendation systems [202, 203, 204]. The motivation for considering equity in this study comes from the well-known trade-off between efficiency and equity in the network flow problems. As known from the canonical static traffic assignment literature, under system optimal solutions (best efficiency), passengers with the same OD pair may have different travel times due to using different paths. Though the total travel time of all passengers is minimized, those passengers who use longer paths are worse-off in the system. On the contrary, under user equilibrium conditions (best equity[1]), all passengers with the same OD pair have the same travel times[2]. User equilibrium is assumed to represent real-world flow patterns without interventions (i.e., the Nash equilibrium from the game theory perspective). When considering controls to reduce the system travel time, equity issues arise. Two well-known tools to increase efficiency and ensure equity is through congestion pricing [205, 206] and tradable credits [207, 208]. The former focuses on charging drivers with advantages in travel time (i.e., those who use shorter paths) and the latter compensates drivers in longer paths.

[1] Here we refer to the simplest horizontal equity and we understand that there are more discussions on the definition of equity

[2] More precisely, the travel time of all paths with passengers are the same in this status.

This study also falls into the topic of using control strategies (i.e., path recommendation) to reduce the system travel time (though in disruption scenarios). Therefore, it is likely that the system optimal recommendations would suggest some passengers use longer paths in order to increase efficiency. resulting in equity issues. None of the previous studies in transit systems have considered equity in the individual path recommendations. For the evacuation literature, some studies have pointed out that there are equity issues in evacuation planning [209, 210, 211, 212, 213]. However, these papers focus on the equity of opportunities in evacuation, aiming to help careless or vulnerable populations who cannot receive services during incidents. In this study, the focus is on travel time equity in the context of using the information as a control mechanism under transit disruptions.

6.3 Problem description

6.3.1 Conceptual description

Consider a service disruption in an urban rail system. During the disruption, some stations in the incident line (or the whole line) are blocked. Passengers in the blocked trains are usually offloaded to the nearest platforms. To respond to the incident, some operating changes are made, such as dispatching shuttle buses, rerouting existing services, short-turning in the incident line, headway adjustment, etc. Assume that all information about the operating changes is available. These changes define a new PT service network and available path sets. Our objective is to develop an individual-based path recommendation model that, when an incident happens, provides a recommended path to every passenger who uses their phones, websites, or electronic boards at stations to enter their origin, destination, and departure time. The recommendation considers the individual's preferences and behavioral histories. Hence, passengers with the same origin, destination, and departure time may get different recommended paths. The overall system aims to minimize the total travel time for all passengers, including passengers in nearby lines or bus routes without incidents (note that these passengers may experience additional crowding due to transfer passengers from the incident line).

Figure 6-1 shows a simple example of the path recommendation problem. In this example, Rail Line 1 has an incident and cannot provide service for a period of time. Both of the two passengers at station A want to go to station C. Assuming that they request path recommendations. The alternative paths include using the bus route (blue dashed line), using the Rail Line 2 (green dashed line), or waiting for the system to recover (i.e., still using Rail Line 1). Note that using either the bus route or Rail Line 2 will take away capacity from passengers who originally use these two services (i.e., the orange passengers in the figure). Hence, the model should consider the total travel time of all four passengers in the system to design recommendation strategies.

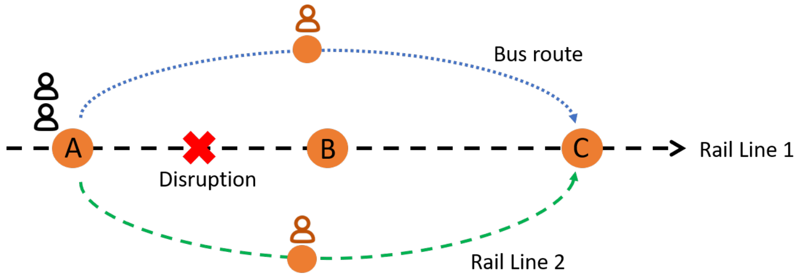


Figure 6-1: Example of the individual path recommendation problem

Moreover, as mentioned in the introduction, behavior uncertainty and equity need to be considered. In this example, if we recommend a passenger to use a bus route, he/she may not follow the recommendation and choose Rail Line 2 instead. Although these two passengers share the same origin, destination, and departure time, they may receive different recommended paths. For example, one of them is recommended to use a bus route and another Rail Line 2. The model should ensure that the travel times of these two paths are similar, otherwise, there will be equity issues.

6.3.2 Analytical description

Let us divide the analysis period into several time intervals with equal length τ (e.g., $\tau = 5$ min). Let t be the integer time index. $t = 1$ is the start of the incident and $t \leq 0$ indicates the time before the incident. Let $\mathcal{P}$ be the set of passengers that will receive path recommendations. We assume $\mathcal{P}$ is known as we can obtain passengers' requests before running the model. Given the revised operation during the incident, let $\mathcal{R}_p$ be the feasible path set for each passenger $p \in \mathcal{P}$. Note that $\mathcal{R}_p$ includes all feasible services that are provided by the PT operator. A path $r \in \mathcal{R}_p$ may be waiting for the system to recover (i.e., using the incident line), or transfer to nearby bus lines, using shuttle services, etc. We do not consider non-PT modes such as TNC or driving for the following reasons: 1) This study aims to design a path recommendation system used by PT operators. The major audience should be all PT users. Considering non-PT modes needs the supply information of all other travel modes and even consider non-PT users (such as the impact of traffic congestion on drivers), which is beyond the scope of this study. Future research may consider a multi-modal path recommendation system. 2) Passengers using non-PT modes can be simply treated as demand reduction for the PT system. So their impact on the PT system can still be captured.

Given a passenger $p \in \mathcal{P}$, we aim to determine $x_{p,r}$ for each p, where $x_{p,r}$ indicates whether path $r \in \mathcal{R}_p$ is recommended to passenger p or not. Assume only one path is recommended to each passenger, we have

$$\sum_{r \in \mathcal{R}_p} x_{p,r} = 1 \quad \forall p \in \mathcal{P} \tag{6.1}$$

Note that we can relax this assumption by designing the recommendation to a passenger as a "composition" including multiple paths or travel times. This generalization is discussed in Section 6.5.1.

$\mathcal{P}$ includes passengers with different origins, destinations, and departure times. If an incident ends at t^{end}, the recommendation should consider a time horizon after t^{end} because there is remaining congestion in the system. Hence, we provide recommendations until time $T^D > t^{\text{end}}$ (e.g., T^D can be one hour after t^{end}). Therefore, the departure times for passenger $p \in \mathcal{P}$ range from $[1, T^D]$ (T^D and t^{end} are both time indices).

The recommendation model will be solved at $t = 1$ and will generate the recommendation strategies $\boldsymbol{x} = (x_{p,r})_{p\in\mathcal{P},r\in\mathcal{R}_p}$ for passengers who depart at time $t \in [1, T^D]$. In reality, the model can be implemented in a rolling horizon manner. Specifically, at each time interval $t \geq 1$, we first update the demand and supply information in the system, including new demand estimates, new to-be-recommended passenger set $\mathcal{P}$, new available path sets $\mathcal{R}_p$, new service routes and frequencies, new incident duration estimates, new onboard passenger estimates, etc. Based on this information, we solve the model to obtain recommendations for passengers with departure time in $[t, T^D]$. But we only implement the recommendation strategies for passengers who depart at the current time t.

Therefore, in the following formulation, we only focus on solving the model at $t = 1$, which is the start of the incident. The whole analysis period includes warm-up and cool-down periods to better estimate the system states (e.g., vehicle loads, passenger travel times, etc.). Therefore, the analysis period is defined as $[t^{\min}, T]$, where $t^{\min} < 1$ (time before the incident) and $T > T^D$. For example, $t^{\min}$ and T can be one hour before and after $t = 1$ and T^D, respectively. And we define all time intervals in the analysis period as $\mathcal{T} = \{t^{\min}, t^{\min} + 1, ..., T\}$. The overall path recommendation framework can be summarized in Figure 6-2.

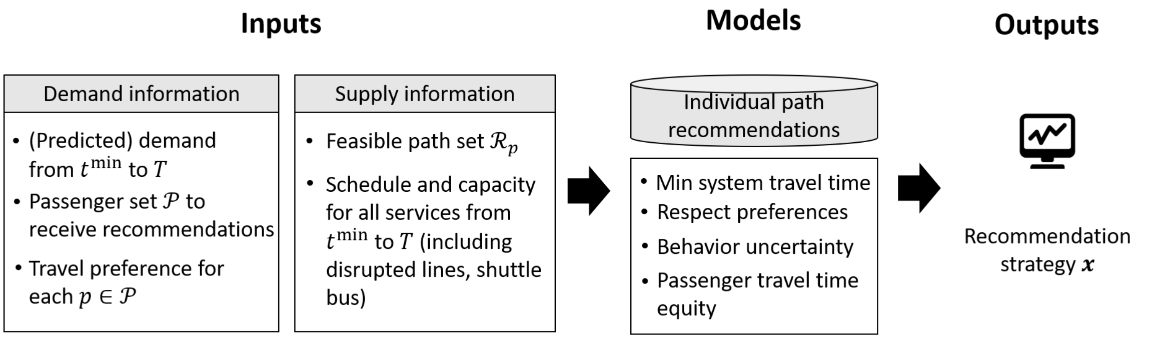

Inputs
Models
Outputs
Demand information
• (Predicted) demand from $t^{\min}$ to T
• Passenger set $\mathcal{P}$ to receive recommendations
• Travel preference for each $p \in \mathcal{P}$
Supply information
• Feasible path set $\mathcal{R}_p$
• Schedule and capacity for all services from $t^{\min}$ to T (including disrupted lines, shuttle bus)
Individual path recommendations
• Min system travel time
• Respect preferences
• Behavior uncertainty
• Passenger travel time equity
Recommendation strategy $\boldsymbol{x}$

Figure 6-2: Problem description and model framework

6.4 Formulation

In this section, we elaborate on the detailed formulation of the individual path recommendation model. Section 6.4.1 develops an optimization model to solve the optimal flow distribution over a public transit network with disruptions. Section 6.4.2 describes how passenger's behavior uncertainty (i.e., non-compliance to recommendation) is modeled based on a random utility maximization framework. Section 6.4.3 provides the overall formulation of the individual path recommendation model by combining the optimal flow model in Section 6.4.1 and the behavior uncertainty component in Section 6.4.2. Section 6.4.4 shows how the individual path recommendation model can be solved efficiently using Benders decomposition. Section 6.4.5 proposes a post-adjustment method to obtain the final recommendation strategy with equity (i.e., the travel times of passengers with the same origin, destination, and departure time do not differ a lot).

The notations used in the chapter are summarized in Table 6.4 (Section 6.9.1).

6.4.1 Optimal flow during disruptions

In this section, we formulate a linear programming (LP) model to solve the optimal flow distribution in a public transit system with service disruptions. Consider an OD pair (u, v) and departure time t. Let $\mathcal{R}^{u,v}$ be the set of feasible paths for OD pair (u, v). Define $q_t^{u,v,r}$ (resp. $f_t^{u,v,r}$) as the number of passengers **in** (resp. **not in**) $\mathcal{P}$ with OD pair (u, v) and departure time t, who use path $r \in \mathcal{R}^{u,v}$. Specifically, $q_t^{u,v,r}$ represents the passenger flows that receive recommendations while $f_t^{u,v,r}$ those do not. Hence, the total path flow in $r \in \mathcal{R}^{u,v}$ is $q_t^{u,v,r} + f_t^{u,v,r}$. Let $d_t^{u,v}$ be the total demand of OD pair (u, v) at time t, we have

$$q_t^{u,v,r} + f_t^{u,v,r} = d_t^{u,v} \quad \forall (u, v) \in \mathcal{W}, t \in \mathcal{T} \tag{6.2}$$

where $\mathcal{W}$ is the set of all OD pairs. As we focus on path recommendations for $\mathcal{P}$, in this study, $q_t^{u,v,r}$ is the decision variable while $f_t^{u,v,r}$ is a known constant (i.e., the estimated demand information). For mathematical convenience, we define $\mathcal{F}$ as the set of all triplets (u, v, r) in the system. And the objective in this section is to find the optimal flows $q_t^{u,v,r}$ $(\forall(u, v, r) \in \mathcal{F}, t \in \mathcal{T})$ that minimize the total system travel time.

The LP-based optimal flow model is adapted from Bertsimas et al. [184] with the following differences: 1) Bertsimas et al. [184]'s model only considers waiting time in the system while ignoring in-vehicle time. In this study, we extend their formulation to include the in-vehicle times. 2) In Bertsimas et al. [184], the capacity constraints are formulated for a vehicle without time indices, which neglects the fact that alighting passengers can release capacity. In this study, we modify the capacity constraints to a vehicle-time-based formulation, which considers the occupancy of only onboard passengers in each time interval. 3) We adapt the model from normal scenarios to incident scenarios by pre-processing the supply of incident lines and pre-defining the system state before the incident. These two operations will be described later.

Consider a path r for OD pair (u, v). A path may include multiple legs, where each leg is associated with the service in a rail or a bus line. For example, the path in Figure 6-3 (indicated by green arrows) has two legs: the first one in the rail line and the second in the bus line. Every leg has a boarding and an alighting station. For example, Leg 1 (resp. 2) in this example has boarding station A (resp. C) and alighting station B (resp. D). Let $\mathcal{I}^{u,v,r} = \{1, ..., |\mathcal{I}^{u,v,r}|\}$ be the set of legs for path r. We use a four-element tuple (u, v, r, i) to represent a leg i of path r for OD pair (u, v), where $i \in \mathcal{I}^{u,v,r}$.

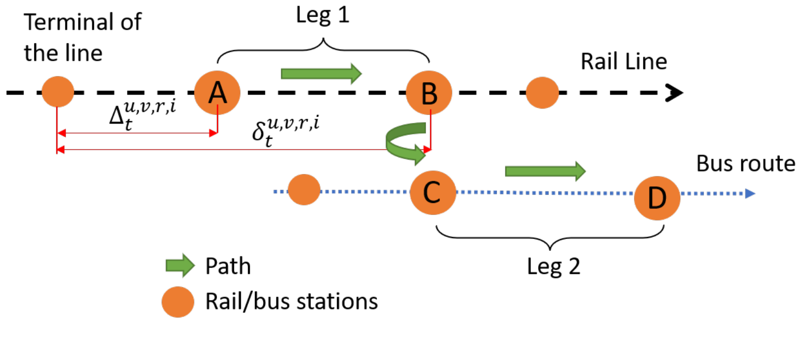


Figure 6-3: Definition of paths and legs

Let $\Delta_t^{u,v,r,i}$ (resp. $\delta_t^{u,v,r,i}$) be the travel time between the **terminal** and the **boarding** (resp. **alighting**) station of leg (u, v, r, i) for a vehicle **departing** from the terminal at time t. Hence, the vehicle's arrival time at the boarding (resp. alighting) station of leg (u, v, r, i) is $t + \Delta_t^{u,v,r,i}$ (resp. $t + \delta_t^{u,v,r,i}$). $\delta_t^{u,v,r,i} - \Delta_t^{u,v,r,i}$ represents the total in-vehicle time of leg (u, v, r, i) for the vehicle. Define $z_t^{u,v,r,i}$ (decision variable) as the total number of on-board passengers in leg (u, v, r, i) who board a vehicle that **had departed** from the terminal at time t.

There are three types of constraints for the network flow description: 1) existing flows constraints, 2) vehicle capacity constraints, and 3) flow conservation constraints.

Existing flows constraints: Although the path recommendations starts at time $t = 1$, there are passengers that already boarded the vehicles. Ignoring these existing flows may lead to overestimation of the system's available capacity. To capture the existing onboard flows at $t = 1$, we define the set of onboard flow indices as

$$\Omega_1 = \{(u, v, r, i, t) : t + \Delta_t^{u,v,r,i} \leq 1 \leq t + \delta_t^{u,v,r,i}\} \tag{6.3}$$

And the existing flow constraints can be expressed as

$$z_t^{u,v,r,i} = \hat{z}_t^{u,v,r,i} \quad \forall (u, v, r, i, t) \in \Omega_1 \tag{6.4}$$

where $\hat{z}_t^{u,v,r,i}$ are constants that capture the existing onboard flows when the incident happens. These flows can be directly obtained from a simulation model or real-time passenger counting data.

Capacity constraints: Transit vehicles have limited capacity. Consider a vehicle departing at time t on line l (referred to as vehicle (l, t)). We denote its total number of onboard passengers at time t' as $O_{l,t,t'}$. Specifically, $O_{l,t,t'}$ can be expressed as

$$O_{l,t,t'}(\boldsymbol{z}) = \sum_{\{(u,v,r,i,t) \in \texttt{Onboard}(l,t')\}} z_t^{u,v,r,i} \quad \forall l \in \mathcal{L}, \forall t \in \mathcal{T}, t' = t, t+1, ..., T_{l,t} \tag{6.5}$$

where $T_{l,t}$ is the time index that vehicle (l, t) arrives at the last station of line l.

$\texttt{Onboard}(l,t')$ is the set of onboard flow indices for vehicle (l,t), defined as

$$\texttt{Onboard}(l,t') = \{(u,v,r,i,t) : \text{Leg } (u,v,r,i) \text{ on line } l, \text{ and } t+\Delta_t^{u,v,r,i} \leq t' \leq t+\delta_t^{u,v,r,i}\} \tag{6.6}$$

Then the capacity constraint is:

$$O_{l,t,t'}(\boldsymbol{z}) \leq K_{l,t} \quad \forall l \in \mathcal{L}, t \in \mathcal{T}, t' = t, t+1, ..., T_{l,t} \tag{6.7}$$

where $K_{l,t}$ is the capacity of the vehicle (l,t). $\mathcal{L}$ is the set of all lines.

Flow conservation constraint: There are two different flow conservation constraints: 1) flow conservation at origin stations and 2) at transfer stations. To ensure the origin flow conservation, the cumulative number of arrival passengers should be larger than the cumulative number of boarding passengers at an origin at any time. This indicates that not all arrival passengers can board due to potentially being left behind because of capacity constraints.

The number of arriving passengers (i.e., demand) for (u,v,r) at time t is $q_t^{u,v,r} + f_t^{u,v,r}$. And the number of boarding passengers at the origin station (i.e., u) at time t is $z_{t'}^{u,v,r,1}$ (i.e., the first leg) with $t' + \Delta_{t'}^{u,v,r,1} = t$. t' is the vehicle departure time from the terminal and $t' + \Delta_{t'}^{u,v,r,1}$ is the time when the vehicle arrives at the boarding station. Therefore, the origin flow conservation constraint can be written as:

$$\sum_{\{t' : t^{\min} \leq t' + \Delta_{t'}^{u,v,r,1} \leq t\}} z_{t'}^{u,v,r,1} \leq \sum_{t'=t^{\min}}^{t} (f_{t'}^{u,v,r} + q_{t'}^{u,v,r}) \quad \forall (u,v,r) \in \mathcal{F}, t \in \mathcal{T} \tag{6.8}$$

Now consider the flow conservation at a transfer station. All arrival passengers at a transfer station of a path are the onboard passengers from the last leg. Therefore, we use a similar way to define the transfer flow conservation: the cumulative number of onboard passengers from the last leg should be larger than the cumulative number of boarding passengers at the transfer station. And the number of boarding passengers at the transfer station is simply $z_{t'}^{u,v,r,i}$ with $i \geq 2$. Hence, flow conservation constraints at a transfer station are:

$$\sum_{\{t':t^{\min}\leq t'+\Delta_{t'}^{u,v,r,i}\leq t\}} z_{t'}^{u,v,r,i} \leq \sum_{\{t':t^{\min}\leq t'+\delta_{t'}^{u,v,r,i-1}\leq t\}} z_{t'}^{u,v,r,i-1} \quad \forall (u,v,r)\in\mathcal{F}, i\in\mathcal{I}^{(u,v,r)}\setminus\{1\}, t\in\mathcal{T} \tag{6.9}$$

Note that $z_{t'}^{u,v,r,i}$ is defined as the onboard passengers for vehicles **departing** at time t'. Therefore, $t'+\delta_{t'}^{u,v,r,i-1}$ is the alighting time for passengers at leg $i-1$ (which is also the transfer demand arrival time at leg i as we assume transfer walk time is within a time interval τ and is negligible). $t'+\Delta_{t'}^{u,v,r,i}$ is the boarding time for passengers at leg i.

The objective is to minimize the total travel time for all passengers in the system. Total travel time can be decomposed into waiting time and in-vehicle time.

In-vehicle time: Total in-vehicle time is simply the onboard flow multiplied by the travel time on each leg:

$$IVT(\boldsymbol{z}) = \sum_{(u,v,r)\in\mathcal{F}} \sum_{i\in\mathcal{I}^{u,v,r}} \sum_{t\in\mathcal{T}} z_t^{u,v,r,i} \cdot T_{u,v,r,i,t}^{\text{IVT}} \tag{6.10}$$

where $T_{u,v,r,i,t}^{\text{IVT}}$ is the in-vehicle time of leg (u,v,r,i) of the vehicle departing at time t.

Waiting time: There are two causes of waiting time: 1) waiting time because of vehicle headways, and 2) waiting time resulting from being left behind. During a specific time interval t, all left behind passengers would have a waiting time of τ. All boarding passengers, assuming uniform arrival, have an average waiting time that is half of the time interval (i.e., $\frac{\tau}{2}$). Therefore, the total waiting time for passengers at station s and time t can be formulated as

$$WT_{s,t} = \tau(AD_{s,t} + XD_{s,t} - BD_{s,t}) + \frac{\tau}{2}(BD_{s,t+1} - BD_{s,t}) \tag{6.11}$$

where $AD_{s,t}$ represents the **cumulative arriving demand** at station s **up to** time t, $XD_{s,t}$ represents the **cumulative transferring demand** at station s **up to** time t, and $BD_{s,t}$ represents the **cumulative boarded demand** at station s **up to** time

t. Hence, $(BD_{s,t+1} - BD_{s,t})$ represents the total number of boarding passengers at time t and station s, and $(AD_{s,t} + XD_{s,t} - BD_{s,t})$ represents the total number of left behind passengers at station s and time t. Finally, the total system waiting time is

$$WT(\boldsymbol{q}, \boldsymbol{z}) = \sum_{s \in \mathcal{S}} \sum_{t=1}^{T} WT_{s,t} \tag{6.12}$$

The cumulative arriving demand $AD_{s,t}$ is simply all arriving passengers with origin s up to time t:

$$AD_{s,t} = \sum_{\{(u,v,r):u=s\}} \sum_{t'=t^{\min}}^{t} (f_{t'}^{u,v,r} + q_{t'}^{u,v,r}) \quad \forall s \in \mathcal{S}, t \in \mathcal{T} \tag{6.13}$$

where $\mathcal{S}$ is the set of all stations.

The cumulative transferring demand is all passengers alighting at station s from their previous leg $i-1$ for their next leg i:

$$XD_{s,t} = \sum_{\{(u,v,r,i) \in \texttt{Xth}(s)\}} \sum_{\{t':t^{\min} \leq t' + \delta_{t'}^{u,v,r,i-1} \leq t\}} z_{t'}^{u,v,r,i-1} \quad \forall t = t^{\min}, ..., T \tag{6.14}$$

where $\texttt{Xth}(s)$ is the set of legs that transfer at station s.

The cumulative boarded demand is all passengers that successfully board a vehicle at station s at time t. Define $\texttt{Bdat}(s)$ as the set of all legs with boarding station s, we have

$$BD_{s,t} = \sum_{\{(u,v,r,i) \in \texttt{Bdat}(s)\}} \sum_{\{t':t^{\min} \leq t' + \Delta_{t'}^{u,v,r,i} \leq t\}} z_{t'}^{u,v,r,i} \quad \forall t = t^{\min}, ..., T \tag{6.15}$$

Taking everything into consideration, the total travel time in the system is $WT(\boldsymbol{x}, \boldsymbol{z})+$

$IVT(\boldsymbol{z})$. The optimal flow problem is:

$$(OF) \quad \min_{\boldsymbol{q},\boldsymbol{z}} \quad WT(\boldsymbol{q},\boldsymbol{z}) + IVT(\boldsymbol{z}) \tag{6.16a}$$

$$\text{s.t.} \quad O_{l,t,t'}(\boldsymbol{z}) \leq K_{l,t} \quad \forall l \in \mathcal{L}, t \in \mathcal{T}, t' = t, t+1, ..., T_{l,t} \tag{6.16b}$$

$$\sum_{\{t': t^{\min} \leq t' + \Delta_{t'}^{u,v,r,1} \leq t\}} z_{t'}^{u,v,r,1} \leq \sum_{t'=t^{\min}}^{t} (f_{t'}^{u,v,r} + q_{t'}^{u,v,r}) \quad \forall (u,v,r) \in \mathcal{F}, t \in \mathcal{T} \tag{6.16c}$$

$$\sum_{\{t': t^{\min} \leq t' + \Delta_{t'}^{u,v,r,i} \leq t\}} z_{t'}^{u,v,r,i} \leq \sum_{\{t': t^{\min} \leq t' + \delta_{t'}^{u,v,r,i-1} \leq t\}} z_{t'}^{u,v,r,i-1} \quad \forall (u,v,r) \in \mathcal{F}, i \in \mathcal{I}^{(u,v,r)} \setminus \{1\}, t \in \mathcal{T} \tag{6.16d}$$

$$\sum_{r \in \mathcal{R}^{u,v}} q_t^{u,v,r} + f_t^{u,v,r} = d_t^{u,v} \quad \forall (u,v) \in \mathcal{W}, t \in \mathcal{T} \tag{6.16e}$$

$$z_t^{u,v,r,i} = \hat{z}_t^{u,v,r,i} \quad \forall (u,v,r,i,t) \in \Omega_1 \tag{6.16f}$$

$$z_t^{u,v,r,i} \geq 0 \quad \forall t \in \mathcal{T}, (u,v,r) \in \mathcal{F}, i \in \mathcal{I}^{u,v,r} \tag{6.16g}$$

$$q_t^{u,v,r} \geq 0 \quad \forall t \in \mathcal{T}, (u,v,r) \in \mathcal{F}, \tag{6.16h}$$

As the objective function is minimizing the system travel time, this formulation will automatically load passengers to a train as long as there is available capacity [184].

Path travel time calculation: It is worth noting that Eq. 6.16 does not explicitly output the travel time of different paths. The travel time of a path (u,v,r) for trips departs at time t (denoted as $TT_t^{u,v,r}$) has to be obtained from the network flow patterns **after** solving Eq. 6.16. Specifically, consider the group of passengers using path (u,v,r) and departing at time t. Their arrival time at the destination (denoted as $AT_t^{u,v,r}$) can be calculated as

$$AT_t^{u,v,r} = \min \left\{ \tilde{t} \in \mathcal{T}_t^{u,v,r} : \sum_{t'=t^{\min}}^{t} (f_{t'}^{u,v,r} + q_{t'}^{u,v,r}) \leq \sum_{t^{\min} \leq t' + \delta_{t'}^{u,v,r,|\mathcal{I}^{u,v,r}|} \leq \tilde{t}} z_{t'}^{u,v,r,|\mathcal{I}^{u,v,r}|} \right\} \quad \forall t \in \mathcal{T}, (u,v,r) \in \mathcal{F} \tag{6.17}$$

where $\mathcal{T}_t^{u,v,r}$ is the set of possible arrival time indices, defined as $\mathcal{T}_t^{u,v,r} = \{t' : t \leq t' \leq T\}$. Eq. 6.17 represents the travel time calculation with cumulative demand curves at origins and destinations. $\sum_{t'=t^{\min}}^{t} (f_{t'}^{u,v,r} + q_{t'}^{u,v,r})$ is the cumulative demand up to time t at the origin. $\sum_{t^{\min} \leq t' + \delta_{t'}^{u,v,r,|\mathcal{I}^{u,v,r}|} \leq \tilde{t}} z_{t'}^{u,v,r,|\mathcal{I}^{u,v,r}|}$ is the cumulative passengers arriving at the destination up to time t'. When the cumulative arrivals at the destination are greater or equal to the cumulative demand at the origin (up to time t), all passengers finish the trip. So taking the minimum over t' gives the arrival time for passengers departing at t. The path travel time is then simply:

$$TT_t^{u,v,r} = AT_t^{u,v,r} - t \quad \forall t \in \mathcal{T}, (u,v,r) \in \mathcal{F} \tag{6.18}$$

Figure 6-4 illustrates the travel time calculation.

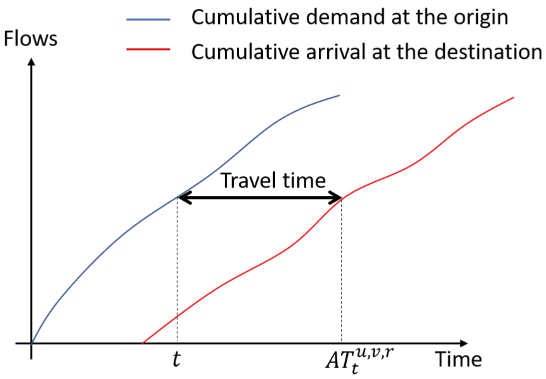


Figure 6-4: Travel time calculation

Incident specification: Eq. 6.16 is a general formulation of the optimal flow problem. Now we will introduce how the incident-specific information is incorporated into this problem. We assume the incident causes a service disruption in a specific line (if only several stations are interrupted, we can separate the line into multiple lines so that the assumption always holds). The service disruption in a line can be seen as stops of vehicles for a period of time. The vehicle stopping can be captured by the parameters $\Delta_t^{u,v,r,i}$, $\delta_t^{u,v,r,i}$, and $K_{l,t}$. Specifically, a long stop due to an incident can be seen as an increase in travel time from the terminal to downstream stations (i.e., increase in $\Delta_t^{u,v,r,i}$ and $\delta_t^{u,v,r,i}$). Moreover, since there is no vehicle dispatching during the incident, we set $K_{l,t} = 0$ for the corresponding time and line. In this way, we can model the incident without changing the formulation.

6.4.2 Behavior uncertainty

Consider a passenger p with a path set $\mathcal{R}_p$. Their inherent preference (utility) of using path r is denoted as V_p^r. If path r' was recommended, the impact of the recommendation on the utility of path r is denoted as $I_{p,r'}^r$. Hence, his/her overall utility of using path r can be represented as

$$U_p^r = V_p^r + \sum_{r' \in \mathcal{R}_p} x_{p,r'} \cdot I_{p,r'}^r + \xi_p^r \quad \forall r \in \mathcal{R}_p,\ p \in \mathcal{P}. \tag{6.19}$$

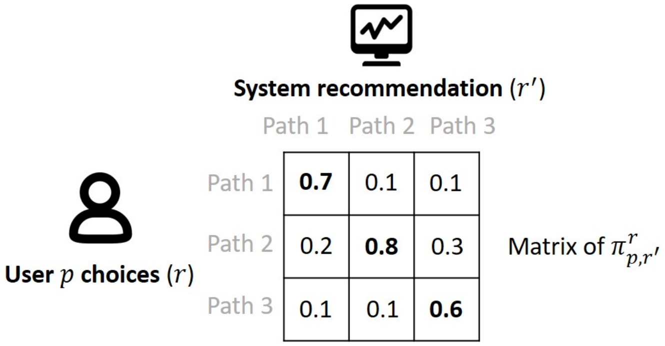
where ξ_p^r is the random error. $x_{p,r'} = 1$ if passenger p is recommended path r', otherwise $x_{p,r'} = 0$. Let $\pi_{p,r'}^r$ be the conditional probability that passenger p chooses path r given that the recommended path is r'. Assuming a utility maximizing behavior, we have

$$\pi_{p,r'}^r = \mathbb{P}(V_p^r + I_{p,r'}^r + \xi_p^r \geq V_p^{r''} + I_{p,r'}^{r''} + \xi_p^{r''},\ \forall r'' \in \mathcal{R}_p) \tag{6.20}$$

Different assumptions for the distribution of ξ_p^r can lead to different expression. For example, if ξ_p^r are i.i.d. Gumbel distributed, the choice probability reduces to multinomial logit model [190] and we have

$$\pi_{p,r'}^r = \frac{\exp(V_p^r + I_{p,r'}^r)}{\sum_{r'' \in \mathcal{R}_p} \exp(V_p^{r''} + I_{p,r'}^{r''})} \tag{6.21}$$

The value of V_p^r and $I_{p,r'}^r$ can be calibrated using data from individual-level survey or smart card, which deserves separate research. When developing the individual path recommendation model, we assume $\pi_{p,r'}^r$ is known. Figure 6-5 shows an example for the conditional probability matrix. The specific values assume that paths with recommendations are more likely to be chosen.



	Path 1	Path 2	Path 3
Path 1	**0.7**	0.1	0.1
Path 2	0.2	**0.8**	0.3
Path 3	0.1	0.1	**0.6**

Figure 6-5: Example of conditional path choice probability

The conditional probability $\pi^r_{p,r'}$ captures the individual's inherent preference for different paths as well as the response to the recommendation system. It varies across individuals and reflects their behavioral uncertainties. This study focuses on design path recommendation systems based on the value of $\pi^r_{p,r'}$.

6.4.3 Individual path recommendation

Let $\mathbb{1}^r_{p,r'}$ be the indicator random variable representing whether passenger p actually chooses path r or not given that he/she is recommended path r'. By definition, $\mathbb{1}^r_{p,r'}$ is a Bernoulli random variable with $\mathbb{E}[\mathbb{1}^r_{p,r'}] = \pi^r_{p,r'}$ and $\text{Var}[\mathbb{1}^r_{p,r'}] = \pi^r_{p,r'} \cdot (1 - \pi^r_{p,r'})$

Therefore, the actual flow for path (u, v, r) at time t is

$$Q_t^{u,v,r} = \sum_{p\in\mathcal{P}_t^{u,v}} \sum_{r'\in\mathcal{R}^{u,v}} x_{p,r'} \cdot \mathbb{1}^r_{p,r'} \tag{6.22}$$

$Q_t^{u,v,r}$ is also a random variable. $\mathcal{P}_t^{u,v} \subseteq \mathcal{P}$ is the set of passengers with OD pair (u, v) arriving at the system at time interval t that receive path recommendations. $\mathcal{R}^{u,v}$ is the set of paths of OD pair (u, v). The mean and variance of the actual flow is

$$\mu_t^{u,v,r}(\boldsymbol{x}) := \mathbb{E}\left[Q_t^{u,v,r}\right] = \sum_{p\in\mathcal{P}_t^{u,v}} \sum_{r'\in\mathcal{R}^{u,v}} x_{p,r'} \cdot \pi^r_{p,r'} \tag{6.23}$$

$$(\sigma_t^{u,v,r}(\boldsymbol{x}))^2 := \text{Var}\left[Q_t^{u,v,r}\right] = \sum_{p\in\mathcal{P}_{t'}^{u,v}} \sum_{r'\in\mathcal{R}^{u,v}} x_{p,r'} \cdot \pi^r_{p,r'} \cdot (1 - \pi^r_{p,r'}) \tag{6.24}$$

Note that Eqs. 6.24 is based on the fact that $x^2_{p,r'} = x_{p,r'}$ and $\text{Cov}[\mathbb{1}^r_{p,r'}, \mathbb{1}^r_{p,r''}] = 0$ if

$r' \neq r''$.

In an optimization model, we cannot use a random variable (e.g., actual flow) as the decision variable. Therefore, let us treat $\boldsymbol{q}$ in Eq. 6.16 as a **realization** of the random variable flow $\boldsymbol{Q}$. To make $\boldsymbol{q}$ a reasonable realization, some constraints need to be considered between the value of $\boldsymbol{q}$ and the distribution of $\boldsymbol{Q}$. We define two new concepts: "ϵ-feasibility" and "Γ-concentratration".

Definition 1 (**ϵ-feasible flows**)**.** A flow $q_t^{u,v,r}$ is ϵ-feasible if and only if

$$|q_t^{u,v,r} - \mu_t^{u,v,r}(\boldsymbol{x})| \leq \epsilon_t^{u,v,r}, \quad \forall (u,v,r) \in \mathcal{F}, t = t^{\min}, ..., T \tag{6.25}$$

where $\epsilon_t^{u,v,r}$ is a small positive constant. This means that $\boldsymbol{q}$ is close to the expectation of the actual flow under recommendation strategy $\boldsymbol{x}$.

Definition 2 (**Γ-concentrated flows**)**.** A flow $q_t^{u,v,r}$ is Γ-concentrated if and only if it is ϵ-feasible and for any constant $a > \epsilon_t^{u,v,r}$, we have

$$\mathbb{P}\left[|Q_t^{u,v,r} - q_t^{u,v,r}| \geq a\right] \leq \left(\frac{\Gamma_t^{u,v,r}}{a - \epsilon_t^{u,v,r}}\right)^2 \quad \forall (u,v,r) \in \mathcal{F}, t = t^{\min}, ..., T \tag{6.26}$$

where $\Gamma_t^{u,v,r}$ is a small positive constant. This means that the probability that $Q_t^{u,v,r}$ and $q_t^{u,v,r}$ are very different (i.e., with difference greater than a) is bounded above, suggesting that $Q_t^{u,v,r}$ is concentrated around $q_t^{u,v,r}$.

Remark 7. The logic of using $\boldsymbol{q}$ as the decision variable and defining the above two concepts is as follows. The objective of this study is to find the best recommendation strategy $\boldsymbol{x}$ that minimizes the system travel time. The system travel time is a function of network flows. Given a recommendation strategy $\boldsymbol{x}$, the actual flow $\boldsymbol{Q}$ is a random variable, which cannot be directly used in the optimization model (as decision variable) to evaluate the system travel time. Hence, we assume that $\boldsymbol{q}$ in Eq. 6.16 is a realization of the actual flow (deterministic variable). We also add two constraints to $\boldsymbol{q}$ so that $\boldsymbol{q}$ is close to the mean of the actual flow, and the distribution of the actual flow is concentrated around $\boldsymbol{q}$. Then, using $\boldsymbol{q}$ to evaluate the system travel time is similar to that of using the actual flows.

Note that one may argue that we can directly use $\mu_t^{u,v,r}(\boldsymbol{x})$ as decision variables to represent network flows and eliminate $\boldsymbol{q}$. This idea is essentially equivalent to setting $\epsilon_t^{u,v,r} = 0$ and does not consider the concentration property (i.e., $\Gamma_t^{u,v,r} = +\infty$), which is a special case of our framework. Our framework has more advantages in controlling the variance. Specifically, ignoring the Γ-concentration may make the model recommendation strategies meaningless. Consider an extreme scenario that there is a recommendation strategy $\boldsymbol{x}$, under which the actual flow is uniformly distributed in $[0, 1]$. Further assume that the system travel time is simply a linear function of the actual flow, say the factor is 1 (i.e., the system travel time is also uniformly distributed in $[0, 1]$). Suppose that the recommendation strategy $\boldsymbol{x}$ minimizes the expected system travel time (now the system travel time is $1 \times \mu_t^{u,v,r}(\boldsymbol{x}) = 0.5$). However, as the actual flow can be any value between 0 and 1 with equal probability, we know that the actual system travel time can also be any value between 0 and 1. Hence, the recommendation strategy $\boldsymbol{x}$, though minimizing the expected system travel time, is meaningless because there are too many variations for the actual system travel time under this recommendation. Γ-concentration is an important property to ensure that the distribution of actual flows is not too dispersed[3] so that the recommendation strategy $\boldsymbol{x}$ is solved based on a reliable estimate of the system travel time.

We will therefore incorporate ϵ-feasibility and Γ-concentration as constraints into the optimization formulation (Eq. 6.16). It turns out that both of them can be modeled as linear constraints. ϵ-feasibility (Eq. 6.25) can be easily transformed into a linear constraint by eliminating the absolute value. To incorporate Γ-concentration (Eq. 6.26), the following Proposition is used:

Proposition 14. *The Γ-concentration inequality (Eq. 6.26) holds if the variance of $Q_t^{u,v,r}$ is bounded above by $(\Gamma_t^{u,v,r})^2$. Mathematically:*

$$\sum_{p \in \mathcal{P}_{t'}^{u,v}} \sum_{r' \in \mathcal{R}^{u,v}} x_{p,r'} \cdot \pi_{p,r'}^{r} \cdot (1 - \pi_{p,r'}^{r}) \leq (\Gamma_t^{u,v,r})^2 \tag{6.27}$$

[3] In reality, as the actual flow is the summation of many Bernoulli random variables, the coefficient of variation will shrink with the increase in passenger size. So in the case of a large number of passengers, the Γ-concentration should be naturally satisfied

Proof. From the triangular inequality, we have:

$$\begin{aligned}\underbrace{|Q_t^{u,v,r} - q_t^{u,v,r}|}_{\text{LHS}} &\leq |Q_t^{u,v,r} - \mu_t^{u,v,r}(\boldsymbol{x})| + |\mu_t^{u,v,r}(\boldsymbol{x}) - q_t^{u,v,r}| \\ &\leq \underbrace{|Q_t^{u,v,r} - \mu_t^{u,v,r}(\boldsymbol{x})| + \epsilon_t^{u,v,r}}_{\text{RHS}} \qquad (6.28)\end{aligned}$$

As LHS $\leq$ RHS, the probability measure satisfies (for all $a > \epsilon_t^{u,v,r}$):

$$\mathbb{P}[\text{LHS} \geq a] \leq \mathbb{P}[\text{RHS} \geq a] \qquad (6.29)$$

Notice that

$$\mathbb{P}[\text{RHS} \geq a] = \mathbb{P}\left[|Q_t^{u,v,r} - \mu_t^{u,v,r}(\boldsymbol{x})| \geq a - \epsilon_t^{u,v,r}\right] \leq \frac{(\sigma_t^{u,v,r}(\boldsymbol{x}))^2}{(a - \epsilon_t^{u,v,r})^2} \qquad (6.30)$$

Eq. 6.30 is based on Chebyshev's inequality. Therefore,

$$\mathbb{P}[\text{LHS} \geq a] = \mathbb{P}[|Q_t^{u,v,r} - q_t^{u,v,r}| \geq a] \leq \frac{(\sigma_t^{u,v,r}(\boldsymbol{x}))^2}{(a - \epsilon_t^{u,v,r})^2} \qquad (6.31)$$

Comparing Eqs. 6.31 and 6.26, we know that to satisfy Eq. 6.26, we only need $\sigma_t^{u,v,r}(\boldsymbol{x}) \leq \Gamma_t^{u,v,r}$, which completes the proof. □

For modeling convenience, we set $\epsilon_t^{u,v,r} = \epsilon \cdot \mu_t^{u,v,r}(\boldsymbol{x})$ and $\Gamma_t^{u,v,r} = \Gamma \cdot d_t^{u,v}$, where ϵ and Γ are hyper-parameters determining how close and concentrated the actual flow should be. Then the final constraint becomes:

$$(1-\epsilon) \sum_{p \in \mathcal{P}_t^{u,v}} \sum_{r' \in \mathcal{R}^{u,v}} x_{p,r'} \cdot \pi_{p,r'}^r \leq q_t^{u,v,r} \leq (1+\epsilon) \sum_{p \in \mathcal{P}_t^{u,v}} \sum_{r' \in \mathcal{R}^{u,v}} x_{p,r'} \cdot \pi_{p,r'}^r \qquad (6.32)$$

and

$$\sum_{p \in \mathcal{P}_{t'}^{u,v}} \sum_{r' \in \mathcal{R}^{u,v}} x_{p,r'} \cdot \pi_{p,r'}^r \cdot (1 - \pi_{p,r'}^r) \leq (\Gamma \cdot d_t^{u,v})^2 \qquad (6.33)$$

Both constraints are linear and can be added into Eq. 6.16.

Besides the total system travel time, many recommendation systems also aim to respect passenger's preferences. That is, if possible, a path with high inherent utility V_p^r should be recommended. Hence the following term is added into the objective function.

$$\max \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}_p} x_{p,r} \cdot V_p^r \Longleftrightarrow \min \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}_p} -x_{p,r} \cdot V_p^r \tag{6.34}$$

The final individual path recommendation (IPR) model can be formulated as:

$$(IPR) \quad \min_{\boldsymbol{x},\boldsymbol{q},\boldsymbol{z}} \quad WT(\boldsymbol{q},\boldsymbol{z}) + IVT(\boldsymbol{z}) + \Psi \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}_p} -x_{p,r} \cdot V_{p,r} \tag{6.35a}$$

$$\text{s.t.} \quad \text{Constraints } (6.16b) - (6.16h) \tag{6.35b}$$

$$(1-\epsilon) \sum_{p \in \mathcal{P}_t^{u,v}} \sum_{r' \in \mathcal{R}^{u,v}} x_{p,r'} \cdot \pi_{p,r'}^r \leq q_t^{u,v,r} \leq (1+\epsilon) \sum_{p \in \mathcal{P}_t^{u,v}} \sum_{r' \in \mathcal{R}^{u,v}} x_{p,r'} \cdot \pi_{p,r'}^r$$

$$\forall t \in \mathcal{T}, (u,v,r) \in \mathcal{F} \tag{6.35c}$$

$$\sum_{p \in \mathcal{P}_{t'}^{u,v}} \sum_{r' \in \mathcal{R}^{u,v}} x_{p,r'} \cdot \pi_{p,r'}^r \cdot (1 - \pi_{p,r'}^r) \leq (\Gamma \cdot d_t^{u,v})^2 \quad \forall t \in \mathcal{T}, (u,v,r) \in \mathcal{F}$$

$$\tag{6.35d}$$

$$\sum_{r \in R_p} x_{p,r} = 1 \quad \forall p \in \mathcal{P} \tag{6.35e}$$

$$x_{p,r} \in \{0,1\} \quad \forall p \in \mathcal{P}, r \in \mathcal{R}_p \tag{6.35f}$$

where Ψ is a hyper-parameter to adjust the scale and balance the trade-off between system efficiency and passenger preferences.

6.4.4 Benders decomposition

Eq. 6.35 is a mixed-integer linear programming (MILP). The structure of Eq. 6.35 allows us to efficiently solve it by Benders decomposition (BD) [214]. The basic idea of BD is to decompose the problem into a master problem and a subproblem and solve these problems iteratively. The decision variables are divided into difficult variables, which in our case are the binary variables $\boldsymbol{x}$, and a set of easier variables, the continuous $\boldsymbol{q}$ and $\boldsymbol{z}$. At each iteration, the master problem determines one possible leader decision $\boldsymbol{x}$. This solution is used in the subproblem to generate optimality-cuts or feasibility-cuts, which are added to the master problem.

Interestingly, in this study, the master problem decides the recommendation strategies, which is a MILP of a smaller scale and can be solved efficiently using existing solvers. The subproblem reduces to the optimal flow problem (Eq. 6.16) with one more linear constraint (still linear programming). This format makes the BD an appropriate algorithm for the original problem.

Subproblem

The subproblem is derived by fixing the decision variables $\boldsymbol{x}$, and only considering the components including $\boldsymbol{q}$ and $\boldsymbol{z}$.

$$
\begin{aligned}
[SP(\boldsymbol{x})] \quad \min_{\boldsymbol{q},\boldsymbol{z}} \quad & WT(\boldsymbol{q},\boldsymbol{z}) + IVT(\boldsymbol{z}) && (6.36a) \\
\text{s.t.} \quad & \text{Constraints } (6.16b) - (6.16h) && (6.36b) \\
& \text{Constraint } (6.35c) && (6.36c)
\end{aligned}
$$

The objective of the dual problem of Eq. 6.36 is

$$
\begin{aligned}
D(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\eta},\boldsymbol{\kappa},\boldsymbol{\rho};\boldsymbol{x}) = & \sum_{l\in\mathcal{L}}\sum_{t\in\mathcal{T}}\sum_{t'=t}^{T_{l,t}} K_{l,t}\alpha_{l,t,t'} + \sum_{(u,v,r)\in\mathcal{F}}\sum_{t\in\mathcal{T}}\sum_{t'=t^{\min}}^{t} f_{t'}^{u,v,r}\beta_t^{u,v,r} \\
& + \sum_{(u,v,r,i,t)\in\Omega_1} \hat{z}_t^{u,v,r,i}\gamma_t^{u,v,r,i} + \sum_{(u,v)\in\mathcal{W}}\sum_{t\in\mathcal{T}} d_t^{u,v}\eta_t^{u,v} \\
& + \sum_{(u,v,r)\in\mathcal{F}}\sum_{t\in\mathcal{T}} \kappa_t^{u,v,r}\cdot(1-\epsilon)\sum_{p\in\mathcal{P}_t^{u,v}}\sum_{r'\in\mathcal{R}^{u,v}} x_{p,r'}\cdot\pi_{p,r'}^{r} \\
& + \sum_{(u,v,r)\in\mathcal{F}}\sum_{t\in\mathcal{T}} \rho_t^{u,v,r}\cdot(1+\epsilon)\sum_{p\in\mathcal{P}_t^{u,v}}\sum_{r'\in\mathcal{R}^{u,v}} x_{p,r'}\cdot\pi_{p,r'}^{r} \qquad (6.37)
\end{aligned}
$$

where $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\eta}$ are the dual variables associated with constraints 6.16b, 6.16c, 6.16f, 6.16e, respectively. $\boldsymbol{\kappa}$, $\boldsymbol{\rho}$ are the dual variables associated with constraint 6.35c. Let $\boldsymbol{\Theta} := (\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\eta},\boldsymbol{\kappa},\ \boldsymbol{\rho})$. If the dual problem of Eq. 6.36 is feasible and bounded with a solution $\mathbf{\Theta}^*$, the following optimality cut is added to the master problem:

$$Y \geq D(\mathbf{\Theta}^*; \boldsymbol{x}) \tag{6.38}$$

where Y is a decision variable in the master problem. If the dual problem of Eq. 6.36 is unbounded, and $\mathbf{\Theta}^*$ is an optimal extreme ray of the dual, the following feasibility cut is added to the master problem:

$$D(\mathbf{\Theta}^*; \boldsymbol{x}) \leq 0 \tag{6.39}$$

Master problem

Let $\mathcal{A}^{\mathrm{O}}$ be the set of solutions $\mathbf{\Theta}^*$ of optimality cuts and $\mathcal{A}^{\mathrm{F}}$ be the set of solutions $\mathbf{\Theta}^*$ of feasibility cuts. At each iteration of the BD, a cut based on the solution of the subproblem is added to the respective set, and the corresponding master problem is defined as follows:

$$\begin{aligned}
[MP(\mathcal{A}^{\mathrm{O}}, \mathcal{A}^{\mathrm{F}})] \quad \min_{\boldsymbol{x}, Y} \quad & \Psi \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}_p} -x_{p,r} \cdot V_{p,r} + Y && \text{(6.40a)} \\
\text{s.t.} \quad & Y \geq D(\mathbf{\Theta}^*; \boldsymbol{x}) \quad \forall \mathbf{\Theta}^* \in \mathcal{A}^{\mathrm{O}} && \text{(6.40b)} \\
& D(\mathbf{\Theta}^*; \boldsymbol{x}) \leq 0 \quad \forall \mathbf{\Theta}^* \in \mathcal{A}^{\mathrm{F}} && \text{(6.40c)} \\
& \text{Constraints } (6.35d) - (6.35f) && \text{(6.40d)}
\end{aligned}$$

Note that the master problem has a smaller scale compared to the original problem (because there are no $\boldsymbol{z}$ and $\boldsymbol{q}$), which can be solved efficiently.

Convergence

Let $(\boldsymbol{x}^{(k)}, Y^{(k)})$ and $(\boldsymbol{q}^{(k)}, \boldsymbol{z}^{(k)})$ be the solutions of the master problem and subproblem, respectively, in the k-th iteration. Then, the upper $(UB^{(k)})$ and lower $(LB^{(k)})$ bounds at the k-th iteration are given by:

$$UB^{(k)} = \Psi \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}_p} -x_{p,r}^{(k)} \cdot V_{p,r} + WT(\boldsymbol{q}^{(k)}, \boldsymbol{z}^{(k)}) + IVT(\boldsymbol{z}^{(k)}) \tag{6.41}$$

$$LB^{(k)} = \Psi \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}_p} -x_{p,r}^{(k)} \cdot V_{p,r} + Y^{(k)} \tag{6.42}$$

$LB^{(k)}$ will keep increasing as k increases because more cuts are added into the master problem. $UB^{(k)}$ does not necessarily decrease at every iteration. The convergence criterion is

$$\text{Gap}^{(k)} = \frac{UB^{(k)} - LB^{(k)}}{LB^{(k)}} \leq \text{Predetermined threshold} \tag{6.43}$$

6.4.5 Post-adjustment for equity

In this study, we define the equity requirement of a recommendation strategy as *passengers with the same OD (u, v) and departure time t should not have too much difference in travel time if they follow the recommendations.* Mathematically:

$$w_p := \sum_{r \in \mathcal{R}_p} TT_{p,r} \cdot x_{p,r} - \min_{p' \in \mathcal{P}_t^{u,v}} \{ \sum_{r \in \mathcal{R}_{p'}} TT_{p',r} \cdot x_{p',r} \} \leq E \cdot \tau \quad \forall p \in \mathcal{P} \tag{6.44}$$

where $TT_{p,r}$ is the travel time of path $r \in \mathcal{R}_p$ for passenger p, which can be obtained from Eq. 6.18. $\sum_{r \in \mathcal{R}_p} TT_{p,r} \cdot x_{p,r}$ is the travel time on the recommended path. w_p represents the difference between passenger p' travel time (if he/she follows the recommendation) and the minimum travel time of all passengers with the same (u, v, t). $E \cdot \tau$ is a predetermined threshold for the travel time difference (where $E \in \mathbb{N}$).

As mentioned in Section 6.4.1, we cannot formulate path travel time in the optimal flow model. Hence, Eq. 6.44 cannot be added into Eq. 6.35 as a constraint directly. In this section, we propose a post-adjustment heuristic method to address the equity constraint. The basic idea of the post-adjustment is to change the recommendations of high-cost paths to low-cost ones after solving Eq. 6.35. Notice that if $E = 0$ in Eq. 6.44, we obtain a user equilibrium recommendation pattern. This motivates us to direct the current system optimal recommendations toward user equilibrium recommendations. The user equilibrium solution is usually obtained by the method of successive averages (MSA):

$$(q_t^{u,v,r})^{(k+1)} = (1 - \lambda_k) \cdot (q_t^{u,v,r})^{(k)} + \lambda_k \cdot (\tilde{q}_t^{u,v,r})^{(k)} \tag{6.45}$$

where $(\tilde{q}_t^{u,v,r})^{(k)}$ is the flow of all-or-nothing assignment to the shortest paths at iteration k. $(q_t^{u,v,r})^{(k)}$ is the current flow at iteration k, and λ_k is the step size at iteration k. Eq. 6.45 means at each iteration, some of the flows are moved to the shortest path so that we expect that ultimately all paths have similar travel time.

However, directly using MSA may violate the ϵ-feasibility and Γ-concentration because Eq. 6.45 does not guarantee that $(q_t^{u,v,r})^{(k+1)}$ satisfies these two constraints. Therefore, we propose a new MILP formulation that works similar to the MSA-based flow redistribution but guarantees ϵ-feasibility and Γ-concentration at each iteration.

Note that the post-adjustment is initialized at $(q_t^{u,v,r})^{(0)} = (q_t^{u,v,r})^*$, where $(q_t^{u,v,r})^*$ is the optimal flow solution from Eq. 6.35. By definition, $(q_t^{u,v,r})^{(0)}$ satisfies ϵ-feasibility and Γ-concentration. Now we need to develop a method that satisfies $(q_t^{u,v,r})^{(k)}$ is ϵ-feasible and Γ-concentrated for all $k \geq 1$.

Let the w_p at iteration k be:

$$w_p^{(k)} := \sum_{r \in \mathcal{R}_p} TT_{p,r}^{(k)} \cdot x_{p,r}^{(k)} - \min_{p' \in \mathcal{P}_t^{u,v}} \{ \sum_{r \in \mathcal{R}_{p'}} TT_{p',r}^{(k)} \cdot x_{p',r}^{(k)} \} \tag{6.46}$$

where $w_p^{(k)}$ represents the "degree of unfairness" for passenger p at iteration k because the larger the value, the longer the travel time of his/her recommended path compared to the shortest one. Let $r_{p,\min}^{(k)} = \arg\min_{r \in \mathcal{R}_p} \{TT_{p,r}\}$ be the shortest path ID for passenger p at iteration k. We formulate the post-adjustment MILP at iteration k as:

$$(PA(\boldsymbol{x}^{(k)}, \boldsymbol{q}^{(k)})) \quad \max_{\boldsymbol{x}^{(k+1)}, \boldsymbol{q}^{(k+1)}} \quad \sum_{p \in \mathcal{P}} w_p^{(k)} \cdot \sum_{r \in \mathcal{R}_p} \mathbb{1}_{\{r = r_{p,\min}^{(k)}\}} \cdot (x_{p,r})^{(k+1)} \tag{6.47a}$$

$$\text{s.t.} \quad (q_t^{u,v,r})^{(k)} \leq (q_t^{u,v,r})^{(k+1)} \leq (1 - \lambda_k) \cdot (q_t^{u,v,r})^{(k)} + \lambda_k \cdot (\tilde{q}_t^{u,v,r})^{(k)} \quad \forall (u,v,r,t) \in \mathcal{U}_+^{(k)} \tag{6.47b}$$

$$(1 - \lambda_k) \cdot (q_t^{u,v,r})^{(k)} + \lambda_k \cdot (\tilde{q}_t^{u,v,r})^{(k)} \leq (q_t^{u,v,r})^{(k+1)} \leq (q_t^{u,v,r})^{(k)} \quad \forall (u,v,r,t) \in \mathcal{U}_-^{(k)} \tag{6.47c}$$

$$\sum_{r \in \mathcal{R}^{u,v}} (q_t^{u,v,r})^{(k+1)} = d_t^{u,v} \quad \forall (u,v,t) \in \mathcal{W} \tag{6.47d}$$

$$(q_t^{u,v,r})^{(k+1)} \geq 0 \quad \forall t \in \mathcal{T}, (u,v,r) \in \mathcal{F} \tag{6.47e}$$

$$(x_{p,r})^{(k+1)} = x_{p,r}^{(k)} \quad \forall p \in \mathcal{P}_{\text{Eq}}{}^{(k)} \tag{6.47f}$$

$$\text{Constraints } (6.35c) - (6.35f) \quad (\text{replacing } \boldsymbol{x}, \, \boldsymbol{q} \text{ with } \boldsymbol{x}^{(k+1)}, \boldsymbol{q}^{(k+1)}) \tag{6.47g}$$

Constraints 6.47b and 6.47c mean that, at each step, the new flows at iteration $k+1$ (i.e., $\boldsymbol{q}^{(k+1)}$) are between $\boldsymbol{q}^{(k)}$ and the values obtained by MSA, where $\mathcal{U}_+^{(k)}$ (resp. $\mathcal{U}_-^{(k)}$) is the set of (u,v,r,t) indices that make $(\tilde{q}_t^{u,v,r})^{(k)} > 0$ (resp. $(\tilde{q}_t^{u,v,r})^{(k)} = 0$). With constraints 6.47g, the solution $\boldsymbol{q}^{(k+1)}$ and $\boldsymbol{x}^{(k+1)}$ satisfy ϵ-feasibility and Γ-concentration.

The objective function is maximized (without constraints) if $x_{p,r}^{(k+1)} = 1$ for all $r = r_{p,\min}^{(k)}$, which means we should change as many of the recommendations as possible to the shortest path. However, 6.47b and 6.47c do not allow to change all recommendations to the shortest path. The unfairness weight $w_p^{(k)}$ suggests that the recommendations to the passengers with the largest degree of unfairness should be changed first.

$\mathcal{P}_{\text{Eq}}{}^{(k)}$ is the set of passengers whose recommendations already satisfy the equity requirement (i.e., $\mathcal{P}_{\text{Eq}}{}^{(k)} = \{p \in \mathcal{P} : w_p^{(k)} \leq E \cdot \tau\}$). Constraint 6.47f helps to fix the recommendations that already satisfy the equity requirement, which reduces the scale of the problem.

Remark 8. The post-adjustment problem (Eq. 6.47) aims to find a new recommendation strategy $\boldsymbol{x}^{(k+1)}$ and flow patterns $\boldsymbol{q}^{(k+1)}$ closer to the user equilibrium solutions (i.e., with more equity). And the solutions are guaranteed to satisfy ϵ-feasibility and Γ-concentration. The post-adjustment problem is always feasible because $(\boldsymbol{x}^{(k)}, \boldsymbol{q}^{(k)})$ is a feasible solution (associated with no flow and recommendation changes). And it can be solved efficiently using branch-and-cut algorithms supported by many off-the-shelf solvers because the size of the problem is much smaller given that many integer variables are fixed (due to constraint 6.47f). Moreover, we can use $(\boldsymbol{x}^{(k)}, \boldsymbol{q}^{(k)})$ as a warm-start to further accelerate the solution.

Eq. 6.47 presents the formulation at iteration k. The complete algorithm for the post-adjustment problem is shown in Algorithm 5. There are two stopping criteria: 1) $(\boldsymbol{x}^{(k)},\ \boldsymbol{q}^{(k)}) = (\boldsymbol{x}^{(k-1)},\ \boldsymbol{q}^{(k-1)})$ or 2) ${\mathcal{P}_{\text{Eq}}}^{(k)} = \mathcal{P}$. The first one means that we cannot find a new recommendation strategy and flow pattern that satisfies equity, ϵ-feasibility, and Γ-concentration requirements. And the second criterion means that all passenger's recommendations satisfy the equity requirement.

Algorithm 5 Solution procedure of the post-adjustment approach for equity

```
Initialize $(\boldsymbol{x}^{(0)}, \boldsymbol{q}^{(0)})$ as the optimal solution of Eq. 6.35.
Set the iteration counter $k = 0$.
do
    Solve the post-adjustment problem (Eq. 6.47) with $(\boldsymbol{x}^{(k)}, \boldsymbol{q}^{(k)})$ as inputs, and return $(\boldsymbol{x}^{(k+1)}, \boldsymbol{q}^{(k+1)})$
    $k = k + 1$
while $(\boldsymbol{x}^{(k)}, \boldsymbol{q}^{(k)}) \neq (\boldsymbol{x}^{(k-1)}, \boldsymbol{q}^{(k-1)})$ and ${\mathcal{P}_{\text{Eq}}}^{(k)} \neq \mathcal{P}$
return $(\boldsymbol{x}^{(k)}, \boldsymbol{q}^{(k)})$
```

Remark 9. Post-adjustment is a heuristic method to find a recommendation strategy that satisfies the equity requirements. Ideally, we also want the system travel time to be minimized under the equity constraint. But the proposed heuristic method cannot guarantee travel time optimality. With the updating by the post-adjustment procedure, the flow patterns are closer to the user equilibrium, and the system travel time will be increased. We may control the step size λ_k to make sure the new flows are not too different from the system optimal one, so that the new travel time is not too bad. However, a too small value of λ_k may limit the flow updating range and lead to a situation where not all passengers with $w_p \leq E \cdot \tau$ (i.e., Algorithm 5 stops with the first criterion). Hence, the step size λ_k plays a trade-off role between satisfying equity and controlling the increase in system travel time.

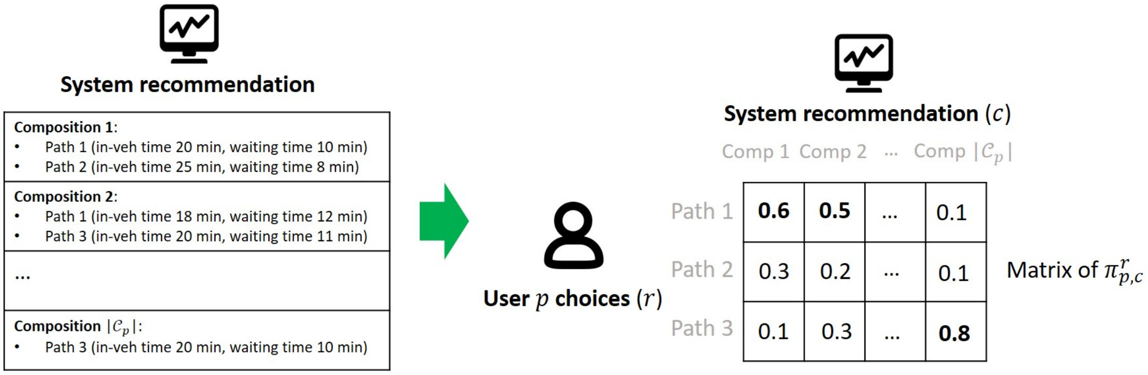
6.5 Model extension

In this section, we discuss several extensions of the model to accommodate more realistic/general scenarios.

6.5.1 Generalization of recommendations

In this study, we assume the information given to passengers is a recommended path. In reality, the recommendation system may provide a bundle of recommended paths with information like estimated in-vehicle time, waiting time, travel cost, etc. The proposed framework can be extended to handle different recommendation typologies. Figure 6-6 shows an example where the recommendation system will provide a composition of path and travel time information, where each composition can include different paths, different estimated waiting/in-vehicle times, etc. Then, we can change $x_{p,r}$ to $x_{p,c}$, where $x_{p,c}$ indicates whether we will present composition c to passenger p. Similarly, each c is associated with a conditional probability $\pi^r_{p,c}$ as shown in Figure 6-6 (the probability for passenger p to choose path r given that he/she is recommended composition c).

In this way, we only need to calibrate $\pi^r_{p,c}$ and predetermine the composition set $\mathcal{C}_p$ for each passenger p. The overall framework proposed above can be easily adapted to the new recommendation typology by replacing $x_{p,r}$ and $\pi^r_{p,r'}$ with $x_{p,c}$ and $\pi^r_{p,c}$, respectively.



Composition 1:
- Path 1 (in-veh time 20 min, waiting time 10 min)
- Path 2 (in-veh time 25 min, waiting time 8 min)

Composition 2:
- Path 1 (in-veh time 18 min, waiting time 12 min)
- Path 3 (in-veh time 20 min, waiting time 11 min)

…

Composition $|\mathcal{C}_p|$:
- Path 3 (in-veh time 20 min, waiting time 10 min)

	Comp 1	Comp 2	…	Comp $|\mathcal{C}_p|$
Path 1	**0.6**	**0.5**	…	0.1
Path 2	0.3	0.2	…	0.1
Path 3	0.1	0.3	…	**0.8**

Figure 6-6: Illustration of the generalized recommendation typology. $\mathcal{C}_p$ is the predetermined recommendation composition sets for passenger p

6.5.2 Feedback and rolling-horizon

As mentioned in Section 6.3.2, the whole path recommendation problem should be solved in a rolling-horizon manner. At each time interval $t \geq 1$, we update the demand, supply, and system state information, and solve the proposed framework above to get a recommendation strategy $\boldsymbol{x}$. But we only implement the $x_{p,r}$ for $p \in \mathcal{P}_t^{u,v}$, $\forall (u,v)$ (i.e., passengers departing at current time t).

The rolling horizon requires updating the estimated demand and system state information. The recommendation system can ask for passenger feedback to facilitate the estimation. For example, after providing a recommendation, we can ask the passenger to respond whether he/she will actually use it or not. This feedback can be used to update the demand predictions.

6.5.3 Integrated formulation with behavior uncertainty and equity

As mentioned in Section 6.4.5, the equity requirement (Eq. 6.44) cannot be incorporated into the recommendation model (Eq. 6.35) due to the fact that it is hard to formulate path travel time directly. In this section, we show a possible MILP formulation that incorporates the equity requirement. However, this formulation is far more complicated than Eq. 6.35 and is hard to solve, representing a future research activity.

Let $y_{\tilde{t}}^{u,v,r,t}$ represent a binary variable indicating whether the arrival time of a passenger departing at time t using (u, v, r) is $\tilde{t}$ or not. By definition,

$$\sum_{\tilde{t} \in \mathcal{T}_t^{u,v,r}} y_{\tilde{t}}^{u,v,r,t} = 1 \quad \forall t \in \mathcal{T}, (u, v, r) \in \mathcal{F} \tag{6.48}$$

$$y_{\tilde{t}}^{u,v,r,t} \in \{0, 1\} \quad \forall t \in \mathcal{T}, (u, v, r) \in \mathcal{F}, \tilde{t} \in \mathcal{T}_t^{u,v,r} \tag{6.49}$$

where $\mathcal{T}_t^{u,v,r}$ is the set of possible arrival time indices (see Eq. 6.17). From the relationship between arrival time and cumulative flows (Eq. 6.17), we have

$$\sum_{t'=t^{\min}}^{t} (f_{t'}^{u,v,r} + q_{t'}^{u,v,r}) \leq \sum_{t^{\min} \leq t' + \delta_{t'}^{u,v,r,|\mathcal{L}^{u,v,r}|} \leq \tilde{t}} z_{t'}^{u,v,r,|\mathcal{L}^{u,v,r}|} + (1 - y_{\tilde{t}}^{u,v,r,t})M$$
$$\forall t \in \mathcal{T}, (u, v, r) \in \mathcal{F}, \tilde{t} \in \mathcal{T}_t^{u,v,r} \tag{6.50}$$

where M is a big positive constant. When $y_{\tilde{t}}^{u,v,r,t} = 0$, $\tilde{t}$ is not the arrival time, and Eq. 6.50 is not binding. As the arrival time is the minimum $\tilde{t}$ that satisfies Eq. 6.50, the following component should be added into the objective function

$$\min \sum_{t \in \mathcal{T}} \sum_{(u,v,r) \in \mathcal{F}} \sum_{\tilde{t} \in \mathcal{T}_t^{u,v,r}} y_{\tilde{t}}^{u,v,r,t} \cdot \tilde{t} \tag{6.51}$$

Eqs 6.48 $\sim$ 6.51 provide a possible way to model path travel times by defining the binary variable $\boldsymbol{y}$. To incorporate the equity constraint, let the earliest arrival time over all paths of OD pair (u, v) and time t be $EAT_t^{u,v}$:

$$EAT_t^{u,v} = \min_{r \in \mathcal{R}^{u,v}} \{ \sum_{\tilde{t} \in \mathcal{T}_t^{u,v,r}} y_{\tilde{t}}^{u,v,r,t} \cdot \tilde{t} \} \quad \forall t \in \mathcal{T}, (u, v) \in \mathcal{W} \tag{6.52}$$

As the objective is minimizing $\sum_{\tilde{t} \in \mathcal{T}_t^{u,v,r}} y_{\tilde{t}}^{u,v,r,t}$, Eq. 6.52 can be transformed to a linear constraint:

$$EAT_t^{u,v} \leq \sum_{\tilde{t} \in \mathcal{T}_t^{u,v,r}} y_{\tilde{t}}^{u,v,r,t} \cdot \tilde{t} \quad \forall t \in \mathcal{T}, (u, v) \in \mathcal{W}, r \in \mathcal{R}^{u,v} \tag{6.53}$$

And the equity requirement can be formulated as

$$\sum_{\tilde{t}\in\mathcal{T}_t^{u,v,r}} y_{\tilde{t}}^{u,v,r,t}\cdot\tilde{t} - EAT_t^{u,v} \leq E\cdot\tau + (1-x_{p,r})\cdot M \quad \forall t\in\mathcal{T}, (u,v)\in\mathcal{W}, p\in\mathcal{P}_t^{u,v}, r\in\mathcal{R}^{u,v} \tag{6.54}$$

If $x_{p,r} = 0$ for all $p \in \mathcal{P}_t^{u,v}$, path r is not recommended and Eq. 6.54 becomes ineffective.

Adding all these constraints and the objective component into Eq. 6.35 results in the integrated formulation. However, it is extremely hard to solve due to a large number of integer variables $\boldsymbol{y}$ and the complicated constraint interactions. We show this formulation to demonstrate the difficulty of incorporating equity requirements and highlight the importance of the heuristic post-adjustment approach in Section 6.4.5.

6.6 Case study

6.6.1 Case study design

CTA Blue Line disruption

For the case study, we consider an actual incident in the Blue Line of the Chicago Transit Authority (CTA) urban rail system (Figure 6-7). The incident starts at 8:14 AM and ends at 9:13 AM on Feb 1st, 2019 due to infrastructure issues between Harlem and Jefferson Park stations (the red X in the figure) that led to a whole Blue Line suspension. During the disruption (morning hours), the destination for most of the passengers is the “Loop” in the CBD area in Chicago. There are four alternative paths to the Loop: 1) using the Blue Line (i.e., waiting for the system to recover), 2) using the parallel bus lines, 3) using the North-South (NS) bus lines to transfer to the Green Line, and 4) using the West-East (WE) bus lines to transfer to the Brown Line. Based on the service structure, the route sets $\mathcal{R}^{(u,v)}$ for each OD pair (u,v) can be constructed.

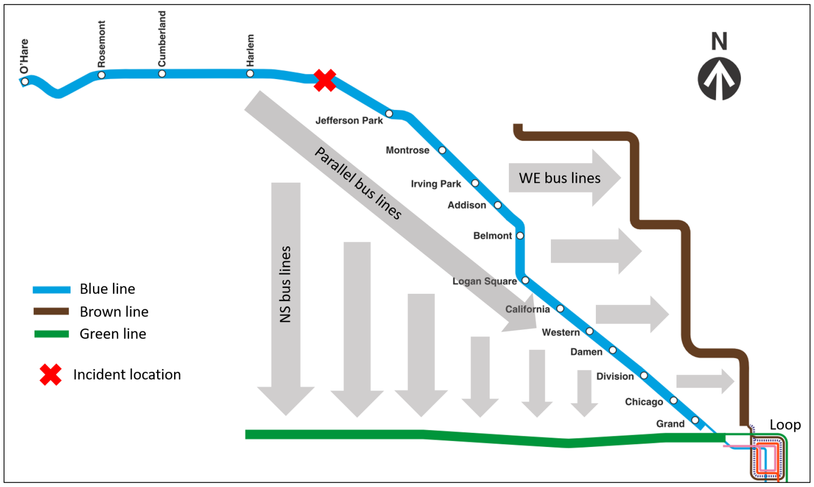


Figure 6-7: Case study network

In the case study, we divide the time into $\tau = 5$ mins equal-length intervals, and focus on solving the problem at $t = 1$ (i.e., beginning of the incident). We assume that the set of passengers to receive recommendations ($\mathcal{P}$) consists of all passengers with their intended origins at the Blue Line and destinations in the Loop. A simulation model [4] is used to get the system state up to time $t = 1$ (i.e., the incident time 8:14 AM) and generate $\hat{z}_t^{u,v,r,i}$ and Ω_1. The recommendation strategy covers passengers departing between $t = 1$ and $T^D = 23$, approximately one hour after the end of the incident (9:13 AM). The analysis period is set as $t^{\min} = -13$ and $T = 34$, approximately one hour before $t = 1$ and after T^D, providing enough buffer (warm-up and cool-down time) for passengers in $\mathcal{P}$ to finish their trips. As demand and incident duration predictions are out of the scope of this study, we simply use the actual demand and incident duration for all experiments. Our other work [45] proposes to use robust and stochastic optimization to address demand and incident duration uncertainty, respectively.

Conditional probability matrix $\boldsymbol{\pi}$

In this section, we describe how to generate the synthetic conditional probability matrix $\boldsymbol{\pi}$ used for the case study. During the incident, CTA does not provide specific path recommendation information. For every individual, we assume that their actual path choices (referred to as the "status quo" choices) reflect their inherent preferences. Section 6.9.2 presents the method of inferring passengers' status quo choices during the disruption using smart card data [215]. The basic idea is to track their tap-in records when entering the Blue Line and nearby bus routes, and compare them with their historical travel histories to get the transfer information.

Given the status quo choices, we assume that the "true" passenger p's inherent preference for path r is given by

$$V_p^r = \begin{cases} 1 + v_p^r & \text{if } r \text{ is } p\text{'s actual path choice} \\ v_p^r & \text{otherwise,} \end{cases} \quad \forall\, p \in \mathcal{P}, r \in \mathcal{R}_p \tag{6.55}$$

where v_p^r is drawn uniformly from $\mathcal{U}[0, 1]$. Eq. 6.55 indicates every path has a random utility v_p^r normalized to $0 \sim 1$. And the chosen path has an additional utility value of 1. We assume that the impact of the recommendation of r' on the utility of path r is

$$I_{p,r'}^r = \begin{cases} \text{Drawn from } \mathcal{U}[0, 5] & \text{if } r = r' \\ 0 & \text{otherwise,} \end{cases} \quad \forall\, p \in \mathcal{P}, r, r' \in \mathcal{R}_p \tag{6.56}$$

Eq. 6.56 means that the utility of the path recommended (i.e., $r = r'$) has an additional positive impact drawn uniformly from $\mathcal{U}[0, 5]$. The utilities of paths not being recommended ($r \neq r'$) do not change. Given Eqs. 6.55 and 6.56, we can generate the conditional probability $\boldsymbol{\pi}$ using Eq. 6.21. It is worth mentioning that the above assumptions for generating synthetic passenger prior preferences are based on two reasonable principles: 1) Passenger's actual chosen path has a higher inherent utility. 2) Recommendations of a path can increase its probability of being chosen.

6.6.2 Parameter settings

The ϵ-feasibility and Γ-concentration parameters are set as $\epsilon = 0.05$ and $\Gamma = 0.3$, indicating 5% and 30% variation constraints in mean and variance. The equity requirement parameter is set as $E = 2$, meaning that we allow at most 10 minutes difference in travel time for passengers with the same OD and departure time. The convergence gap threshold for Benders decomposition is set as 1×10^{-8}. The post-adjustment updating step is set as $\lambda_k = \frac{1}{4}$ based on numerical tests.

6.6.3 Benchmark models

There are two benchmark path choice scenarios we use for comparison purposes:

Status-quo path choices. This scenario provides the status quo situation which does not include any recommendations. It represents the worst case. In this scenario, no behavior uncertainty is considered because this is based on the actual path choices realized by passengers.

Capacity-based path recommendations. The capacity-based path recommendations aim to recommend passengers to different paths according to the available capacity of paths. Specifically, for a path in OD pair (u, v) and time t, its capacity is the total available capacity of all vehicles passing through the first boarding station of the path during the time period. For example, for a path consisting of an NS bus route and the Green Line, the path capacity is the total available capacity of all buses at the boarding station of the NS bus route during time interval t. The available capacity can be obtained from a simulation model using historical demand as the input or using historical passenger counting data. The available capacity for the Blue line (the incident line) depends on modified operations during the incident (i.e., the service suspension is considered). When no vehicles operate in the Blue line during time interval t, the path capacity is zero.

6.6.4 System travel time evaluation

Given a recommendation strategy $\boldsymbol{x}$, as mentioned above, the actual system travel time is a random variable because of the passenger behavior uncertainty. To obtain the mean and standard deviation of the system travel time, we generate multiple passenger choice realizations based on $\boldsymbol{\pi}$ and $\boldsymbol{x}$. For each generated passenger choice

($\hat{\mathbb{1}}^r_{p,r'}$), the realized path flows are

$$\hat{q}^{u,v,r}_t = \sum_{p \in \mathcal{P}^{u,v}_t} \sum_{r' \in \mathcal{R}^{u,v}} x_{p,r'} \cdot \mathbb{1}^r_{p,r'} \quad \forall (u,v,r) \in \mathcal{F}, t \in \mathcal{T}. \tag{6.57}$$

The system travel time for the above passenger choice realization is calculated by solving the solving the optimal flow problem (Eq. 6.16) with the constraints $q^{u,v,r}_t = \hat{q}^{u,v,r}_t$ for all $(u,v,r) \in \mathcal{F}$ and $t \in \mathcal{T}$. This process is repeated with multiple realizations, providing the sample mean and standard deviation of the system travel time under recommendation strategy $\boldsymbol{x}$.

6.6.5 Experimental design

As this research considers various components (such as the optimal flow optimization, passengers' path preferences, behavior uncertainty, equity, etc.), it is useful to test different components separately to identify the impact of each one. Hence, we design the following test cases, each one with specific parameter settings to systematically evaluate the impacts of each component.

Model performance compared to benchmark models. The most straightforward model validation is to evaluate the effect on reducing system travel time. In this test case, we set $\Psi = 0$, meaning that we ignore the passengers' preference and focus only on minimizing system travel time. We also solve the individual path recommendation model (Eq. 6.35) without the post-adjustment process. The impact of the equity adjustments will be evaluated in another experiment introduced later. The results of this test case are discussed in Section 6.7.2

The benefit of considering behavior uncertainty. In this test case, we evaluate the importance of incorporating behavior uncertainty in the model. The model without behavior uncertainty assumes that passengers take the recommended path. The recommendation strategy is obtained by solving Eq. 6.35 with $\pi^r_{p,r'} = 1$ if $r = r'$. Similarly, we set $\Psi = 0$ and ignore the post-adjustment process. Note that, when we evaluate the recommendation strategy, the behavior uncertainty is still considered in generating the system travel time (see Section 6.6.4). The results of this test case are shown in Section 6.7.3

Impact of considering travel time equity. In this test case, the post-adjustment method is used to obtain "equity"-constrained path recommendations. The results before and after the post-adjustment are compared. The results of this test case are shown in Section 6.7.4.

Impact of considering passenger preference. In all the above tests, $\Psi = 0$ is used, focusing on the system travel time. In this test case, we evaluate the model performance under different values of Ψ in order to assess the impact of considering passenger preferences. The results of this test case are discussed in Section 6.7.5.

6.7 Results

6.7.1 Model convergence

Convergence of Benders decomposition and computational time

Figure 6-8 shows the convergence of the BD algorithm. As expected, the lower bound of the model keeps increasing, while the upper bound, after dropping significantly in early iterations, exhibits some fluctuations. The model converges after 28 iterations with a relative gap of less than $1{\times}10^{-8}$. The number of optimality cuts was 28 and no feasibility cut was generated.

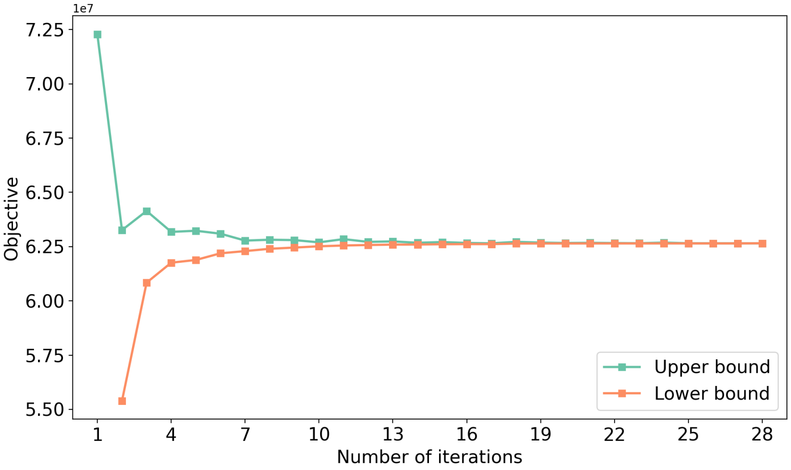


Figure 6-8: Convergence of the Benders decomposition

Table 6.1 compares the computational time of the Benders decomposition and off-the-shelf solvers. The BD algorithm was implemented using Julia 1.6 with the Gurobi 9.1 solver [216] on a personal computer with the I9-9900K CPU. The total computational time is 17.8 seconds (master problem 8.2 seconds + subproblem 9.6 seconds), which is more efficient than directly using the Mixed integer programming (MIP) solvers, including Gurobi [216], CPLEX [217], GLPK (GNU Linear Programming Kit) [218], and CBC (Coin-or branch and cut) [219].

Table 6.1: Computational time comparison

Solver	CPU time (sec)	Gap	Solver	CPU time (sec)	Gap
BD	17.8	0.000%	Gurobi	55.1	0.000%
CPLEX	65.7	0.000%	CBC	425.4	0.000%
GLPK	562.6	0.000%			

Convergence of post-adjustment algorithm

Figure 6-9 shows the convergence process of the post-adjustment algorithm. Specifically, we show the number of non-equity passengers (i.e., $|\mathcal{P}| - |{\mathcal{P}_{\text{Eq}}}^{(k)}|$) and the average value of w_p (i.e., $\frac{\sum_{p\in\mathcal{P}} w_p}{|\mathcal{P}|}$). Iteration 0 indicates the system optimal state (i.e., the solution of Eq. 6.35). We observe that, after only three iterations, the post-adjustment algorithm terminates and all passengers in $\mathcal{P}$ have a travel time difference relative to the shortest path (i.e., w_p) smaller than 10 minutes, satisfying the equity requirement. The average w_p keeps decreasing with the post-adjustment process, meaning that more and more passengers have similar travel times.

6.7.2 Model performance compared to benchmark models

In this section, we compare the system travel time under the proposed individual path recommendations (without post-adjustment) and two benchmark models. All travel times (except for the status quo that is deterministic) are calculated based on 10 replications using the randomly sampled actual path choices based on the given recommendation (see Section 6.6.4).

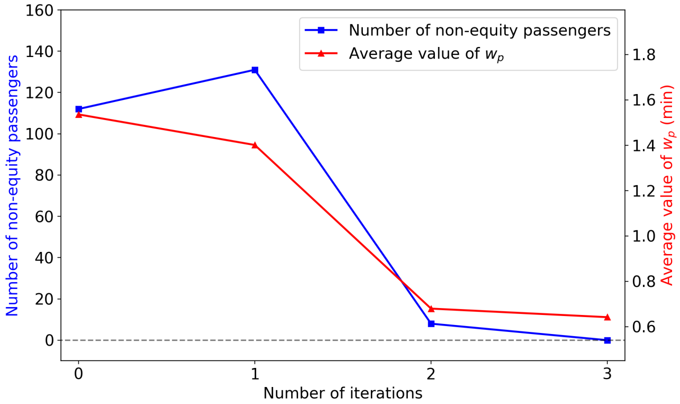


Figure 6-9: Convergence of the post-adjustment procedure

Table 6.2 shows that the proposed model (IPR) significantly reduces the average travel time in the system compared to the status quo. Specifically, there is a 6.6% reduction in travel times of all passengers in the system. And for passengers in the incident line (i.e., passengers who received the recommendation, $\mathcal{P}$), the average travel time reduction is 19.0%. Our model also outperforms the capacity-based benchmark path recommendation strategy, which reduces the travel time of all passengers by 2.5% and incident line passengers by 15.9%. It is also worth noting that the standard deviation is small, meaning that variations due to behavior uncertainty are not significant.

Table 6.2: Average travel time comparison for different models

Models	Average travel time (all passengers)		Average travel time (incident line passengers)	
	Mean (min)	Std. (min)	Mean (min)	Std. (min)
Status quo	28.318	N.A.	40.255	N.A.
Capacity-based	27.609 (-2.5%)	0.033	33.848 (-15.9%)	0.165
IPR model	26.457 (-6.6%)	0.018	32.626 (-19.0%)	0.187

Numbers in parentheses represent percentage travel time reduction compared to the status quo

6.7.3 Benefits of considering behavior uncertainty

In this section, we aim to compare the model with and without considering the behavior uncertainty. The model without behavior uncertainty assumes that all passengers follow the recommended path when designing the recommendation (but they may not in reality).

Table 6.3 shows the comparison of average travel time for the two models. As expected, considering behavior uncertainty in the path recommendation design achieves smaller travel time for all passengers and incident line passengers. Note that, though the 0.93% reduction (around 15 seconds saving per passenger) is relatively small, considering the large number of passengers in the system, the total travel time savings are still significant.

Table 6.3: Average travel time comparison with and without behavior uncertainty (BU)

Models	Average travel time (all passengers)		Average travel time (incident line passengers)	
	Mean (min)	Std. (min)	Mean (min)	Std. (min)
IPR model (w.o. BU)	26.706	0.026	32.852	0.122
IPR model (w. BU)	26.457 (-0.93%)	0.018	32.626 (-0.69%)	0.187

Numbers in parentheses represent percentage travel time reduction compared to the IPR model w.o. BU

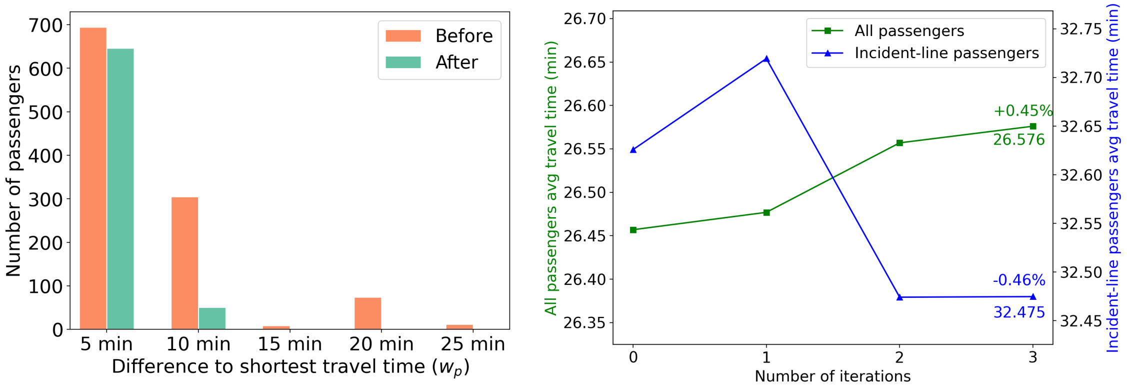
6.7.4 Impact of considering travel time equity

Figure 6-10 shows the comparison before and after the post-adjustment, which reflects the impact of considering passenger equity. Figure 6-10a shows that before the post-adjustment, there are hundreds of passengers with more than 10 minutes longer travel time than the shortest path travel time, showing equity issues. After the post-adjustment, w_p is less than 10 minutes for all passengers in $\mathcal{P}$. Furthermore, the distribution of w_p is shifted to smaller values. Note that $|\mathcal{P}|$ is around 5,800, so most passengers are recommended a path with the shortest travel time (i.e., $w_p = 0$)[4].

Figure 6-10b compares the system travel time before and after the post-adjustment as a function of the number of iterations. Since the system optimal solution has the smallest travel time, the post-adjustment, as expected, slightly increases the average travel time of all passengers by 0.45% (from 26.457 to 26.576 minutes), suggesting that a small sacrifice is enough to satisfy the equity requirement. Interestingly, the average travel time of the incident line passengers decreased (from 32.626 to 32.475 minutes), which implies that the increase in travel time mainly happens to passengers

[4] $w_p = 0$ is not shown in the distribution because it is too large and will distort the figure

in nearby lines who are indirectly affected by the incidents, rather than the incident-line passengers.



(a) Distribution of w_p before and after post-adjustment

(b) Average travel time change during post-adjustment

Figure 6-10: Comparison before and after the post-adjustment

6.7.5 Impact of respecting passenger's prior preferences

In this section, we evaluate the impact of different values of Ψ in terms of respecting passenger's prior preferences. Besides the system travel time, we also evaluate the total utility, defined as the sum of the prior utilities of the recommended path:

$$TU(\boldsymbol{x}) = \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}_p} x_{p,r} \cdot V_{p,r}. \tag{6.58}$$

Note that the maximum value of $TU(\boldsymbol{x})$ is achieved when every passenger is recommended their preferred path (i.e., the path with the highest prior utility, $V_{p,r}$). Denote this maximum value as $TU^{\max}$. The relative ratio of total utility, $\frac{TU(\boldsymbol{x})}{TU^{\max}}$, represents the fraction of the total (prior) utility that the recommendation has achieved.

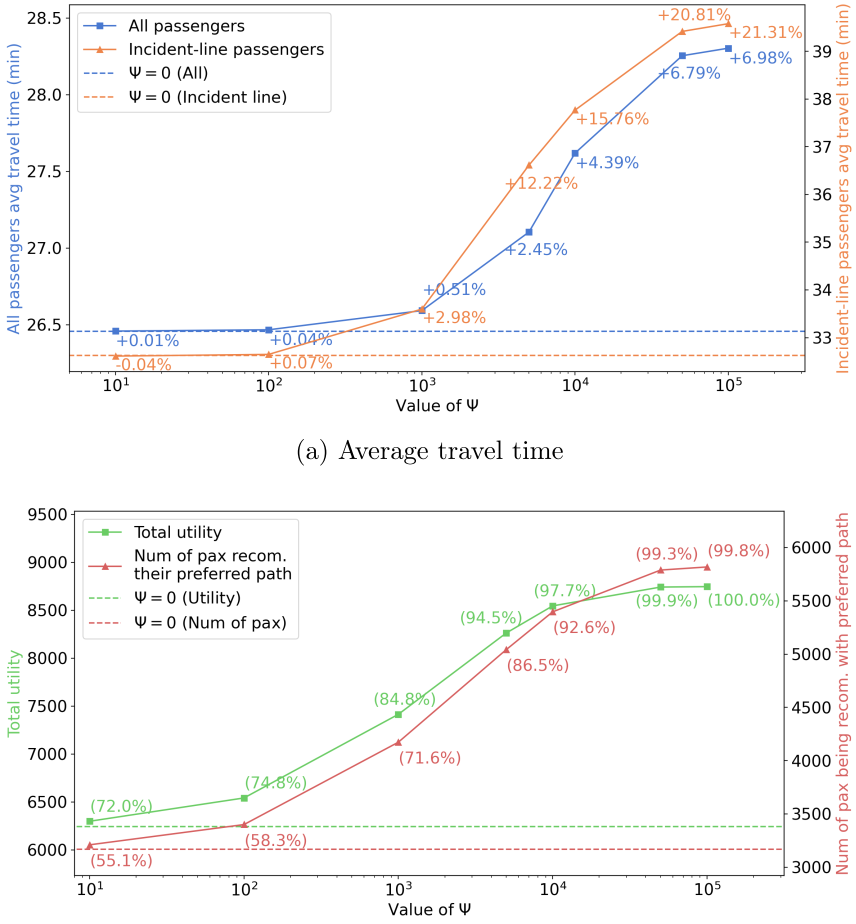
Another indicator is the number of passengers recommended their preferred path (denoted as $NP(\boldsymbol{x})$). Similarly, we also define the proportion of passengers recommended their preferred path (i.e., $\frac{NP(\boldsymbol{x})}{|\mathcal{P}|}$, where $|\mathcal{P}| = 5,827$ in the case study).

Figure 6-11 shows the results for different value of Ψ. The x-axis is plotted in log-scale. In Figure 6-11a, the average travel time for all passengers and incident-line passengers increases with the increase of Ψ, which is as expected because the larger value of Ψ means that the recommendation generation focuses more on satisfying passenger's inherent preferences rather than minimizing the system travel time. Similarly, in Figure 6-11b, as expected, both $TU(\boldsymbol{x})$ and $NP(\boldsymbol{x})$ increase with the increase in Ψ. When $\Psi = 10^5$, the average travel time of the incident line passengers increased by 21.3%, which is close to the status quo scenario. This is because we generate passengers' prior utilities based on the status quo choices. Figure 6-11b shows that nearly all passengers in $\mathcal{P}$ are recommended with their preferred path when $\Psi = 10^5$.

Figure 6-11 illustrates the trade-off between respecting passenger's preferences and reducing system congestion. When the value of Ψ is relatively small (e.g., less than 10^3), increasing Ψ can effectively increase the total utility and number of passengers recommended their preferred path. Meanwhile, the system travel time only slightly increases. But when Ψ is large (e.g., greater than 10^4), increasing Ψ significantly increases the system travel time, but the impact on increasing passenger's utility is limited. The reason may be that, in the system, there are some passengers whose preferred paths are not at the capacity bottlenecks. Hence, when Ψ is small, the optimal solution recommends those passengers use their preferred paths without significantly impacting the system travel time. When Ψ is large, passengers are recommended to use their preferred paths even if these paths are highly congested, causing a significant increase in the system travel time. The results imply that a reasonable value of Ψ should be relatively small. With small Ψ, most of the passengers (e.g., more than 70%) are recommended to use their preferred paths without significantly reducing the system efficiency.

6.8 Conclusion and discussion

This study proposes a mixed-integer programming (MIP) formulation to model the individual-based path (IPR) recommendation problem during PT service disruptions with the objective of minimizing total system travel time and respecting passengers'



(a) Average travel time

(b) Total utility and number of passengers being recommended with preferred path

Figure 6-11: Impact of different values of Ψ on results. The percentage change in Figure (a) is compared with the scenario of $\Psi = 0$. The percentage in parentheses in Figure (b) represents the relative ratio of total utility and proportion of passengers recommended their preferred path, respectively.

path choice preference. Passengers' behavior uncertainty in path choices given recommendations and their travel time equity are also considered in the formulation. We first formulate the optimal flow distribution problem in PT systems as a linear programming, which outputs the optimal path flows for each OD pair and time interval that minimize the total system travel time. Then, we model the behavior uncertainty based on passenger's prior preferences and posterior path choice probability distribution with two new concepts: ϵ-feasible flows and Γ-concentrated flows, which control the mean and variance of path flows in the optimization problem. We show that these two concepts can be transformed into linear constraints using Chebyshev's inequality. The individual path recommendation problem with behavior uncertainty is solved using Benders decomposition (BD) efficiently. The master problem of BD is a small-scale integer programming and the subproblem of BD reduces the optimal flow problem that is a linear program. The BD is more efficient than many off-the-shelf MIP solvers. Finally, we use a post-adjustment heuristic to address equity requirements.

The proposed approach is demonstrated in a case study using data from a real-world urban rail disruption in the CTA system. The results show that the proposed IPR model significantly reduces the average travel times in the system compared to the status quo. Specifically, there is a 6.6% reduction in travel times for all passengers in the system. Passengers in the incident line (i.e., passengers who received the recommendation), experience a 19.0% average travel time reduction. Our model also outperforms the capacity-based benchmark path recommendation strategy. Compared to the model that assumes all passengers would follow the recommendations, considering behavior uncertainty in the path recommendation design can achieve smaller system travel time. Post-adjustment effectively reduces the difference in passengers' travel times and increases equity. After the post-adjustment, the travel time difference to the shortest path is within 10 minutes for all passengers. The equity requirement slightly increases the system travel time by 0.46%, showing the trade-off between efficiency and equity. In terms of respecting passenger's preferences, we show that it is possible that most of the passengers (e.g., more than 70%) are recommended their preferred paths while only increasing the system travel time by 0.51%.

Following the discussion in Section 6.5, future studies can be pursued in the following directions. First, as shown in Section 6.5.1, it is possible to extend the current framework with more complex recommendation compositions. The challenges in implementing the more general framework stem from the quantification of the posterior path choice probabilities. Future studies may conduct corresponding surveys to calibrate passengers' responses to the recommendations. Second, the integrated formulation with both behavior uncertainty and equity is hard to solve (Section 6.5.3).

Though the post-adjustment heuristic works well, it is still a methodological challenge to develop a direct solution algorithm for the integrated formulation. Finally, future studies may consider different sources of uncertainty (including incident duration, in-vehicle time, demand, etc.) for a more realistic modeling framework.

6.9 Appendix

6.9.1 Notation

Table 6.4: Notation summary

Notation	**Description**
Model Parameters	
(u, v, r, i)	The i-th leg of path r for OD pair (u, v)
t	Integer time index, $t = 1$ represents the start of the incident. Non-positive time indices indicate time before the incident.
T^D	Time index at which the recommendation system stops working
t^{min}	Start time index of the whole analysis period (negative by definition)
t^{end}	End time index of the incident
T	End time index of the whole analysis period (greater than T^D)
$\mathcal{T}$	The set of time indices of analysis and $\mathcal{T} = \{t^{\text{min}}, t^{\text{min}} + 1, ..., T\}$
τ	The time duration that each time index represents
$\mathcal{P}$	The set of passengers that will receive the path recommendation
$\mathcal{R}_p$	The set of feasible paths for passenger $p \in \mathcal{P}$
$\mathcal{R}^{u,v}$	The set of feasible paths for OD pair (u, v)
$\Delta_t^{u,v,r,i}$	Travel time between the terminal station and the **boarding** station of leg $((u, v, r, i))$ for vehicle departing at time t
$\delta_t^{u,v,r,i}$	Travel time between the terminal station and the **alighting** station of leg $((u, v, r, i))$ for vehicle departing at time t

$f_t^{u,v,r}$	Number of passengers with OD pair (u, v) and departure time t and using path r who are not provided path recommendations
$d_t^{u,v}$	Total number of passengers with OD pair (u, v) and departure time t
Ω_1	The set of onboard flow indices at time $t = 1$
$\hat{z}_t^{u,v,r,i}$	Number of onboard passengers in the vehicle departing at time t in leg (u, v, r, i)
$\mathcal{P}_t^{u,v}$	The set of passengers with OD pair (u, v) and departure time t that will receive the path recommendation (a subset of $\mathcal{P}$)
$\mathcal{F}$	The set of all (u, v, r) indices
$\mathcal{S}$	The set of all stations (stops)
$\mathcal{W}$	The set of all OD pairs
$\mathcal{L}$	The set of all transit lines (routes) in the system
$\mathcal{I}^{u,v,r}$	The set of legs for path r of OD pair (u, v)
Vehicle (l, t)	The vehicle departing at time t on line l
$T_{l,t}$	The time that vehicle (l, t) arrives the last station of line l
$O_{l,t,t'}$	Total number of onboard passengers at time t' for vehicle (l, t)
$K_{l,t}$	The capacity of vehicles (l, t)
$T^{\text{IVT}}_{u,v,r,i,t}$	In-vehicle time of leg (u, v, r, i) of vehicle departing at time t
$AD_{s,t}$	Cumulative arriving demand at station s up to time t
$XD_{s,t}$	Cumulative transferring demand at station s up to time t
$BD_{s,t}$	Cumulative boarded demand at station s up to time t
$TT_t^{u,v,r}$	Travel time of a path (u, v, r) at time t
$\mathcal{T}_t^{u,v,r}$	The set of possible arrival times for path (u, v, r) at time t
$AT_t^{u,v,r}$	Arrival time at the destination for the group of passengers using path (u, v, r) and departing at time t
V_p^r	Passenger p's inherent preference (utility) of using path r
$I_{p,r'}^r$	The impact of the recommendation of path r' for passenger p on his/her utility of path r

$\pi_{p,r'}^{r}$	Conditional probability for passenger p to choose path r given that he/she is recommended with path r'
$\mu_t^{u,v,r}$	Expectation of $Q_t^{u,v,r}$
$\sigma_t^{u,v,r}$	Standard deviation of $Q_t^{u,v,r}$
ϵ	Threshold parameter for ϵ-feasibility
Γ	Threshold parameter for Γ-concentration
Ψ	The hyper-parameter to adjust the scale and balance the trade-off between system efficiency and passenger preference
w_p	Degree of unfairness for passenger p
E	A Predetermined integer threshold for the travel time difference
λ_k	Step size for post-adjustment at iteration k
$\mathcal{P}_{\text{Eq}}{}^{(k)}$	The set of passengers whose recommendations already satisfy the equity requirement at iteration k
Random Variables	
$\mathbb{1}_{p,r}^{r'}$	Binary variable indicating whether passenger p has chosen route r or not
$Q_t^{u,v,r}$	Actual flow for path (u,v,r) at time t
Decision Variables for Optimization Models	
$x_{p,r}$	Binary variable indicating whether recommending passenger p to use route r or not
$q_t^{u,v,r}$	Number of passengers in $\mathcal{P}$ with OD pair (u,v) and departure time t and using path r
$z_t^{u,v,r,i}$	Number of passengers boarding a vehicle that had started at time t on leg (u,v,r,i)
$y_{\tilde{t}}^{u,v,r,t}$	Binary decision variable indicating whether the arrival time of passenger departing at time t using (u,v,r) is $\tilde{t}$ or not
Y	Decision variable in the master problem of the BD, representing the tentative objective function of the subproblem

$(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{\rho})$	Dual variables in the subproblem of the BD

6.9.2 Inference of status quo choices

The status quo path choice inference method is based on our previous study [215], which is also similar to the trip-train method used for destination inference in open public transit systems (i.e., no tap-out).

[In the system when the incident happens]: Consider a passenger $p \in \mathcal{P}$ with an incident line tap-in record before the end of the incident, meaning that he/she were in the transit system when the incident happens. We then track his/her next tap-in record. If he/she next tap-in is a transfer at nearby bus or rail stations, we can identify his/her chosen path based on the transfer station. We can also identify the waiting passenger if he/she continues to use the incident line to his/her intended destinations inferred by his/her next tap-in records.

[Out of the system when the incident happens]: For a passenger $p \in \mathcal{P}$ with only a tap-in record in nearby bus or rail stations. He/she may be affected by the incident to change the tap-in station, or just use the service as a normal commute. To identify whether he/she was affected, we extract his/her travel histories on previous days without incidents to get the normal commute trajectories. If his/her tap-in time and location on the incident day has never appeared in the historical records before, we treat him/her as a passenger affected by the incident and identify his/her chosen path based on the tap-in station.

For passengers in $\mathcal{P}$ without next tap-in records or travel histories, we randomly assign him/her a status quo path based on the proportion of inferred passengers.

Chapter 7

Conclusion

7.1 Summary of results

7.1.1 System performance evaluation under short-term service suspensions using a bulk-service queue model

This chapter proposes a stochastic framework to model the resilience of public transit systems under short random service suspensions. Two aspects regarding resilience are evaluated: 1) system stability and 2) system performance changes (queue length and waiting time) under random service suspensions. We adopt a bulk-service queue model to formulate the queuing behavior at a station. A Markov chain model is used to model passenger flow dynamics across stations. We model the random service suspension as a two-state (failure and normal) Markov process, where vehicles are assumed to stop in the failure state. Under certain assumptions, we prove that headway can be represented as the difference between two compound Poisson exponential variables. Assuming no vehicle overtaking, we approximate the headway as a zero-inflated truncated normal distribution to obtain a closed-form moment generating function. Based on the headway distribution, the number of arrival passengers within a headway is derived. It is then used to calculate the mean and variance of queue length and waiting time at each station with analytical formulations. These analytical formulations allow efficient calculation of system performance and quantify the impact of random service suspensions. We also derive stability conditions of the system with a closed-form formula that implies the system is more likely to be unstable with high incident rates and long incident duration. The proposed model is implemented on a bus network with sensitivity analysis of different parameters (such as incident rate, incident duration, planned headway, etc.). Results show that higher incident rates and higher average incident duration will increase both the mean and variance of queue length and waiting time. Crowding stations (i.e., stations with high demand but low available capacity) are more vulnerable to random service suspensions. The theoretical results are validated with a simulation model, showing consistency between the two outcomes.

7.1.2 Empirical analysis for the impact of service disruptions

This chapter proposes a general incident analysis framework both from the supply and demand sides using automatically-collected data (AFC and AVL) in public transit systems. Specifically, from the supply side, we propose an incident-based network redundancy index to analyze the network's ability to provide alternative services under a specific rail disruption. The impacts on service operations are analyzed through the headway changes. From the demand side, we calculate the demand changes at different rail lines, rail stations, bus routes, and bus stops to understand the passenger flow redistribution under incidents. Individual behavior is analyzed using a binary logit model based on inferred passengers' mode choices and socio-demographics inferred from AFC and sale transaction data. Two incidents in the CTA public transit system are used as case studies. The two rail disruption cases have different attributes, one at a location with high network redundancy and the other with low network redundancy.

Results show that the service frequency of the incident line was largely reduced during the incident time. Nearby lines with substitutional functions are also slightly affected. Depending on the incident location, the network's redundancies are different, as well as the passengers' behavior. In the low redundancy scenario, most of the passengers chose to use nearby buses to move, either to their destinations or to the nearby rail lines. In the high redundancy scenario, most of the passengers transferred directly to nearby rail lines. The results of the case study provide useful insights into operations when dealing with incidents.

7.1.3 Inferring passenger behavioral responses under disruptions

This chapter proposes a probabilistic framework to infer passengers' response behavior to an unplanned rail service disruption using smart card data in a tap-in-only public transit system. We enumerate 19 possible behaviors that passengers may have based on the stages of their trips when the incident happened. A probabilistic model is proposed to estimate the mean and variance of the number of passengers in each of the 19 groups using passengers' historical and subsequent trip information. Based on the information used and the context of the behavior, four cases of formulations are used in the probabilistic model. Data from the CTA public transit system (bus and urban rail) is used for the case study with a rail incident. The model is implemented with both synthetic data (consistent with the CTA AFC data) and real-world data.

The proposed approach can estimate passengers' behavior well and outperform the rule-based benchmark model. Results with synthetic data show that the RMSE and MAPE for the estimated expected number of passengers in each behavior group are 143.9 and 20.5%, respectively. The RMSE and MAPE for the estimated standard deviation are 4.4 and 69.8%, respectively. The estimation results with real-world data are consistent with the incident's context. An indirect model validation using ridership change information and incident log data demonstrates the reasonableness of the results. Results with real-world data find that most of the passengers (97.43%) are not affected by the incident. This is reasonable because the incident only affected a small area. The incident we analyzed has high service redundancy with the Red line substituting the blocked Brown and Purple lines. Our model results show that in the high redundancy case, most of the affected passengers (69.51%) choose to use rail by changing routes. Based on the results, CTA operators can confirm that the Red line is a good alternative and quantify the impact. To relieve the incident impact, operators can increase service frequency in the Red line. The model indicates that only 8.1% of passengers choose to leave the public transit system. This number can help CTA conduct the service loss analysis due to the incident.

7.1.4 Station-based path recommendations under demand uncertainty

In this chapter, we propose a station-based path recommendation model to mitigate the congestion during public transit disruptions. Passengers with different ODs and departure times are recommended alternative paths to use such that the total system travel time is minimized. To tackle the non-analytical formulation of travel times due to left behind, we propose a simulation-based first-order approximation to transform the original problem into a linear program and solve the new problem iteratively with the method of successive average (MSA). Uncertainties in demand are modeled using RO techniques to protect the path recommendation strategies against inaccurate estimates. A real-world rail disruption scenario in the CTA system is used as a case study. Results show that even without considering uncertainty, the nominal model can reduce the system travel time by 9.1% (compared to the status quo), and outperforms the benchmark capacity-based path recommendation. The average travel time of passengers in the incident line is reduced more (-20.6% compared to the status quo). After incorporating the demand uncertainty, the robust model further reduces the system travel time. The best robust model with $\rho_{1-\epsilon} = 0.84$ decreases the average travel time of incident-line passengers by 2.91% compared to the nominal model.

The performance improvement by incorporating demand uncertainty is not very significant. The reason may be that demand variations in the incident situation have a limited impact on the optimal path shares. Notice that the demand during an incident is already very high for the system (due to the reduced supply level). Hence, the path recommendation patterns under nominal and worst-case demand may be similar. However, the methodology presented in this study provides a general way to deal with PT demand uncertainty. It can be used for other operations control, optimization, planning, or recommendation applications.

7.1.5 Individual-based path recommendation considering behavior uncertainty and equity

This chapter proposes a mixed-integer programming (MIP) formulation to model the individual-based path (IPR) recommendation problem during PT service disruptions with the objective of minimizing total system travel time and respecting passengers' path choice preferences. Passengers' behavior uncertainty in path choices given recommendations and their travel time equity are also considered in the formulation. We first formulate the optimal flow distribution problem in PT systems as linear programming, which outputs the optimal path flows for each OD pair and time interval that minimize the total system travel time. Then, we model the behavior uncertainty based on passenger's prior preferences and posterior path choice probability distribution with two new concepts: ϵ-feasible flows and Γ-concentrated flows, which control the mean and variance of path flows in the optimization problem. We show that these two concepts can be transformed into linear constraints using Chebyshev's inequality. The individual path recommendation problem with behavior uncertainty is solved using Benders decomposition (BD) efficiently. The master problem of BD is small-scale integer programming and the subproblem of BD reduces the optimal flow problem that is a linear program. The BD is more efficient than many off-the-shelf MIP solvers. Finally, we use a post-adjustment heuristic to address equity requirements.

The proposed approach is demonstrated in a case study using data from a real-world urban rail disruption in the CTA system. The results show that the proposed IPR model significantly reduces the average travel times in the system compared to the status quo. Specifically, there is a 6.6% reduction in travel times for all passengers in the system. Passengers in the incident line (i.e., passengers who received the recommendation), experience a 19.0% average travel time reduction. Our model also outperforms the capacity-based benchmark path recommendation strategy. Compared to the model that assumes all passengers would follow the recommendations, considering behavior uncertainty in the path recommendation design can achieve smaller system travel time. Post-adjustment effectively reduces the difference in passengers' travel times and increases equity. After the post-adjustment, the travel time difference to the shortest path is within 10 minutes for all passengers. The equity requirement slightly increases the system travel time by 0.46%, showing the trade-off between efficiency and equity. In terms of respecting passengers' preferences, we show that it is possible that most of the passengers (e.g., more than 70%) are recommended their preferred paths while only increasing the system travel time by 0.51%.

7.2 Summary of contributions

This thesis provides a framework about what and how we should do during the service disruptions for public transportation systems. We summarize different tasks in monitoring, control, and planning to build a resilient public transit system. All previous works regarding the resilience of public transit can be involved in the proposed framework. The proposed framework can guide many interesting future studies as a continuity of the five chapters included in the thesis.

7.2.1 New methodologies

From the methodology point of view, there are many new approaches proposed in the thesis to solve various challenges. These methods can also be applied to other research.

Closed-form headway distribution. In Chapter 2, we derive the headway distribution for a public transit system under short random service suspensions. These short-term incidents can be treated as disturbances in the system. Therefore, the derivation is general to normal public transit systems with perturbations. Typical headway modeling assumes a specific distribution for vehicles' travel times. This study provides a new framework by modeling disturbances of vehicle speed to derive the headway distribution. The headway distribution can be used to analyze a lot of queuing characteristics at a station, providing an efficient way to evaluate the level of services.

Though the derivation is based on public transit systems, we may also apply the headway (service interval) analysis to other bulk-service systems such as airport, logistics, and internet.

Interpolation-based roots-solving method. In Chapter 2, we propose an interpolation-based roots-solving method. Root-solving is an important component in queuing analysis. The proposed method can be used to solve other queuing systems for the steady-state distributions.

Simulation-based first-order approximation for travel time calculation in public transit system In Chapter 5, we mentioned that one of the key challenges in the public transit network modeling is that the system travel time has no analytical formulation. The proposed simulation-based first-order approximation provides a way to model the system travel time as a linear function, which allows for implementing advanced optimization techniques (e.g., robust optimization) or incorporating different model purposes (e.g., modeling of control strategies in the network).

Solving optimization problems with random decision variables. In Chapter 6, the individual path recommendation model needs to solve for the path flow and individual recommendation simultaneously. However, path flow becomes a random variable given behavior uncertainty. Optimization problems with random decision variables cannot be solved directly. Consider a general optimization problem where the decision variables are random variables with density function $f(\cdot \mid \boldsymbol{\theta})$ (Eq. 7.1).

$$\min_{\boldsymbol{x} \sim f(\cdot|\boldsymbol{\theta})} \quad g(\boldsymbol{x}) \tag{7.1a}$$

$$\text{s.t.} \quad h(\boldsymbol{x}) \leq \boldsymbol{b} \tag{7.1b}$$

The typical way to transform this problem into a deterministic problem is to take the expectation of the objective function and constraints (or consider the probability guarantee of the constraints with a pre-defined parameter η, such as $\eta = 0.95$), as shown in Eq. 7.2. That is, instead of solving for random variable $\boldsymbol{x}$, we treat the distribution parameters $\boldsymbol{\theta}$ as the decision variables, which is deterministic.

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}_{\boldsymbol{x}\sim f(\cdot|\boldsymbol{\theta})}[g(\boldsymbol{x}) \mid \boldsymbol{\theta}] \tag{7.2a}$$

$$\text{s.t.} \quad \mathbb{E}_{\boldsymbol{x}\sim f(\cdot|\boldsymbol{\theta})}[h(\boldsymbol{x})] \leq \boldsymbol{b} \ \text{ or } \ \mathbb{P}_{\boldsymbol{x}\sim f(\cdot|\boldsymbol{\theta})}[h(\boldsymbol{x}) \leq \boldsymbol{b}] \geq \eta \tag{7.2b}$$

However, the formulations in Eq. 7.2 is in general hard to solve except for that we have the closed-form expressions for $\mathbb{E}_{\boldsymbol{x}\sim f(\cdot|\boldsymbol{\theta})}[\cdot]$ and $\mathbb{P}_{\boldsymbol{x}\sim f(\cdot|\boldsymbol{\theta})}[\cdot]$.

In this study, we propose two concepts, ϵ-feasibility and Γ-concentration, to model random decision variables. Hence, instead of solving Eq. 7.2, we reformulate the problem to:

$$\min_{\hat{\boldsymbol{x}},\boldsymbol{\theta}} \quad g(\hat{\boldsymbol{x}}) \tag{7.3a}$$

$$\text{s.t.} \quad h(\hat{\boldsymbol{x}}) \leq \boldsymbol{b} \tag{7.3b}$$

$$|\hat{\boldsymbol{x}} - \mathbb{E}[\boldsymbol{x}]| \leq \epsilon \tag{7.3c}$$

$$\mathrm{Var}[\boldsymbol{x}] \leq \Gamma \tag{7.3d}$$

where $\hat{\boldsymbol{x}}$ in Eq. 7.3 is treated as a realization of $\boldsymbol{x}$ (instead of a random variable). $\boldsymbol{\theta} = (\mathbb{E}[\boldsymbol{x}], \mathrm{Var}[\boldsymbol{x}])$

The formulation in Eq. 7.3 has relationship with Eq. 7.2. When $\epsilon = 0$, we have $\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}]$. The objective function in Eq. 7.3 becomes $g(\mathbb{E}[\boldsymbol{x}])$ and the constraint becomes $h(\mathbb{E}[\boldsymbol{x}]) \leq \boldsymbol{b}$. If $g(\cdot)$ and $h(\cdot)$ are both convex functions (corresponding to convex optimization), according to Jensen's inequality, we have:

$$g(\mathbb{E}[\boldsymbol{x}]) \leq \mathbb{E}[g(\boldsymbol{x})] \tag{7.4}$$

$$h(\mathbb{E}[\boldsymbol{x}]) \leq \mathbb{E}[h(\boldsymbol{x})] \tag{7.5}$$

Then the optimal value of Eq. 7.3 is a lower bound of the optimal value of Eq. 7.2. Note that, in this case, Eq. 7.3 is also called certainty-equivalent (or mean-field) problem of a stochastic optimization problem.

On the other hand, since we also bound the variance of $\boldsymbol{x}$ in Eq. 7.3, we can also derive from the variance bound to a similar formula as $\mathbb{P}_{\boldsymbol{x}\sim f(\cdot|\boldsymbol{\theta})}[h(\boldsymbol{x}) \leq \boldsymbol{b}] \geq \eta$ (based on Chebyshev's and Jensen's inequalities), which shows another relationship between Eq. 7.3 and 7.2.

7.2.2 Contributions of each chapter

The main contributions of the five chapters are summarized below.

System performance evaluation under short-term service suspensions using a bulk-service queue model. This is the first study to explore analytically the bulk-service queuing problem involving short random service suspensions applied to PT systems. We model the service suspension in PT systems by analyzing vehicles' speed profiles, which is a novel and practical way to consider "server breakdown" in PT systems. 2) We prove that the headway under random service suspensions can be represented as the difference between two compound Poisson exponential variables. We assume there is no vehicle overtaking and approximate the headway distribution as a zero-inflated truncated normal distribution to obtain a closed-form moment generating function. Based on this we derive the PGF and corresponding moments of the number of arrival passengers within a headway (these are critical components for the bulk-service queue model). This is a new analytical contribution to the bulk-service queuing theory. 3) Based on Islam et al. [49]'s work, we introduce a Markov chain model to capture the inter-station passenger flow dynamics, which extends the typical bulk-service queuing analysis from the station level to the route level. 4) We propose an interpolation-based roots-solving method to find all complex roots for this study's model specification. Roots-solving is an essential step to obtain the queue length and waiting time for the bulk-service queuing model.

Empirical analysis for the impact of service disruptions. This chapter 1) proposes an incident-based network redundancy index to reflect the system's ability to provide alternative services considering the integrated bus and rail systems. The index leverages the proposed concept of path throughput to incorporate the impact of the incident duration on the redundancy calculation. 2) We develop an incident analysis framework using AFC and AVL data and apply it to incidents with different characteristics. Specifically, we analyze two types of incidents with high and low redundancy separately from both demand and supply perspectives. 3) An individual mode choice analysis method using AFC data is proposed. The approach includes a travel mode inference model and a passenger demographics extraction model. To the best of our knowledge, this is the first study that adopts AFC data for individual mode choice analysis during incidents. 4) We conduct an empirical study to demonstrate the proposed framework using AFC and AVL data from two real-world incidents in the CTA system. The corresponding policy implications and operation suggestions are also discussed.

Inferring passenger behavioral responses under disruptions. This chapter 1) provides a comprehensive framework of passengers' behavior under service disruptions. A total of 19 possible behavior groups for passengers at different stages of their trips are considered, which enables a more detailed modeling framework. The behavior identification is based on when and where passengers are making their decisions during a disruption. The method is general and can be used for other transit systems. 2) We propose a probabilistic behavior inference model with a specific formulation for each of the 19 behavior groups. The model enables the estimation of the mean and variance of the number of passengers in each group to capture passengers' behavior uncertainty. To the best of the authors' knowledge, this is the first article providing the estimation for both mean and variance of post-incident behaviors using AFC data. 3) The proposed approach leverages both passengers' historical travel trajectories and their subsequent tap-in records after the incident to facilitate behavior inference. This is contrary to previous studies where only the AFC data on the incident day is used.

Station-based path recommendations under demand uncertainty. The robust path recommendations study has two major contributions. 1) First, to tackle the non-analytical system travel time calculation, we propose a simulation-based linearization to convert the total system travel time to a linear function of path flows using first-order approximation. Importantly, we utilize the physical interaction between passengers and vehicles in a public transit system to efficiently calculate the gradient (i.e., marginal change of travel time) without running the simulation multiple times (as opposed to traditional black-box optimization). 2) Second, we use robust optimization (RO) to model the demand uncertainty which protects the model against inaccurate demand estimation. Specifically, we derive the closed-form robust counterpart with respect to the intersection of one ellipsoidal and three polyhedral uncertainty sets. These uncertainties capture the demand variations and the potential demand reduction during an incident. We also provide a feasible way of combining historical and survey data to quantify the uncertainty parameters.

Individual-based path recommendation considering behavior uncertainty and equity. This chapter is 1) the first study dealing with individual path recommendations under public transit service disruptions considering behavior uncertainty and equity. Previous studies only considered uncertainty in demand [45] or incident duration [161]. And for the objective function, they either focus on minimizing travel time or maximizing individual preferences [155]. Equity has not been considered in the literature. 2) To model behavior uncertainty, this chapter proposes a framework with prior path utility and posterior path choice distribution given recommendations. We use two new concepts: ϵ-feasibility and Γ-concentration, to control the mean and variance of path flows due to behavior uncertainty and transform these two requirements to linear constraints in the optimization model using Chebyshev's inequality. 3) Benders decomposition (BD) is used to solve the mixed-integer individual path recommendation problem efficiently. Under BD, the master problem becomes a small-scale integer program and the sub-problem reduces to a linear program. 4) This chapter mathematically defines the equity requirement in the individual path recommendations, and proposes a post-adjustment heuristic method to solve it. We also propose an integrated mixed-integer programming formulation with both behavior uncertainty and equity requirement, discuss the difficulty in solving the corresponding problem, and highlight the importance of the post-adjustment heuristic.

7.3 Application discussion

7.3.1 Implementation of path recommendation

To implement the path recommendation models proposed in Chapters 5 and 6, we need collaborations among different components in an agency. Figure 7-1 provides a possible workflow diagram for the implementation of path recommendation models. When an incident happens, the engineering team of a transit agency needs to first inspect the conditions and provide the estimated duration of the incident. These inspection results will be sent to the operation team to adjust the system's supply, including short-turn of trains, dispatching shuttle buses, etc. Note that supply-side optimal control models can be used to facilitate the operation adjustment. Given the supply and incident information, the proposed path recommendation models (both station-based and individual-based) can output the recommendations for passengers, which will be shown on the agency's App, websites, and electronic boards at stations. The software engineering team will support the design and maintenance of the App or other displayed platforms.

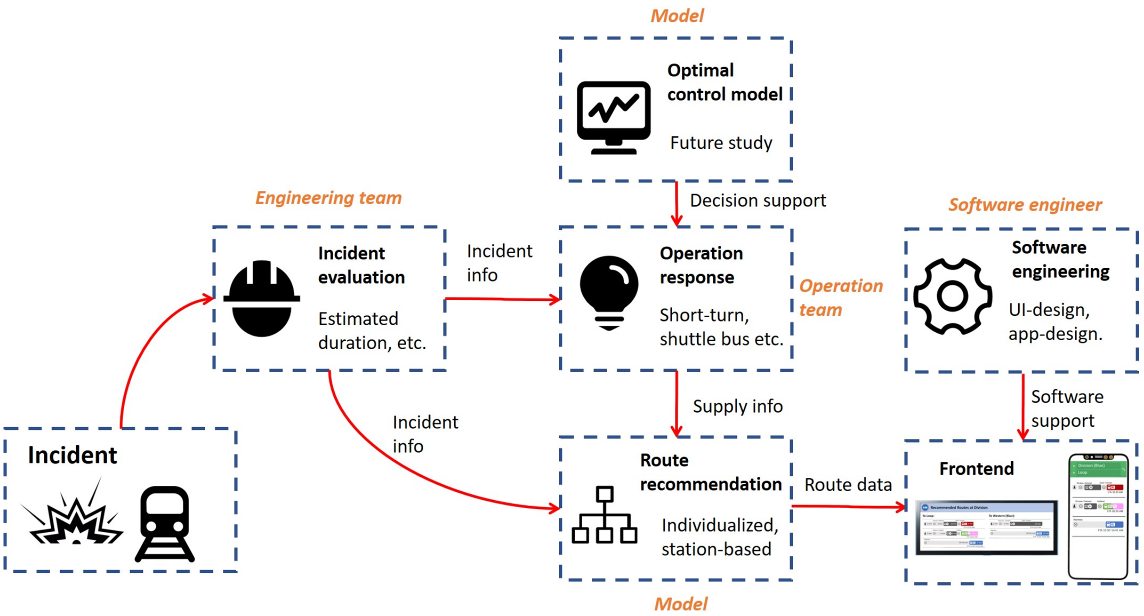


Figure 7-1: Framework for path recommendation implementation

Station-based path recommendation models

In the station-based path recommendation, our model can output the optimal path proportions for each OD pair and departure time (see Chapter 5). However, to make passengers' actual path choices in line with the optimal proportions, the implementation strategies need to be designed carefully. As mentioned in Section 5.3.5, one strategy is to randomly recommend to each passenger a specific path with the probability same as the optimal proportions. However, this may raise equity issues because some passengers may be recommended paths with significantly longer travel times than others at the same station.

An alternative implementation of station-based path recommendation is to list multiple paths with estimated travel time information. An example mock screen is shown in Figure 7-2. Since all passengers receive the same information, there are no aforementioned equity issues. However, in this setting, we need to determine which routes to show at a specific station and in which displaying formats. Different displays may result in different passengers' route choice probabilities. The objective is to make the choice probabilities as close to the optimal proportions as possible. However, this needs separate research to quantify passengers' behavior responses to different presented information and optimally decide the displayed information.

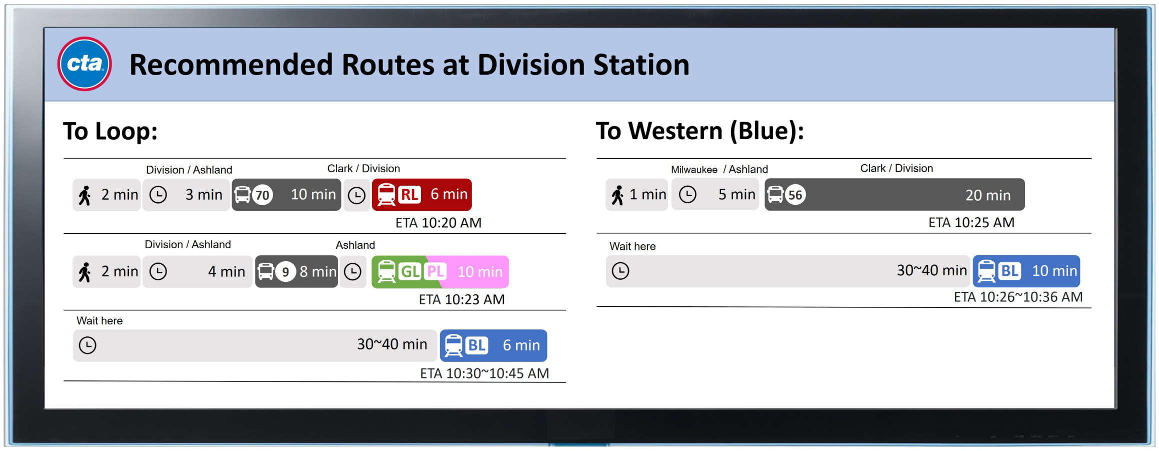
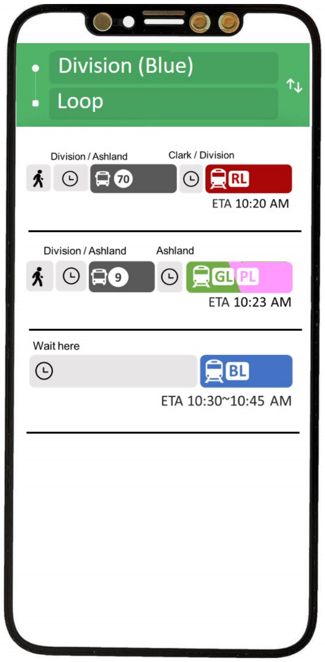


Figure 7-2: Mock electronic board at a station

Individual-based path recommendation

For the individual-based path recommendation, we can implement the model in a smart-phone based app. A mock screen for the recommendation system is shown in Figure 7-3. In the app, passengers who need recommendations can input their origin, destination, and departure times. And the system may recommend one or more paths with travel time information. Note that to enable recommendations of multiple paths, we need a model to quantify passengers' posterior choice probabilities with respect to different compositions of recommendation information (as discussed in Section 6.5).



Division (Blue)

Loop

Figure 7-3: Mock screen of the individual path recommendation app

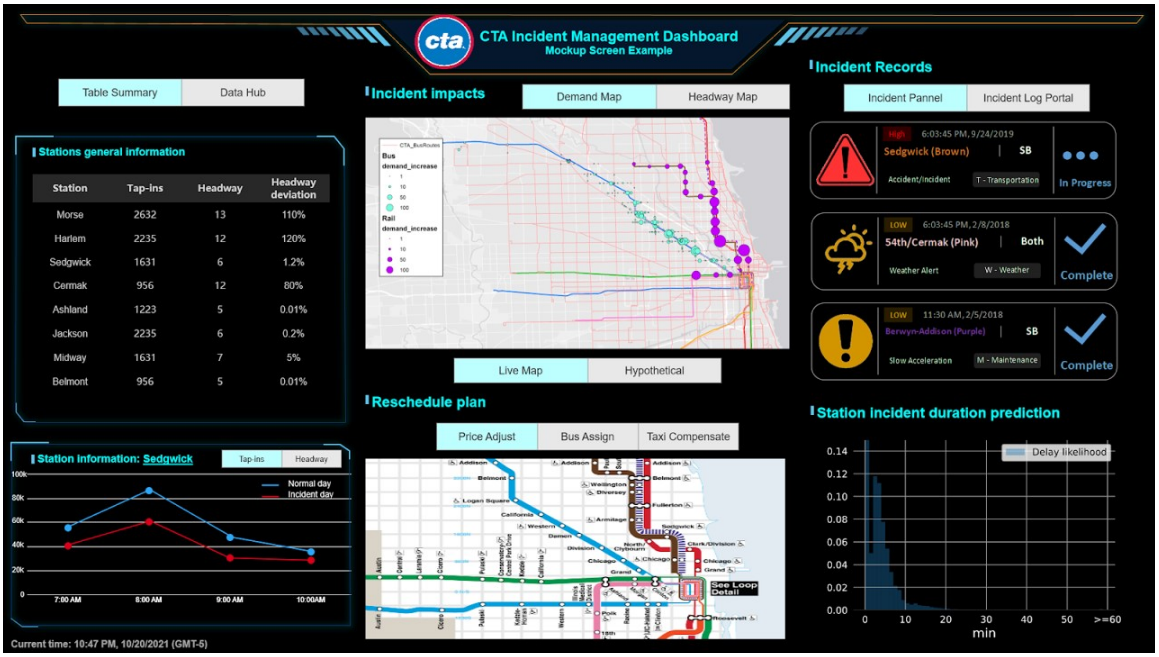
Pros and cons for two recommendation schemes

Station-based path recommendation, in general, is easier to implement because operators neither need to develop a smartphone-based app nor collect individual-level preference information. The drawback is that it cannot capture individual heterogeneity in responding to recommendations, and may lead to equity issues. The individual-based path recommendation model enables customized information provision, which may result in better performance and higher passenger satisfaction. However, the implementation is harder due to the challenges in collecting information at the individual level.

7.3.2 Incident management

Combining both monitoring and controlling models proposed in the dissertation (as well as other models discussed in Chapter 1), we can develop a holistic incident management tool for transit agencies. Figure 7-4 shows a mock screen for the incident management tool. On the screen, we can present real-time headway and its deviations (compared to the schedules), as well as real-time demand and its comparison to normal days. We can also visualize the impacts of the incident, showing the affected scopes and passenger flow redistribution. The tool may also provide guidance in operation adjustment and path recommendations.

However, to implement the automatic incident management platform, we need fast data collection and more efficient solution algorithms to enable real-time decisions. These are challenges that should be addressed in the future.



Stations general information

Station	Tap-ins	Headway	Headway deviation
Morse	2632	13	110%
Harlem	2235	12	120%
Sedgwick	1631	6	1.2%
Cermak	956	12	80%
Ashland	1223	5	0.01%
Jackson	2235	6	0.2%
Midway	1631	7	5%
Belmont	956	5	0.01%

High 6:03:45 PM, 9/24/2019
Sedgwick (Brown) | SB
Accident/incident T - Transportation
In Progress

LOW 6:03:45 PM, 2/8/2018
54th/Cermak (Pink) | Both
Weather Alert W - Weather
Complete

LOW 11:30 AM, 2/5/2018
Berwyn-Addison (Purple) | SB
Slow Acceleration M - Maintenance
Complete

Current time: 10:47 PM, 10/20/2021 (GMT-5)

Figure 7-4: Mock screen of the incident management system (credited to Dingyi Zhuang)

7.4 Future studies

7.4.1 Overcome limitations in current studies

There are several limitations in the above studies that can be improved in future studies. In Chapter 2, we assume that the incident durations for two consecutive vehicles are independent to facilitate closed-form derivations. In reality, if the incidents are caused by road congestion or infrastructure issues, it is possible that the incident durations for two consecutive vehicles passing through the same route segment are correlated. One possible way to capture these correlations is to model headway intervals as two different sequences to capture the first-order correlation. One may refer to Powell [48] for more details. Another improvement that can be done for Chapter 2 is to develop a better way to find the steady distribution of queue length. The current algorithm needs to solve for complex roots. It may also suffer from precision issues. Alternative methods include approximations of the queue length [220, 221], matrix-analytic method [222], or more recently, root-free methods [223].

In Chapter 3, we propose a redundancy index. An extension to the proposed index is to consider demand information. This can be done by replacing C_p as available capacity in a path considering onboard passengers. In that chapter, we also conducted the individual choice analysis. However, one of the drawbacks of individual analysis is that only a limited number of samples can be obtained due to the strict definition of regular passengers. Future studies may consider modifying the definition of regular passengers and enhancing the behavior inference model to get more samples. Alternatively, they can conduct research for aggregate demand analysis under disruptions (e.g., regression) by collecting many incident cases and the associated demand changes.

In Chapter 4, due to the data limitation in the open transit systems (i.e., no tap-out record), a lot of strong assumptions are imposed in the behavior response inference model in order to differentiate some behavior groups. Future studies may extend the current model to a closed transit system with tap-out information to allow a more granular behavior inference.

In Chapters 5 and 6, we consider demand and behavior uncertainty in the system. However, the most critical uncertainty in transit disruptions is the incident duration uncertainty. Due to the slow process of inspection, the control center may not fully understand the incident's characteristics (e.g., duration) before making operating decisions. Hence, future studies may consider a general framework for modeling incidents' uncertainties, including duration, affecting areas, etc.

7.4.2 Other control strategies

As discussed in Chapter 1, there are other control strategies in addition to path recommendations.

From the demand side, operators may control the inflow in front of the entering gates at each station. The objective of gate control is to limit the number of entering passengers, reduce downstream congestion, decrease risks at platforms, and force passengers to use other transportation modes (such as buses or taxis) so as to increase total system efficiency. Figure 7-5 shows the illustration of gate control. Considering a rail line with service disruptions, we may select several stations as the controlled stations. In each controlled station, we may design an algorithm to determine, for every five minutes, how many passengers (or what proportion of the total queuing passengers) should be allowed to enter the station. The waiting passengers may leave the queue and take alternative buses, which has the potential for better system capacity utilization.

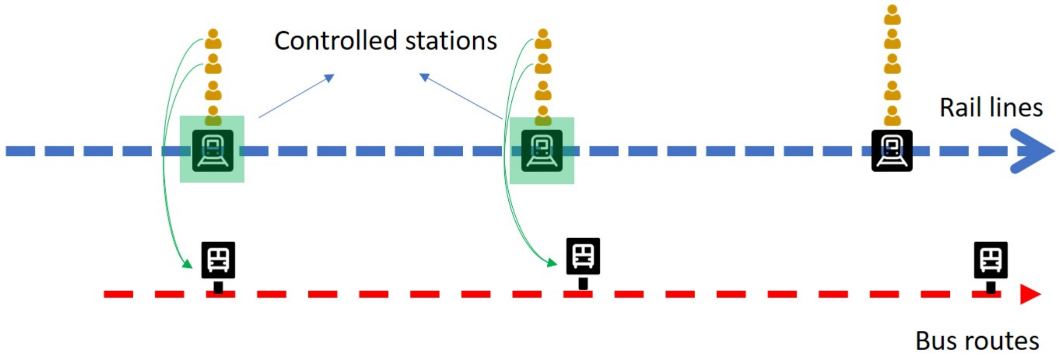


Figure 7-5: Illustration of gate flow control

Another demand-side control strategy may be passenger incentive design. In our individual path recommendation model, we introduce an equity constraint to ensure that passengers in the same situation (origin, destination, and departure time) have similar travel times. Another way to ensure equity is to provide subsidies (such as fare credits and digital tokens) to passengers with longer travel times. Hence, future studies may consider incentive design with path recommendations.

From the supply side, we may design algorithms for route re-scheduling, rolling stock adjustment, re-routing, shuttle bus dispatching, etc. Supply-based service adjustment has been extensively studied before, we thus do not discuss more details here.

7.4.3 Extension to planning tasks

Future studies may also explore proactive planning in response to potential service disruptions. Example tasks in planning include schedule (timetable) design, fleet size design, vehicle type and size design, crew scheduling, and service design (e.g., network extension). Incidents may result in uncertainties in these planning tasks. For example, timetable design may be affected due to uncertainties in vehicle travel time (e.g., buses travel slowly due to road accidents), dwelling time (e.g., crime risk causes trains to stop at a station for inspection), and the number of available vehicles (e.g., absence of bus drivers). Future studies may consider schedule design with respect to these uncertainties. Robust and stochastic optimization techniques can be used to solve these planning problems. For other planning tasks, it is also possible to consider incident-caused uncertainties and service degradation in the corresponding design.

7.4.4 Resilience quantification

In the dissertation, we mostly talk about resilience in a qualitative way. One of the future studies is to quantitatively define and evaluate resilience in a public transit system. Resilience quantification is commonly illustrated as in Figure 7-6, where the function of the system ($Q(t)$) is monitored over time. A shock initiates degradation of system function, followed by a period of consolidation, and eventual recovery of the nominal condition (Q_0). We evaluate the function loss as the lost areas due to system function degradation (the gray area in the figure). A higher resilience means a smaller function loss given a shock (or incidents). It can be achieved by lowering the magnitude of function degradation, or reducing the duration of the consolidation and recovery periods.

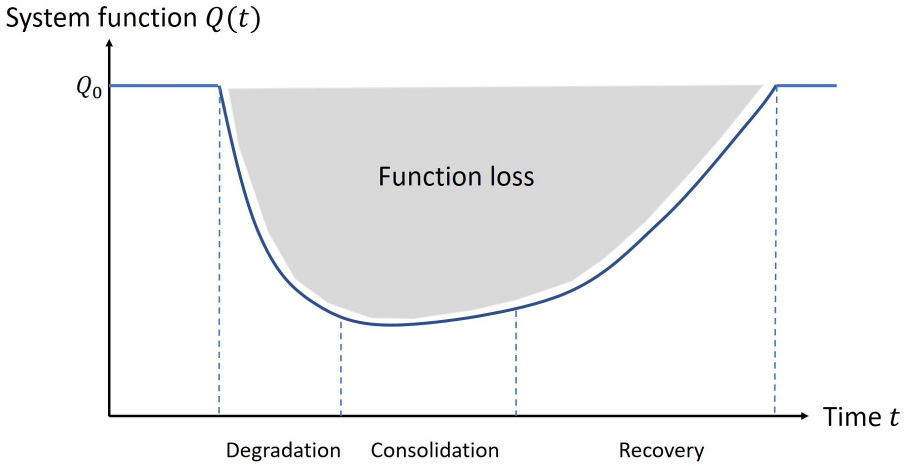


Figure 7-6: Illustration of resilience quantification (adapted from Jenelius and Mattsson [5])

Quantitative resilience analysis requires a measure of system function. Jenelius and Mattsson [5] proposed a framework for transport system resilience analysis that incorporates both supply and demand shocks. In their framework, system function loss corresponds to a shortage of supply in relation to demand. This can occur due to either a reduction in supply or an increase in demand. Figure 7-7 shows an example of supply losses, where a disruption causes a sudden supply cut while the demand remains relatively unchanged. The lack of resilience is represented by the total loss of function until supply is restored to the baseline level. This framework enables us to quantitatively evaluate the system's resilience based on the function loss.

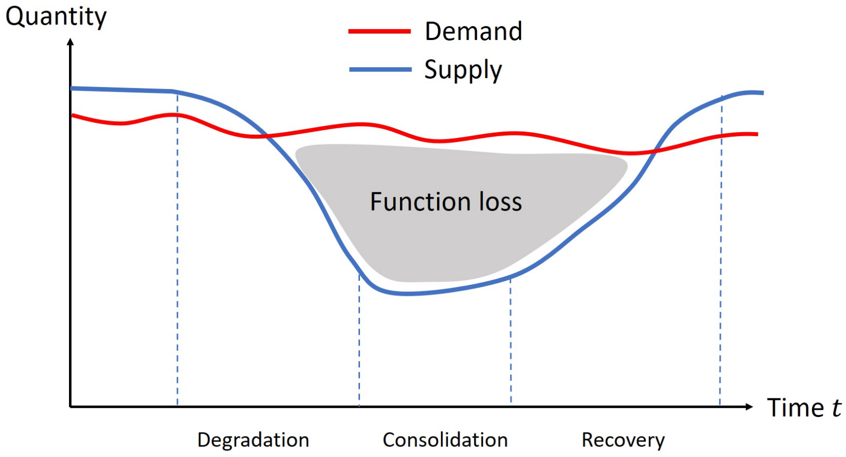


Figure 7-7: Resilience quantification for transportation systems (adapted from Jenelius and Mattsson [5])

7.4.5 Extension to multimodal system

Although this dissertation focuses on public transit systems, resilience can be extended to a multimodal transportation system as future research. To consider a multimodal resilient system, we should first define the source of incidents and service disruptions. In general, the multimodal system has more potential causes of incidents because every mode segment's function degradation will affect the integrated system. For example, the service disruption of bikes may affect the first and last-mile trips in the multimodal system. However, on another hand, a multimodal system may also be more resilient compared to a single-mode system because it has more capacities provided by various transportation modes.

After defining the potential incidents, we then can consider different tasks needed to maintain the operation of the multimodal system, and how these tasks should be modified under the conditions of incidents. For example, TNC companies need to run matching, pricing, and rebalancing algorithms during normal operations. Future studies may design more advanced algorithms for these tasks when there is significant demand (e.g., Olympic games, concerts) and supply (driver strike) fluctuations.

Bibliography

[1] Jesse M. Keenan. Theories of resilience, lecture presented at the harvard university graduate school of design, February 2019.

[2] Michael Vincent Martello. *Resilience of rapid transit networks in the context of climate change*. PhD thesis, Massachusetts Institute of Technology, 2020.

[3] Nikola Bešinović. Resilience in railway transport systems: a literature review and research agenda. *Transport Reviews*, 40(4):457–478, 2020.

[4] Baichuan Mo, Zhenliang Ma, Haris N Koutsopoulos, and Jinhua Zhao. Capacity-constrained network performance model for urban rail systems. *Transportation Research Record*, page 0361198120914309, 2020.

[5] Erik Jenelius and Lars-Göran Mattsson. Resilience of transport systems. *Encyclopedia of Transportation*, 2020.

[6] Andrew Cox, Fynnwin Prager, and Adam Rose. Transportation security and the role of resilience: A foundation for operational metrics. *Transport policy*, 18(2):307–317, 2011.

[7] Seyedmohsen Hosseini, Kash Barker, and Jose E Ramirez-Marquez. A review of definitions and measures of system resilience. *Reliability Engineering & System Safety*, 145:47–61, 2016.

[8] Jeryang Park, Thomas P Seager, Palakurth Suresh Chandra Rao, Matteo Convertino, and Igor Linkov. Integrating risk and resilience approaches to catastrophe management in engineering systems. *Risk analysis*, 33(3):356–367, 2013.

[9] Donald R Nelson. Adaptation and resilience: responding to a changing climate. *Wiley Interdisciplinary Reviews: Climate Change*, 2(1):113–120, 2011.

[10] Christopher B Field and Vicente R Barros. *Climate change 2014–Impacts, adaptation and vulnerability: Regional aspects*. Cambridge University Press, 2014.

[11] Milan Janić. Reprint of "modelling the resilience, friability and costs of an air transport network affected by a large-scale disruptive event". *Transportation Research Part A: Policy and Practice*, 81:77–92, 2015.

[12] David ZW Wang, Haoxiang Liu, WY Szeto, and Andy HF Chow. Identification of critical combination of vulnerable links in transportation networks–a global optimisation approach. *Transportmetrica A Transport Science*, 12(4):346–365, 2016.

[13] Serhiy Y Ponomarov and Mary C Holcomb. Understanding the concept of supply chain resilience. *The international journal of logistics management*, 2009.

[14] Mo Mansouri, Brian Sauser, and John Boardman. Applications of systems thinking for resilience study in maritime transportation system of systems. In *2009 3rd Annual IEEE Systems Conference*, pages 211–217. IEEE, 2009.

[15] Abdullah A Khaled, Mingzhou Jin, David B Clarke, and Mohammad A Hoque. Train design and routing optimization for evaluating criticality of freight railroad infrastructures. *Transportation Research Part B: Methodological*, 71:71–84, 2015.

[16] Min Zhou, Hairong Dong, Yanbo Zhao, Petros A Ioannou, and Fei-Yue Wang. Optimization of crowd evacuation with leaders in urban rail transit stations. *IEEE transactions on intelligent transportation systems*, 20(12):4476–4487, 2019.

[17] Yaoming Zhou, Junwei Wang, and Hai Yang. Resilience of transportation systems: concepts and comprehensive review. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4262–4276, 2019.

[18] Mostafa Bababeik, Navid Khademi, Anthony Chen, and M Mahdi Nasiri. Vulnerability analysis of railway networks in case of multi-link blockage. *Transportation Research Procedia*, 22:275–284, 2017.

[19] Raymond Chan and Joseph L Schofer. Measuring transportation system resilience: Response of rail transit to weather disruptions. *Natural Hazards Review*, 17(1):05015004, 2016.

[20] Qing-Chang Lu. Modeling network resilience of rail transit under operational incidents. *Transportation Research Part A: Policy and Practice*, 117:227–237, 2018.

[21] Kpotissan Adjetey-Bahun, Babiga Birregah, Eric Châtelet, and Jean-Luc Planchet. A model to quantify the resilience of mass railway transportation systems. *Reliability Engineering & System Safety*, 153:1–14, 2016.

[22] Ehab Diab and Amer Shalaby. Metro transit system resilience: Understanding the impacts of outdoor tracks and weather conditions on metro system interruptions. *International Journal of Sustainable Transportation*, 14(9):657–670, 2020.

[23] Emma Ferranti, Lee Chapman, Caroline Lowe, Steve McCulloch, David Jaroszweski, and Andrew Quinn. Heat-related failures on southeast england's railway network: Insights and implications for heat risk management. *Weather, Climate, and Society*, 8(2):177–191, 2016.

[24] David Dawson, Jon Shaw, and W Roland Gehrels. Sea-level rise impacts on transport infrastructure: The notorious case of the coastal railway line at dawlish, england. *Journal of Transport Geography*, 51:97–109, 2016.

[25] Robert Dorbritz. Assessing the resilience of transportation systems in case of large-scale disastrous events. In *Environmental Engineering. Proceedings of the International Conference on Environmental Engineering. ICEE*, volume 8, page 1070. Vilnius Gediminas Technical University, Department of Construction Economics ..., 2011.

[26] Michel Bruneau, Stephanie E Chang, Ronald T Eguchi, George C Lee, Thomas D O'Rourke, Andrei M Reinhorn, Masanobu Shinozuka, Kathleen Tierney, William A Wallace, and Detlof Von Winterfeldt. A framework to quantitatively assess and enhance the seismic resilience of communities. *Earthquake spectra*, 19(4):733–752, 2003.

[27] Michelle Oswald Beiler, Sue McNeil, David Ames, and Rebekah Gayley. Identifying resiliency performance measures for megaregional planning: Case study of the transportation corridor between boston, massachusetts, and washington, dc. *Transportation research record*, 2397(1):153–160, 2013.

[28] Paolo Bocchini, Dan M Frangopol, Thomas Ummenhofer, and Tim Zinke. Resilience and sustainability of civil infrastructure: Toward a unified approach. *Journal of Infrastructure Systems*, 20(2):04014004, 2014.

[29] Yalda Saadat, Yanjie Zhang, Dongming Zhang, Bilal M Ayyub, and Hongwei Huang. Post-failure recovery strategies for metrorail transit networks with washington dc as a case study. In *ASME International Mechanical Engineering Congress and Exposition*, volume 52187, page V013T05A060. American Society of Mechanical Engineers, 2018.

[30] Dong-ming Zhang, Fei Du, Hongwei Huang, Fan Zhang, Bilal M Ayyub, and Michael Beer. Resiliency assessment of urban rail transit networks: Shanghai metro as an example. *Safety Science*, 106:230–243, 2018.

[31] Baichuan Mo. *Network performance model for urban rail systems.* PhD thesis, Massachusetts Institute of Technology, 2020.

[32] Xiaolei Ma, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12, 2013.

[33] Baichuan Mo, Zhan Zhao, Haris N Koutsopoulos, and Jinhua Zhao. Individual mobility prediction in mass transit systems using smart card data: An interpretable activity-based hidden markov approach. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[34] Mariko Utsunomiya, John Attanucci, and Nigel Wilson. Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation research record*, 1971(1):118–126, 2006.

[35] Marie-Pier Pelletier, Martin Trépanier, and Catherine Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568, 2011.

[36] E Mazloumi, G Currie, and M Sarvi. Assessing measures of transit travel time variability and reliability using avl data. In *Transportation Research Board 87th Annual MeetingTransportation Research Board*, 2008.

[37] Amer Shalaby and Ali Farhan. Prediction model of bus arrival and departure times using avl and apc data. *Journal of Public Transportation*, 7(1):3, 2004.

[38] Fabian Cevallos, Xiaobo Wang, Zhenmin Chen, and Albert Gan. Using avl data to improve transit on-time performance. *Journal of Public Transportation*, 14 (3):2, 2011.

[39] Dan Levy and Llew Lawrence. *The Use of Automatic Vehicle Location for Planning and Management Information.* Number STRP# 4. 1991.

[40] CTA. Cta monthly ridership report for september 2019, 2019.

[41] Warren B Powell. Analysis of vehicle holding and cancellation strategies in bulk arrival, bulk service queues. *Transportation Science*, 19(4):352–377, 1985.

[42] Baichuan Mo, Li Jin, Haris N. Koutsopoulos, Zuo-Jun Max Shen, and Jinhua Zhao. Resilience of public transit systems under short random service suspension. In *Working paper*, 2022.

[43] Baichuan Mo, Max Y von Franque, Haris N Koutsopoulosc, John Attanuccid, and Jinhua Zhao. Impact of unplanned service disruptions on urban public transit systems. *arXiv preprint arXiv:2201.01229*, 2022.

[44] Baichuan Mo, Haris N Koutsopoulos, and Jinhua Zhao. Inferring passenger responses to urban rail disruptions using smart card data: A probabilistic framework. *Transportation Research Part E: Logistics and Transportation Review*, 159:102628, 2022.

[45] Baichuan Mo, Haris N Koutsopoulos, Max Zuo-Jun Shen, and Jinhua Zhao. Robust path recommendations during public transit disruptions under demand uncertainty. *arXiv preprint arXiv:2201.01437*, 2022.

[46] Baichuan Mo, Haris N. Koutsopoulos, and Jinhua Zhao. Individual path recommendation under public transit service disruptions considering behavior uncertainty and equity. In *Working paper*, 2022.

[47] Rajendra K Pachauri, Myles R Allen, Vicente R Barros, John Broome, Wolfgang Cramer, Renate Christ, John A Church, Leon Clarke, Qin Dahe, Purnamita Dasgupta, et al. *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change.* Ipcc, 2014.

[48] Warren Buckler Powell. *Stochastic delays in transportation terminals: New results in the theory and application of bulk queues.* PhD thesis, Massachusetts Institute of Technology, 1981.

[49] Md Kamrul Islam, Upali Vandebona, Vinayak V Dixit, and Ashish Sharma. A bulk queue model for the evaluation of impact of headway variations and passenger waiting behavior on public transit performance. *IEEE Transactions on Intelligent Transportation Systems*, 15(6):2432–2442, 2014.

[50] Yibing Wang, Jingqiu Guo, Avishai Avi Ceder, Graham Currie, Wei Dong, and Hao Yuan. Waiting for public transport services: Queueing analysis with balking and reneging behaviors of impatient passengers. *Transportation Research Part B: Methodological*, 63:53–76, 2014.

[51] Aykut Kahraman and Abhijit Gosavi. On the distribution of the number stranded in bulk-arrival, bulk-service queues of the m/g/1 form. *European journal of operational research*, 212(2):352–360, 2011.

[52] Ali Selvi and Matthew Rosenshine. A queueing system for airport buses. *Transportation Research Part B: Methodological*, 17(6):427–434, 1983.

[53] KC Madan. A single channel queue with bulk service subject to interruptions. *Microelectronics Reliability*, 29(5):813–818, 1989.

[54] IP Singh and Chhotu Ram. Three-server bulk service queue with service interruptions and exponential repairs. *Microelectronics Reliability*, 31(2-3):257–259, 1991.

[55] KC Madan. A bulk queueing system with random failures and two phase repairs. *Microelectronics Reliability*, 32(5):669–677, 1992.

[56] G Ayyappan and S Karpagam. Analysis of a bulk service queue with unreliable server, multiple vacation, overloading and stand-by server. *International Journal of Mathematics in Operational Research*, 16(3):291–315, 2020.

[57] D Jayaraman, R Nadarajan, and MR Sitrarasu. A general bulk service queue with arrival rate dependent on server breakdowns. *Applied mathematical modelling*, 18(3):156–160, 1994.

[58] Lotfi Tadj and Gautam Choudhury. A quorum queueing system with an unreliable server. *Applied mathematics letters*, 22(11):1710–1714, 2009.

[59] Lotfi Tadj, Gautam Choudhury, and Kamel Rekab. A two-phase quorum queueing system with bernoulli vacation schedule, setup, and n-policy for an unreliable server with delaying repair. *International Journal of Services and Operations Management*, 12(2):139–164, 2012.

[60] Norman TJ Bailey. On queueing processes with bulk service. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(1):80–87, 1954.

[61] F Downton. Waiting time in bulk service queues. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):256–261, 1955.

[62] NK Jaiswal. Time-dependent solution of the bulk-service queuing problem. *Operations Research*, 8(6):773–781, 1960.

[63] Marcel F Neuts. A general class of bulk queues with poisson input. *The Annals of Mathematical Statistics*, 38(3):759–770, 1967.

[64] ML Chaudhry and James GC Templeton. First course in bulk queues. 1983.

[65] S Sasikala and K Indhira. Bulk service queueing models–a survey. *International Journal of Pure and Applied Mathematics*, 106(6):43–56, 2016.

[66] Achyutha Krishnamoorthy, Padinhare K Pramod, and Srinivas R Chakravarthy. Queues with interruptions: a survey. *Top*, 22(1):290–320, 2014.

[67] Robert B Dial. Transit path finder algorithm. *Highway Research Record*, (205), 1967.

[68] FP Clerq. A public transport assignment method. *Verkeerstechniek, Netherlands*, 23(6), 1972.

[69] SC Wirasinghe. Nearly optimal parameters for a rail/feeder-bus system on a rectangular grid. *Transportation Research Part A: General*, 14(1):33–40, 1980.

[70] PI Welding. The instability of a close-interval service. *Journal of the operational research society*, 8(3):133–142, 1957.

[71] EE Osuna and Gordon F Newell. Control strategies for an idealized public transportation system. *Transportation Science*, 6(1):52–72, 1972.

[72] Warren B Powell. Bulk service queues with deviations from departure schedules: The problem of correlated headways. *Transportation Research Part B: Methodological*, 17(3):221–232, 1983.

[73] Amnon Rapoport, William E Stein, Vincent Mak, Rami Zwick, and Darryl A Seale. Endogenous arrivals in batch queues with constant or variable capacity. *Transportation Research Part B: Methodological*, 44(10):1166–1185, 2010.

[74] Philippe Henri Joseph Marguier. *Bus route performance evaluation under stochastic considerations.* PhD thesis, Massachusetts Institute of Technology, 1985.

[75] Mark D Hickman. An analytic stochastic model for the transit vehicle holding problem. *Transportation Science*, 35(3):215–237, 2001.

[76] Md Kamrul Islam, Upali Vandebona, Vinayak V Dixit, and Ashish Sharma. A model to evaluate the impact of headway variation and vehicle size on the reliability of public transit. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):1840–1850, 2015.

[77] Giuseppe Bellei and Konstantinos Gkoumas. Transit vehicles' headway distribution and service irregularity. *Public transport*, 2(4):269–289, 2010.

[78] Haodong Yin, Baoming Han, Dewei Li, and Ying Wang. Evaluating disruption in rail transit network: a case study of beijing subway. *Procedia Engineering*, 137:49–58, 2016.

[79] Xiangdong Xu, Anthony Chen, Sarawut Jansuwan, Kevin Heaslip, and Chao Yang. Modeling transportation network redundancy. *Transportation research procedia*, 9:283–302, 2015.

[80] Katja Berdica. An introduction to road vulnerability: what has been done, is done and should be done. *Transport policy*, 9(2):117–127, 2002.

[81] Graham Currie and Carlyn Muir. Understanding passenger perceptions and behaviors during unplanned rail disruptions. *Transportation Research Procedia*, 25:4396–4406, 12 2017.

[82] Pamela Murray-Tuite, Kris Wernstedt, and Weihao Yin. Behavioral shifts after a fatal rapid transit accident: A multinomial logit model. *Transportation research part F: traffic psychology and behaviour*, 24:218–230, 2014.

[83] Noriko Fukasawa, Kana Yamauchi, Akiko Murakoshi, Kohei Fujinami, and Daisuke Tatsui. Provision of forecast train information and consequential impact on decision making for train-choice. *Quarterly Report of RTRI*, 53(3): 141–147, 2012.

[84] Jing Teng and Wang-Rui Liu. Development of a behavior-based passenger flow assignment model for urban rail transit in section interruption circumstance. *Urban Rail Transit*, 1(1):35–46, 2015.

[85] Teddy Lin, Siva Srikukenthiran, Eric Miller, and Amer Shalaby. Subway user behaviour when affected by incidents in toronto (subwait) survey—a joint revealed preference and stated preference survey with a trip planner tool. *Canadian Journal of Civil Engineering*, 45(8):623–633, 2018.

[86] Ramachandran Balakrishna, Yang Wen, Moshe Ben-Akiva, and Constantinos Antoniou. Simulation-based framework for transportation network management in emergencies. *Transportation Research Record*, 2041(1):80–88, 2008.

[87] Pablo Suarez, William Anderson, Vijay Mahal, and TR Lakshmanan. Impacts of flooding and climate change on urban transportation: A systemwide performance assessment of the boston metro area. *Transportation Research Part D: transport and environment*, 10(3):231–244, 2005.

[88] Ling Hong, Jia Gao, and Wei Zhu. Self-evacuation modelling and simulation of passengers in metro stations. *Safety science*, 110:127–133, 2018.

[89] Huijun Sun, Jianjun Wu, Lijuan Wu, Xiaoyong Yan, and Ziyou Gao. Estimating the influence of common disruptions on urban rail transit networks. *Transportation Research Part A: Policy and Practice*, 94:62–75, 2016.

[90] Xiancai Tian and Baihua Zheng. Using smart card data to model commuters' responses upon unexpected train delays. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 831–840. IEEE, 2018.

[91] Jian Gang Jin, Kwong Meng Teo, and Amedeo R Odoni. Optimizing bus bridging services in response to disruptions of urban transit rail networks. *Transportation Science*, 50(3):790–804, 2016.

[92] Chunling Luo, Xinrong Li, Yuan Zhou, Aakil M Caunhye, Umberto Alibrandi, Nazli Y Aydin, Carlo Ratti, David Eckhoff, and Iva Bojic. Data-driven disruption response planning for a mass rapid transit system. In *Smart Transportation Systems 2019*, pages 205–213. Springer, 2019.

[93] Susan W O'Dell and Nigel HM Wilson. Optimal real-time control strategies for rail transit operations during disruptions. In *Computer-aided transit scheduling*, pages 299–323. Springer, 1999.

[94] Jian Gang Jin, Loon Ching Tang, Lijun Sun, and Der-Horng Lee. Enhancing metro network resilience via localized integration with bus services. *Transportation Research Part E: Logistics and Transportation Review*, 63:17–30, 2014.

[95] Leo Kroon and Dennis Huisman. Algorithmic support for railway disruption management. In *Transitions Towards Sustainable Mobility*, pages 193–210. Springer, 2011.

[96] Ehsan Rahimi, Ali Shamshiripour, Ramin Shabanpour, Abolfazl Mohammadian, and Joshua Auld. Analysis of transit users' waiting tolerance in response to unplanned service disruptions. *Transportation Research Part D: Transport and Environment*, 77:639–653, 2019.

[97] Alan F Beardon. *Complex analysis: The argument principle in analysis and topology*. Courier Dover Publications, 2019.

[98] Walter Rudin. *Real and complex analysis.* Tata McGraw-hill education, 2006.

[99] Robert V Hogg, Elliot A Tanis, and Dale L Zimmerman. *Probability and statistical inference.* Pearson/Prentice Hall Upper Saddle River, NJ, USA:, 2010.

[100] Per-Åke Andersson and Gian-Paolo Scalia-Tomba. A mathematical model of an urban bus route. *Transportation Research Part B: Methodological*, 15(4): 249–266, 1981.

[101] Carlos F Daganzo. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B: Methodological*, 43(10):913–921, 2009.

[102] John Burkardt. The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*, pages 1–35, 2014.

[103] Enrique R Villa and Luis A Escobar. Using moment generating functions to derive mixture distributions. *The American Statistician*, 60(1):75–80, 2006.

[104] Howard Wilson. Complex zeros of functions, 2014. MATLAB Central File Exchange. Retrieved December 6, 2020.

[105] Mohan L Chaudhry, BR Madill, and G Briere. Computational analysis of steady-state probabilities of m/g a, b/1 and related nonbulk queues. *Queueing systems*, 2(2):93–114, 1987.

[106] Ehsan Rahimi, Ali Shamshiripour, Ramin Shabanpour, Abolfazl Mohammadian, and Joshua Auld. Analysis of transit users' response behavior in case of unplanned service disruptions. *Transportation Research Record*, page 0361198120911921, 2020.

[107] Patrick O'Connor and Andre Kleyner. *Practical reliability engineering.* John Wiley & Sons, 2012.

[108] David C Wheeler and Morton E O'Kelly. Network topology and city accessibility of the commercial internet. *The Professional Geographer*, 51(3):327–339, 1999.

[109] Paul Kalungi and Tiku T Tanyimboh. Redundancy model for water distribution systems. *Reliability Engineering & System Safety*, 82(3):275–286, 2003.

[110] Yossi Sheffi and James B Rice Jr. A supply chain view of the resilient enterprise. *MIT Sloan management review*, 47(1):41, 2005.

[111] S Wilson-Goure, N Houston, and AV Easton. Assessment of the state of the practice and state of the art in evacuation transportation management. *Task Two: Literature Search for Federal Highway Administration (ITS-JPO). McLean, Virginia*, 2006.

[112] Pamela M Murray-Tuite. A comparison of transportation network resilience under simulated system optimum and user equilibrium conditions. In *Proceedings of the 2006 Winter Simulation Conference*, pages 1398–1405. IEEE, 2006.

[113] Anne Goodchild, Eric Jessup, Edward McCormack, Derik Andreoli, Kelly Pitera, Sunny Rose, Chilan Ta, et al. Development and analysis of a gis-based statewide freight data flow network. Technical report, Washington (State). Dept. of Transportation, 2009.

[114] Andrew Nash and Daniel Huerlimann. Railroad simulation using opentrack. *WIT Transactions on The Built Environment*, 74, 2004.

[115] Jan-Dirk Schmöcker, Shoshana Cooper, and William Adeney. Metro service delay recovery: comparison of strategies and constraints across systems. *Transportation research record*, 1930(1):30–37, 2005.

[116] Yuhang Wu, Baojing Huang, Xue Li, Yingnan Zhang, and Xinyue Xu. A data-driven approach to detect passenger flow anomaly under station closure. *IEEE Access*, 8:149602–149615, 2020.

[117] Tianyou Liu, Zhenliang Ma, and Haris N Koutsopoulos. Unplanned disruption analysis in urban railway systems using smart card data. *Urban Rail Transit*, pages 1–14, 2021.

[118] Ryuji Tsuchiya, Yoichi Sugiyama, and Riichiro Arisawa. A route choice support system for use during disrupted train operation. In *15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual MeetingITS AmericaERTICOITS JapanTransCore*, 2008.

[119] Anastasia Pnevmatikou and Matthew Karlaftis. Demand changes from metro line closures. 2011.

[120] Rohana Kamaruddin, Ismah Osman, and Che Anizaliana Che Pei. Public transport services in klang valley: customer expectations and its relationship using sem. *Procedia-Social and Behavioral Sciences*, 36:431–438, 2012.

[121] Yuan Bai and Lina Kattan. Modeling riders' behavioral responses to real-time information at light rail transit stations. *Transportation Research Record*, 2412 (1):82–92, 2014.

[122] Erik Jenelius and Oded Cats. The value of new public transport links for network robustness and redundancy. *Transportmetrica A: Transport Science*, 11(9):819–835, 2015.

[123] Muhammad Adnan, Francisco C Pereira, Carlos Lima Azevedo, Kakali Basak, Kenneth Koh, Harish Loganathan, Zhang Huai Peng, and Moshe Ben-Akiva. Evaluating disruption management strategies in rail transit using simmobility mid-term simulator: a study of singapore mrt north-east line. In *96th Annual Meeting of the Transportation Research Board, Washington, DC*, 2017.

[124] Anastasia M Pnevmatikou, Matthew G Karlaftis, and Konstantinos Kepaptsoglou. Metro service disruptions: how do people choose to travel? *Transportation*, 42(6):933–949, 2015.

[125] Lutz Leistritz, Thomas Weiss, Karl-Jürgen Bär, Fabrizio De VicoFallani, Fabio Babiloni, Herbert Witte, and Thomas Lehmann. Network redundancy analysis of effective brain networks; a comparison of healthy controls and patients with major depression. *Plos one*, 8(4):e60956, 2013.

[126] Baichuan Mo, Zhenliang Ma, Haris Koutsopoulos, and Jinhua Zhao. Assignment-based path choice estimation for metro system using smart card data. In *24th International Symposium on Transportation & Traffic Theory (ISTTT)*, 2020.

[127] Gabriel Goulet-Langlois, Haris N Koutsopoulos, Zhan Zhao, and Jinhua Zhao. Measuring regularity of individual travel patterns. *IEEE Transactions on Intelligent Transportation Systems*, 19(5):1583–1592, 2017.

[128] Carlos H Mojica. *Examining changes in transit passenger travel behavior through a smart card activity analysis.* PhD thesis, Massachusetts Institute of Technology, 2008.

[129] James J Barry, Robert Newhouser, Adam Rahbee, and Shermeen Sayeda. Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record*, 1817(1):183–187, 2002.

[130] Jinhua Zhao, Adam Rahbee, and Nigel HM 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.

[131] Jason B Gordon, Harilaos N Koutsopoulos, Nigel HM Wilson, and John P Attanucci. Automated inference of linked transit journeys in london using fare-transaction and vehicle location data. *Transportation research record*, 2343(1): 17–24, 2013.

[132] Moshe E Ben-Akiva, Steven R Lerman, Steven R Lerman, et al. *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press, 1985.

[133] Shauhrat S Chopra, Trent Dillon, Melissa M Bilec, and Vikas Khanna. A network-based framework for assessing infrastructure resilience: a case study of the london metro system. *Journal of The Royal Society Interface*, 13(118): 20160113, 2016.

[134] Qian Ye and Hyun Kim. Assessing network vulnerability of heavy rail systems with the impact of partial node failures. *Transportation*, 46(5):1591–1614, 2019.

[135] Nuannuan Leng, Valerio De Martinis, and Francesco Corman. Agent-based simulation approach for disruption management in rail schedule. In *Conference on Advanced Systems in Public Transport and TransitData (CASPT 2018)*, 2018.

[136] Sybil Derrible and Christopher Kennedy. The complexity and robustness of metro networks. *Physica A: Statistical Mechanics and its Applications*, 389 (17):3678–3691, 2010.

[137] X Zhang, E Miller-Hooks, and K Denny. Assessing the role of network topology in transportation network resilience. *Journal of Transport Geography*, 46:35–45, 2015.

[138] Stavri Dimitri Dimitrov and Avishai Avi Ceder. A method of examining the structure and topological properties of public-transport networks. *Physica A: Statistical Mechanics and its Applications*, 451:373–387, 2016.

[139] Fernando A López, Antonio Páez, Juan A Carrasco, and Natalia A Ruminot. Vulnerability of nodes under controlled network topology and flow autocorrelation conditions. *Journal of Transport Geography*, 59:77–87, 2017.

[140] Lucas P Veelenturf, Martin P Kidd, Valentina Cacchiani, Leo G Kroon, and Paolo Toth. A railway timetable rescheduling approach for handling large-scale disruptions. *Transportation Science*, 50(3):841–862, 2016.

[141] Lars Kjær Nielsen, Leo Kroon, and Gábor Maróti. A rolling horizon approach for disruption management of railway rolling stock. *European Journal of Operational Research*, 220(2):496–509, 2012.

[142] Sonia Adelé, Sabine Tréfond-Alexandre, Corinne Dionisio, and Pierre-Alain Hoyau. Exploring the behavior of suburban train users in the event of disruptions. *Transportation research part F: traffic psychology and behaviour*, 65: 344–362, 2019.

[143] Ricardo Silva, Soong Moon Kang, and Edoardo M Airoldi. Predicting traffic volumes and estimating the effects of shocks in massive transportation systems. *Proceedings of the National Academy of Sciences*, 112(18):5643–5648, 2015.

[144] Evelien van der Hurk. *Passengers, information, and disruptions.* Number EPS-2015-345-LIS. 2015.

[145] Baichuan Mo, Max Y von Franque, Haris N Koutsopoulos, John Attanucci, and Jinhua Zhao. Impact of unplanned service disruptions on urban public transit systems. *arXiv preprint arXiv:2201.01229*, 2022.

[146] Teddy Lin, Amer Shalaby, and Eric Miller. Transit user behaviour in response to service disruption: State of knowledge. In *Canadian Transportation Research Forum 51st Annual Conference-North American Transport Challenges in an Era of Change//Les défis des transports en Amérique du Nord à une aire de changement Toronto, Ontario*, 2016.

[147] Zhanhong Cheng, Martin Trépanier, and Lijun Sun. Probabilistic model for destination inference and travel pattern mining from smart card data. *Transportation*, pages 1–19, 2020.

[148] Baichuan Mo, Zhenliang Ma, Haris N Koutsopoulos, and Jinhua Zhao. Calibrating path choices and train capacities for urban rail transit simulation models using smart card and train movement data. *Journal of Advanced Transportation*, 2021, 2021.

[149] James J Barry, Robert Freimer, and Howard Slavin. Use of entry-only automatic fare collection data to estimate linked transit trips in new york city. *Transportation research record*, 2112(1):53–61, 2009.

[150] Yiwen Zhu, Haris N Koutsopoulos, and Nigel HM Wilson. A probabilistic passenger-to-train assignment model based on automated data. *Transportation Research Part B: Methodological*, 104:522–542, 2017.

[151] DU Peng, Liu Chao, and LIU Zhili. Walking time modeling on transfer pedestrians in subway passages. *Journal of Transportation Systems Engineering and Information Technology*, 9(4):103–109, 2009.

[152] Joaquin De Cea and Enrique Fernández. Transit assignment for congested public transport systems: an equilibrium model. *Transportation science*, 27(2): 133–147, 1993.

[153] Maurizio Bruglieri, Francesco Bruschi, Alberto Colorni, Alessandro Luè, Roberto Nocerino, and Vincenzo Rana. A real-time information system for public transport in case of delays and service disruptions. *Transportation Research Procedia*, 10:493–502, 2015.

[154] Katerina Böhmová, Matús Mihalák, Tobias Pröger, Rastislav Srámek, and Peter Widmayer. Robust routing in urban public transportation: How to find reliable journeys based on past observations. In *ATMOS-13th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems-2013*, volume 33, pages 27–41. Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik, 2013.

[155] DS Roelofsen, Oded Cats, Niels van Oort, and SP Hoogendoorn. Assessing disruption management strategies in rail-bound urban public transport systems from a passenger perspective. In *Proceedings of the 14th Conference on Advanced Systems in Public Transport (CASPT), Brisbane, Australia*, 2018.

[156] Xiudan Wang, Shaokuan Chen, Yangfan Zhou, Hongqin Peng, and Yuan Cui. Simulation on passenger evacuation under fire emergency in metro station. In *2013 IEEE International Conference on Intelligent Rail Transportation Proceedings*, pages 259–262. IEEE, 2013.

[157] Shaokuan Chen, Yue Di, Shuang Liu, and Baoshan Wang. Modelling and analysis on emergency evacuation from metro stations. *Mathematical Problems in Engineering*, 2017, 2017.

[158] Erfan Hassannayebi, Mehrdad Memarpour, Soheil Mardani, Masoud Shakibayifar, Iman Bakhshayeshi, and Shervin Espahbod. A hybrid simulation model of passenger emergency evacuation under disruption scenarios: A case study of a large transfer railway station. *Journal of Simulation*, 14(3):204–228, 2020.

[159] Hossam Abdelgawad and Baher Abdulhai. Large-scale evacuation using subway and bus transit: approach and application in city of toronto. *Journal of Transportation Engineering*, 138(10):1215–1232, 2012.

[160] Jiadong Wang, Zhenzhou Yuan, and Yonghao Yin. Optimization of bus bridging service under unexpected metro disruptions with dynamic passenger flows. *Journal of Advanced Transportation*, 2019, 2019.

[161] Zhijia Tan, Min Xu, Qiang Meng, and Zhi-Chun Li. Evacuating metro passengers via the urban bus system under uncertain disruption recovery time and heterogeneous risk-taking behaviour. *Transportation research part C: emerging technologies*, 119:102761, 2020.

[162] Jia Hao Wu, Michael Florian, and Patrice Marcotte. Transit equilibrium assignment: a model and solution algorithms. *Transportation Science*, 28(3):193–203, 1994.

[163] Jan-Dirk Schmöcker, Achille Fonzone, Hiroshi Shimamoto, Fumitaka Kurauchi, and Michael GH Bell. Frequency-based transit assignment considering seat capacities. *Transportation Research Part B: Methodological*, 45(2):392–408, 2011.

[164] Otto Anker Nielsen. A stochastic transit assignment model considering differences in passengers utility functions. *Transportation Research Part B: Methodological*, 34(5):377–402, 2000.

[165] Sang Nguyen, Stefano Pallottino, and Federico Malucelli. A modeling framework for passenger assignment on a transport network with timetables. *Transportation Science*, 35(3):238–249, 2001.

[166] Younes Hamdouch and Siriphong Lawphongpanich. Schedule-based transit assignment model with travel strategies and capacity constraints. *Transportation Research Part B: Methodological*, 42(7-8):663–684, 2008.

[167] Younes Hamdouch, WY Szeto, and Y Jiang. A new schedule-based transit assignment model with travel strategies and supply uncertainties. *Transportation Research Part B: Methodological*, 67:35–67, 2014.

[168] Jan-Dirk Schmöcker, Michael GH Bell, and Fumitaka Kurauchi. A quasi-dynamic capacity constrained frequency-based transit assignment model. *Transportation Research Part B: Methodological*, 42(10):925–945, 2008.

[169] A. Ben-Tal, L. El Ghaoui, and A.S. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

[170] Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[171] Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

[172] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.

[173] Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[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] 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.

[176] 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.

[177] 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.

[178] Dimitris Bertsimas and Dick den Hertog. *Robust and adaptive optimization*. Dynamic Ideas LLC, Belmont, Massachusetts, 2020.

[179] Meghan K Cain, Zhiyong Zhang, and Ke-Hai Yuan. Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior research methods*, 49(5):1716–1735, 2017.

[180] Aharon Ben-Tal, Dick Den Hertog, and Jean-Philippe Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical programming*, 149(1):265–299, 2015.

[181] Siriphong Lawphongpanich and Yafeng Yin. Solving the pareto-improving toll problem via manifold suboptimization. *Transportation Research Part C: Emerging Technologies*, 18(2):234–246, 2010.

[182] Dimitris Bertsimas, Eugene Litvinov, Xu Andy Sun, Jinye Zhao, and Tongxin Zheng. Adaptive robust optimization for the security constrained unit commitment problem. *IEEE transactions on power systems*, 28(1):52–63, 2012.

[183] Y Sheffi. Urban transportation networks. *Prentice-Hall, Inc., Englewood Cliffs, NJ*, 1985.

[184] 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, 2020.

[185] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530, 2007.

[186] Osman Khalid, Muhammad Usman Shahid Khan, Samee U Khan, and Albert Y Zomaya. Omnisuggest: A ubiquitous cloud-based context-aware recommendation system for mobile social networks. *IEEE Transactions on Services Computing*, 7(3):401–414, 2013.

[187] Kam Fung Yeung and Yanyan Yang. A proactive personalized mobile news recommendation system. In *2010 Developments in E-systems Engineering*, pages 207–212. IEEE, 2010.

[188] Abdul Majid, Ling Chen, Gencai Chen, Hamid Turab Mirza, Ibrar Hussain, and John Woodward. A context-aware personalized travel recommendation system based on geotagged social media data mining. *International Journal of Geographical Information Science*, 27(4):662–684, 2013.

[189] Hani S Mahmassani. Uncertainty in transportation systems evaluation: issues and approaches. *Transportation planning and technology*, 9(1):1–12, 1984.

[190] Kenneth E Train. *Discrete choice methods with simulation.* Cambridge university press, 2009.

[191] Baichuan Mo, Qing Yi Wang, Joanna Moody, Yu Shen, and Jinhua Zhao. Impacts of subjective evaluations and inertia from existing travel modes on adoption of autonomous mobility-on-demand. *Transportation Research Part C: Emerging Technologies*, 130:103281, 2021.

[192] Violeta Mirchevska. *Behavior Modeling by Combining Machine Learning and Domain Knowledge.* PhD thesis, PhD Thesis, IPS Jožef Stefan, Ljubljana, Slovenia, 2013.

[193] Shenhao Wang, Baichuan Mo, and Jinhua Zhao. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. *Transportation Research Part C: Emerging Technologies*, 112:234–251, 2020.

[194] Gyugeun Yoon and Joseph YJ Chow. Contextual bandit-based sequential transit route design under demand uncertainty. *Transportation Research Record*, 2674(5):613–625, 2020.

[195] June Young Jung, Gary Blau, Joseph F Pekny, Gintaras V Reklaitis, and David Eversdyk. A simulation based optimization approach to supply chain management under demand uncertainty. *Computers & chemical engineering*, 28(10): 2087–2106, 2004.

[196] Anirudh Subramanyam, Frank Mufalli, José M Laínez-Aguirre, Jose M Pinto, and Chrysanthos E Gounaris. Robust multiperiod vehicle routing under customer order uncertainty. *Operations Research*, 69(1):30–60, 2021.

[197] Matt Horne, Mark Jaccard, and Ken Tiedemann. Improving behavioral realism in hybrid energy-economy models using discrete choice studies of personal transportation decisions. *Energy Economics*, 27(1):59–77, 2005.

[198] Todd Litman. Evaluating transportation equity. *World Transport Policy & Practice*, 8(2):50–65, 2002.

[199] Farideh Ramjerdi. Equity measures and their performance in transportation. *Transportation Research Record*, 1983(1):67–74, 2006.

[200] Karel Martens. Substance precedes methodology: on cost–benefit analysis and equity. *Transportation*, 38(6):959–974, 2011.

[201] 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, 2021.

[202] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 109–116, 2011.

[203] Gediminas Adomavicius and YoungOk Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):896–911, 2011.

[204] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[205] Kenneth J Button and David A Hensher. *Handbook of transport systems and traffic control.* Emerald Group Pub Ltd, 2001.

[206] Hai Yang and Hai-Jun Huang. *Mathematical and economic theory of road pricing.* 2005.

[207] Yu Marco Nie and Yafeng Yin. Managing rush hour travel choices with tradable credit scheme. *Transportation Research Part B: Methodological*, 50:1–19, 2013.

[208] Li-Jun Tian, Hai Yang, and Hai-Jun Huang. Tradable credit schemes for managing bottleneck congestion and modal split with heterogeneous users. *Transportation Research Part E: Logistics and Transportation Review*, 54:1–13, 2013.

[209] John Renne. Evacuation and equity. *Planning*, 72(5), 2006.

[210] Thomas W Sanchez and Marc Brenman. Transportation equity and environmental justice: Lessons from hurricane katrina. *Environmental Justice*, 1(2): 73–80, 2008.

[211] Douglas R Bish. Planning for a bus-based evacuation. *OR spectrum*, 33(3): 629–654, 2011.

[212] John L Renne and Estefania Mayorga. What has america learned since hurricane katrina? evaluating evacuation plans for carless and vulnerable populations in 50 large cities across the united states. Technical report, 2018.

[213] Stephen D Wong, Joan L Walker, and Susan A Shaheen. Bridging the gap between evacuations and the sharing economy. *Transportation*, 48(3):1409–1458, 2021.

[214] Jacques F Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische mathematik*, 4(1):238–252, 1962.

[215] Baichuan Mo, Haris Koutsopoulos, and Jinhua Zhao. Inferring passenger responses to urban rail disruptions using smart card data: A probabilistic framework. *Transportation Research Part E: Logistics and Transportation Review*, 159:102628, 2022.

[216] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2021. URL `https://www.gurobi.com`.

[217] IBM ILOG Cplex. V12. 1: User's manual for cplex. *International Business Machines Corporation*, 46(53):157, 2009.

[218] Andrew Makhorin. Glpk (gnu linear programming kit). *http://www. gnu. org/s/glpk/glpk. html*, 2008.

[219] John Forrest and Robin Lougee-Heimer. Cbc user guide. In *Emerging theory, methods, and applications*, pages 257–277. INFORMS, 2005.

[220] Donald R McNeil. A solution to the fixed-cycle traffic light problem for compound poisson arrivals. *Journal of Applied Probability*, 5(3):624–635, 1968.

[221] Mark S van den Broek, JSH Van Leeuwaarden, Ivo JBF Adan, and Onno J Boxma. Bounds and approximations for the fixed-cycle traffic-light queue. *Transportation Science*, 40(4):484–496, 2006.

[222] Marcel F Neuts. Matrix-analytic methods in queuing theory. *European Journal of Operational Research*, 15(1):2–12, 1984.

[223] Anna Oblakova, Ahmad Al Hanbali, Richard J Boucherie, Jan CW van Ommeren, and WHM Zijm. An exact root-free method for the expected queue length for a class of discrete-time queueing systems. *Queueing systems*, 92(3): 257–292, 2019.