Robust Path Recommendations During Public Transit Disruptions Under Demand Uncertainty

Baichuan Mo[a,*], Haris N. Koutsopoulos[b], Max Zuo-Jun Shen[d], Jinhua Zhao[c]

[a]*Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139*
[b]*Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115*
[c]*Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 20139*
[d]*Department of Industrial Engineering and Operations Research, University of California, Berkeley, Berkeley, CA 94720*

Abstract

When there are significant service disruptions in public transit systems, passengers usually need guidance to find alternative paths. This paper proposes a path recommendation model to mitigate the congestion during public transit disruptions. Passengers with different origin-destination and departure times are recommended with different paths such that the system travel time is minimized. We model the path recommendation as an optimal flow problem with uncertain demand information. 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 linear programming. Uncertainties in demand are modeled with robust optimization to protect the path recommendation strategies against inaccurate estimates. A real-world rail disruption scenario in the Chicago Transit Authority (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 (i.e., passengers receiving recommendations) is reduced more (-20.6% compared to the status quo). After incorporating the demand uncertainty, the robust model can further reduce the system travel time. The best robust model can decrease the average travel time of incident-line passengers by 2.91% compared to the nominal model.

Keywords: Path recommendation; Robust optimization; Rail disruptions; Demand uncertainty

1. Introduction

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 (Cox et al., 2011).

This study considers significant service disruptions in public transit systems where the incident 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 (especially during incident time), the new routes chosen by passengers may not be optimal or even cause more congestion (Mo et al., 2021). For example, during a

*Corresponding author

rail disruption, most of the passengers may choose some 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).

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 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, we need to have an analytical formulation to calculate passengers' travel time of a path. However, passengers' waiting time at the boarding and transfer station is not only determined by other waiting passengers, but only those who already boarded the same line as they reduce the vehicle's capacity (De Cea and Fernández, 1993). 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 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). None of the previous studies have 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 and protect the recommendation results against uncertainties due to inaccurate demand estimates. Different from previous recommendation systems that focus on maximizing individual preference, 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 first-order approximation, which provides 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 with the robust optimization (RO) technique. 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 paper 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 robust optimization 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 paper is organized as follows. Literature review is shown in Section 2. In Section 3, we describe the problem and discuss the solution methods. We apply the proposed framework on the CTA system as a case study in Section 5. The model results are analyzed in Section 6. Finally, we conclude our study, summarize our main findings in Section 7.

2. Literature review

2.1. Path recommendation during incidents

Most previous studies on path recommendation under incidents focus on individual or 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. (2015) 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. (2013) developed a routing algorithm in urban public transportation to find reliable journeys that are robust for system delays. Roelofsen et al. (2018) 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 at different OD pairs and with different departure times so that the system travel time is reduced.

2.2. Passenger evacuation under emergency

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. In general, these evacuation papers can be categorized into micro-level and macro-level based on how passenger flows are modeled and the range of study areas.

The micro-level studies usually use an agent-based simulation model to evaluate different evacuation strategies within some infrastructure. For example, Wang et al. (2013) simulated the passenger evacuation under fire emergency in Metro stations. Chen et al. (2017) implemented four methodologies including a queuing model and an agent-based simulation to calculate the evacuation time under a given number of evacuees in different emergency situations and evacuation plans. Hassannayebi et al. (2020) 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, including train failure in the tunnel and fire at the station gallery. Zhou et al. (2019) proposed a hybrid bi-level model to optimize the number and initial locations of leaders as well as the routes of leaders during the evacuation to guide passenger's evacuation in urban rail transit stations.

The macro-level studies consider a larger study area (e.g., city-level) and aim to evacuate passengers from the incident location through various transportation modes. For example, Abdelgawad and Abdulhai (2012) developed an evacuation model with routing and scheduling of subway and bus transit to alleviate congestion pressure during the evacuation of busy urban areas. Wang et al. (2019a) proposed an optimal bus bridging design method under operational disruptions on a single metro line. Tan et al. (2020) 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 paper, the service disruption is not as severe as the emergency situation. 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 evacuation plans as assumed in evacuation papers. Second, in this study, we do not adjust the operation from the supply side. Instead, we focus on providing information to the passengers to better utilize the existing resources/capacities of the system.

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). But the waiting time is usually hard to calculate if the system is congested with left behind.

Passengers' travel time is usually modeled in the transit assignment literature, where two major frameworks exist: frequency-based (static) and schedule-based (dynamic). In the frequency-based transit assignment model, the waiting time is either assumed to be reversely proportional to the (effective) service frequency (Wu et al., 1994; Schmöcker et al., 2011; Nielsen, 2000), or modeled as a congestion function (e.g., BRP) of previously boarded flows and new arrival flows with exogenously-calibrated parameters (De Cea and Fernández, 1993). The former method does not consider the left behind, and the latter method 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 (Nguyen et al., 2001; Hamdouch and Lawphongpanich, 2008; Hamdouch et al., 2014; Schmöcker et al., 2008), the waiting time can only be obtained after a dynamic network loading (or simulation) process. For example, Schmöcker et al. (2008) used the fail-to-board probability to model the left behind behavior. This probability will be updated after finishing 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.

As 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. And given a tractable linear programming model, the uncertainties can also be incorporated.

2.4. Robust optimization

RO is a common approach to handle data uncertainty in optimization problems. The general approach is to specify a range for an uncertain parameter (the "uncertainty set"), and optimize over the worst-case realizations within the bounded uncertainty set. The method is therefore well suited to applications where there is considerable uncertainty related to the model input parameters, and when data uncertainties can lead to significant penalties or infeasibility in practice. The solution method for robust optimization problems involves generating a deterministic equivalent, called the robust counterpart. Computational tractability of the robust counterpart has been a major practical difficulty (Ben-Tal et al., 2009). A variety of uncertainty sets have been identified for which the robust counterpart to a robust optimization problem is reasonably tractable (Bertsimas et al., 2011).

The RO field has grown substantially over the past two decades. Seminal papers in the late 1990s (Ben-Tal and Nemirovski, 1998, 1999) and early 2000s (Bertsimas and Sim, 2004) established the field. Comprehensive surveys on the early literature were done by Ben-Tal et al. (2009) and Bertsimas et al. (2011). The development of the robust optimization technique has allowed researchers to tackle problems with data uncertainty in a range of fields. Examples can be found for renewable energy network design (Xiong et al.,

2016), supply chain operations (Ma et al., 2018), health care logistics (Wang et al., 2019b), and ride-hailing (Guo et al., 2021).

However, to the best of the authors' knowledge, no existing papers have incorporated robust optimization techniques into path recommendations during service disruptions. This research gap is important to address given the potential inaccurate estimates of demand in public transit networks during an incident.

3. Methodology

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 operation changes 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 operation changes. These changes define a new PT service network and available 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 system, inputting their **origin, destination, and departure time** to get a recommended path. And the recommendation aims to minimize the system travel time, that is, the sum of all passengers' travel time, 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).

Let $\mathcal{K}$ be the predetermined set of all OD pairs that may need path recommendation. $\mathcal{K}$ is defined based on whether passengers with the OD are affected by the incident or not. Note that as the path recommendation starts 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 in $\mathcal{K}$ and departure time in the range from T_s to some time after T_e, because the congestion may last longer than T_e, passengers depart after T_e may also need guidance. We divide the periods of recommendation into a time point (h_0) and several equal-length time intervals ($h_1, ..., h_H$). 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 $(T_s + (t-1)\tau, T_s + t\tau]$, where τ is the length of a time interval (e.g., 10 min). Recommendations at h_t $(t \geq 1)$ focus on passengers who have not entered the system when the incident happens 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 new operation after T_s, we can 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) 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 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 3.4. Let f_{hkr} be the number of passengers departing at time interval h, with OD pair k, and using path $r \in R_k$. By definition:

$$\sum_{r \in R_k} f_{hkr} = d_{hk} \quad \forall h \in \mathcal{H}, k \in \mathcal{K} \tag{1}$$

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 the 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 a set of all path indices. Then the optimal flow problem can be formulated as:

$$\min_{\boldsymbol{f}, \boldsymbol{p}} \quad Z(\boldsymbol{f}) = \text{Sum of all passengers' travel time} \tag{2a}$$

$$\text{s.t.} \quad \sum_{r \in R_k} p_{hkr} = 1 \qquad \forall\, h \in \mathcal{H}, k \in \mathcal{K}, \tag{2b}$$

$$f_{hkr} = d_{hk} \cdot p_{hkr} \quad \forall\, (h, k, r) \in \mathcal{F}, \tag{2c}$$

$$f_{hkr} \geq 0 \qquad \forall\, (h, k, r) \in \mathcal{F}, \tag{2d}$$

$$0 \leq p_{hkr} \leq 1 \qquad \forall\, (h, k, r) \in \mathcal{F} \tag{2e}$$

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 2.3). Note that using both of $\boldsymbol{f}$ and $\boldsymbol{p}$ in the optimization problem is redundant, but it is useful for the methodology explanation.

If there was no uncertainty in the system, the optimal path shares (p^*_{hkr}) solved from Eq. 2 are the recommendation proportion. That is, for all passengers who input OD pair k and departure time h, the system will recommend them to use path r with probability p^*_{hkr}. However, Eq. 2 is just a naive formulation. We cannot solve it 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.

3.2. Event-based public transit simulator

3.2.1. Simulator design

Before introducing the solution procedure for Eq. 2, we first describe an event-based public transit simulator used in this study (Mo et al., 2020). It can be used to evaluate $Z(\boldsymbol{f})$ and facilitate simulation-based linearization.

Figure 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, runs, 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. Two journeys that share a common line can be treated as two different paths. Hence, there is no "common line" problem (De Cea and Fernández, 1993) in this study.

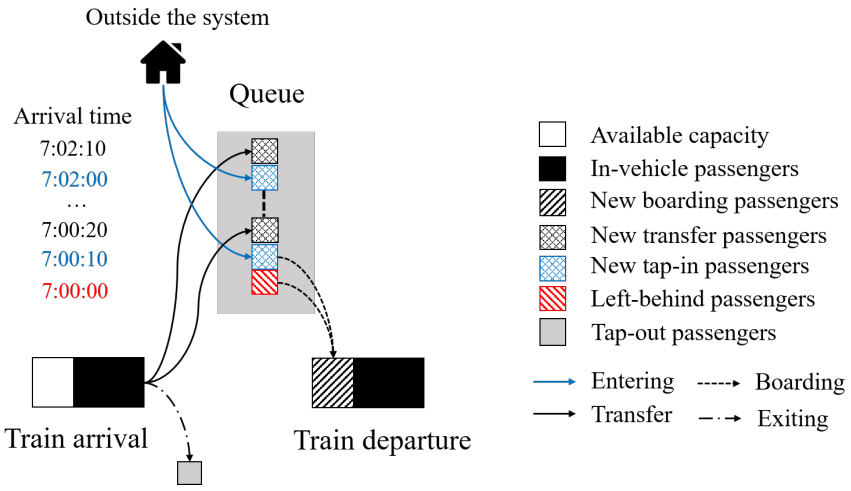


Figure 1: Structure of the network loading model (adapted from Mo et al. (2020)

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 $Z(\boldsymbol{f})$.

3.2.2. 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 seen as some vehicles in the line having long dwelling time 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 the service disruption without changing the

framework.

From the passenger side, when an incident happens, all passengers in the 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 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 $\boldsymbol{p}$).

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 the first-order approximation. Notice that $Z(\boldsymbol{f})$ can be approximated as:

$$\hat{Z}(\boldsymbol{f}) = Z(\tilde{\boldsymbol{f}}) + (\boldsymbol{f} - \tilde{\boldsymbol{f}})^T \frac{\partial Z(\boldsymbol{f})}{\partial \boldsymbol{f}}|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}} \tag{3}$$

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 inputs. $\frac{\partial Z(\boldsymbol{f})}{\partial \boldsymbol{f}} = (\frac{\partial Z(\boldsymbol{f})}{\partial f_{hkr}})_{h,k,r\in\mathcal{F}}$ 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:

$$\frac{\partial Z(\boldsymbol{f})}{\partial f_{hkr}}|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}} \approx \frac{Z(\tilde{\boldsymbol{f}} + \boldsymbol{e}_{hkr}) - Z(\tilde{\boldsymbol{f}})}{1} \tag{4}$$

where $\boldsymbol{e}_{hkr}$ represents a vector with only the (h,k,r)-th element being 1 and others being zero. Eq. 4 represents the numerical approximation of the gradient. Then we only need to calculate $Z(\tilde{\boldsymbol{f}} + \boldsymbol{e}_{hkr}) - Z(\tilde{\boldsymbol{f}})$. One of the naive methods 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. Since we already run a simulation for $\tilde{\boldsymbol{f}}$, it is possible to directly calculate the marginal change without running it again. Given the definition of the gradient, the problem is to 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 2. Let M_{hkr} be the set of passengers that counts the flow of $\tilde{f}_{hkr}$ (i.e., the orange passengers in Figure 2). These passengers have origin of station a_1 and destination of station a_7, and the path includes a transfer from station a_4 to station a_5. Denote the average travel time of $\tilde{f}_{hkr}$ as $T^{\mathrm{A}}_{hkr}(\tilde{\boldsymbol{f}})$. Suppose that we want to add one more passenger to $\tilde{f}_{hkr}$ (i.e., the orange passenger is duplicated in Figure 2). 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.

Besides, it may also affect other passengers' travel time. All passengers in the red-dashed square may be affected. Passengers at station a_1 and a_5 who queue behind the orange passenger may have additional waiting time if the vehicle that M_{hkr} used is full after departure (under the simulation results of $\tilde{\boldsymbol{f}}$), because the increase of one flow in $\tilde{f}_{hkr}$ will occupy the one capacity for these waiting passengers, and one of them have to board the next bus (i.e., wait for one more headway). Mathematically, let V^b_{hkr} be the set of vehicles that M_{hkr} board at station b, adding an additional passenger to M_{hkr} means we have one more passenger board in 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

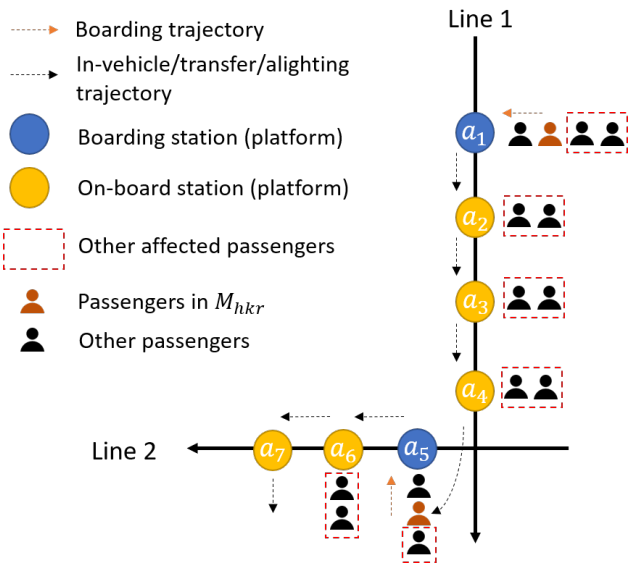


Figure 2: Explanation for the impact of adding additional one unit flow to the system

departure from station b, then the total increase in system travel time for passengers queuing behind M_{hkr} is:

$$T^{\mathrm{Q}}_{hkr}(\tilde{\boldsymbol{f}}) = \sum_{b \in B_{hkr}} \sum_{v \in V^b_{hkr}} \frac{\mathbb{1}_{\{\mathrm{Full}^b_v\}} \cdot W^b_v}{|V^b_{hkr}|} \tag{5}$$

where B_{hkr} is the set of all boarding stations of M_{hkr} (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), the total travel time increase is approximately two headways if the vehicles are full.

For passengers waiting at stations where M_{hkr} are 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 the on-board stations. And 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 flow 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:

$$T^{\mathrm{O}}_{hkr}(\tilde{\boldsymbol{f}}) = \sum_{b \in B_{hkr}} \sum_{v \in V^b_{hkr}} \frac{1}{|V^b_{hkr}|} \sum_{a \in O^v_{hkr}} \mathbb{1}_{\{\mathrm{Full}^a_v\}} \cdot W^a_v \tag{6}$$

Therefore, in this way, depending on whether the vehicle is full or not under flow pattern $\tilde{\boldsymbol{f}}$, we can calculate the increase in system travel time due to adding one flow to $\tilde{f}_{hkr}$ 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, 2) the additional waiting time for passengers queuing behind M_{hkr}, and 3) the additional waiting

time for passengers queuing at M_{hkr}'s on-board stations. Specifically, we have

$$Z(\tilde{\boldsymbol{f}} + \boldsymbol{e}_{hkr}) - Z(\tilde{\boldsymbol{f}}) = T^{\mathrm{A}}_{hkr}(\tilde{\boldsymbol{f}}) + T^{\mathrm{Q}}_{hkr}(\tilde{\boldsymbol{f}}) + T^{\mathrm{O}}_{hkr}(\tilde{\boldsymbol{f}}) \tag{7}$$

Consequently, $\frac{\partial Z(\boldsymbol{f})}{\partial \boldsymbol{f}}|_{\boldsymbol{f}=\tilde{\boldsymbol{f}}}$ can be obtained from Eq. 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:

$$\hat{Z}(\boldsymbol{f}) = Z(\tilde{\boldsymbol{f}}) + \boldsymbol{\beta}(\tilde{\boldsymbol{f}})^T(\boldsymbol{f} - \tilde{\boldsymbol{f}}) \tag{8}$$

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. 8 is a linear function of $\boldsymbol{f}$, which supports for addressing uncertainties in the optimization problem.

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 parts of uncertainties are considered as a whole by introducing an ellipsoidal uncertainty set and three polyhedral uncertainty sets.

From the constraint 2c, we can substitute $f_{hkr} = d_{hk} \cdot p_{hkr}$ to the objective function and rewriting Eq. 8 as:

$$\hat{Z}(\boldsymbol{f}) = \hat{Z}(\boldsymbol{p}) = Z(\tilde{\boldsymbol{f}}) + \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr} \cdot (d_{hk} \cdot p_{hkr} - \tilde{f}_{hkr}) \tag{9}$$

Note that β_{hkr} is a function of $\tilde{\boldsymbol{f}}$, for simplicity we ignore it 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:

$$\min_{\boldsymbol{p},\,t} \quad t \tag{10a}$$

$$\text{s.t.} \quad t \geq Z(\tilde{\boldsymbol{f}}) + \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr} \cdot (d_{hk} \cdot p_{hkr} - \tilde{f}_{hkr}), \tag{10b}$$

$$\text{Other constraints omitted} \tag{10c}$$

Constraint 10b can be rewritten as

$$\sum_{h,k} \sum_{r\in R_k} \beta_{hkr} \cdot d_{hk} \cdot p_{hkr} \leq t - Z(\tilde{\boldsymbol{f}}) + \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr}\tilde{f}_{hkr} \tag{11}$$

And Eq. 11 can be written as a matrix form:

$$\boldsymbol{a}^T\boldsymbol{p} \leq b \tag{12}$$

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 1. *If $\boldsymbol{d}$ is normally distributed with $\boldsymbol{d} \sim \mathcal{N}(\bar{\boldsymbol{d}}, \boldsymbol{\Sigma})$, then in a RO problem where constraint 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:

$$(\boldsymbol{A}\bar{\boldsymbol{d}} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{z})^T\boldsymbol{p} \leq b, \quad \forall \boldsymbol{z} \in \mathcal{Z}_E \tag{13}$$

where $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{F}|\times HK}$ *and the entry is defined as* $A_{hkr,h'k'} = \beta_{hkr}$ *if* $h = h'$ *and* $k = k'$*, otherwise* $A_{hkr,h'k'} = 0$*.* $\boldsymbol{D}$ *is the Cholesky decomposition of* $\boldsymbol{\Sigma}$ *(i.e.,* $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{D}^T$*).* $\boldsymbol{z}$ *is 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

$$\boldsymbol{a}^T\boldsymbol{p} \sim \mathcal{N}((\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p}, \boldsymbol{p}^T\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{p}) \tag{14}$$

If we want constraint 12 to hold with probability at least $1-\varepsilon$, it suffices to have:

$$(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \rho_{1-\varepsilon}\sqrt{\boldsymbol{p}^T\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{p}} \leq b \tag{15}$$

Substituting $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{D}^T$ into Eq. 15 finishes the proof of Step 1.

Step 2: We need to show that the robust counterpart of Eq. 13 is $(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \rho_{1-\varepsilon}\left\|(\boldsymbol{A}\boldsymbol{D})^T\boldsymbol{p}\right\|_2 \leq b$.

Notice that Eq. 13 is equivalent to:

$$(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \max_{\boldsymbol{z}\in\mathcal{Z}_{\mathrm{E}}}(\boldsymbol{A}\boldsymbol{D}\boldsymbol{z})^T\boldsymbol{p} \leq b. \tag{16}$$

Define the indicator function on set $\mathcal{Z}_{\mathrm{E}}$ as

$$\delta(\boldsymbol{z} \mid \mathcal{Z}_{\mathrm{E}}) = \begin{cases} 1, & \text{if } \boldsymbol{z} \in \mathcal{Z}_{\mathrm{E}} \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

Then the convex conjugate of $\delta(\boldsymbol{z} \mid \mathcal{Z}_{\mathrm{E}})$ (also known as the **support function**) can be derived as (Bertsimas and den Hertog, 2020):

$$\delta^*(\boldsymbol{y} \mid \mathcal{Z}_{\mathrm{E}}) = \sup_{\boldsymbol{z}\in\mathbb{R}^{HK}}\{\boldsymbol{y}^T\boldsymbol{z} - \delta(\boldsymbol{z} \mid \mathcal{Z}_{\mathrm{E}})\} = \sup_{\boldsymbol{z}\in\mathcal{Z}_{\mathrm{E}}}\boldsymbol{y}^T\boldsymbol{z} = \rho_{1-\varepsilon}\|\boldsymbol{y}\|_2 \tag{18}$$

Therefore, Eq. 16 can be rewritten with the convex conjugate:

$$(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \delta^*((\boldsymbol{A}\boldsymbol{D})^T\boldsymbol{p} \mid \mathcal{Z}) = (\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \rho_{1-\varepsilon}\left\|(\boldsymbol{A}\boldsymbol{D})^T\boldsymbol{p}\right\|_2 \leq b \tag{19}$$

which finishes the proof of Step 2. Combining step 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 1. 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 $\rho_{1-\varepsilon}$ could have a specific meaning and we can characterize the

probability for that constraint 12 holds. The use of multivariate normality assumption in Proposition 1 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 (Cain et al., 2017) shows that $\boldsymbol{d}$ has no significant skewness.

Eq. 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 to d_{hk}. In reality, the demand level for a specific OD pair and time interval is usually bounded, which can be expressed as:

$$d_{hk}^{\mathrm{L}} \leq d_{hk} \leq d_{hk}^{\mathrm{U}} \tag{20}$$

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. 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

$$\boldsymbol{d}^{\mathrm{L}} - \bar{\boldsymbol{d}} \leq \boldsymbol{D}\boldsymbol{z} \leq \boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}} \tag{21}$$

We can rewritten 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. 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. 20 cannot capture (e.g., all d_{hk} are at the lower or upper bounds). Mathematically:

$$d_h^{\mathrm{L}} \leq \sum_{k\in\mathcal{K}} d_{hk} \leq d_h^{\mathrm{U}} \tag{22}$$

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. 22 can be rewritten in a matrix form:

$$\boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}} - \boldsymbol{S}\bar{\boldsymbol{d}} \leq \boldsymbol{S}\boldsymbol{D}\boldsymbol{z} \leq \boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}} - \boldsymbol{S}\bar{\boldsymbol{d}} \tag{23}$$

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. 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 unreal as 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 captures critical OD pairs, we add an additional constraint on the total demand:

$$\sum_{h\in\mathcal{H},k\in\mathcal{K}} d_{hk} \leq \Gamma \cdot \sum_{h\in\mathcal{H},k\in\mathcal{K}} \bar{d}_{hk} \tag{24}$$

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, and the worst

case scenario will have more demand on critical OD pairs but less demand on others. The value of Γ can be decided by the historical highest total demand or by empirical knowledge.

Similarly, Eq. 24 can be written in a matrix form:

$$\mathbf{1}^T(\bar{\boldsymbol{d}} + \boldsymbol{D}\boldsymbol{z}) \leq \Gamma \cdot \mathbf{1}^T\bar{\boldsymbol{d}} \tag{25}$$

where $\mathbf{1} \in \mathbb{R}^{HK}$ is a vector with all elements one. And we define another polyhedral uncertainty set: $\mathcal{Z}_{\text{P3}} = \left\{\boldsymbol{z} \in \mathbb{R}^{HK} : \mathbf{1}^T(\bar{\boldsymbol{d}} + \boldsymbol{D}\boldsymbol{z}) \leq \Gamma \cdot \mathbf{1}^T\bar{\boldsymbol{d}}\right\}$.

Therefore, the final robust constraint for Eq. 12 is

$$(\boldsymbol{A}\bar{\boldsymbol{d}} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{z})^T\boldsymbol{p} \leq b, \quad \forall \boldsymbol{z} \in \mathcal{Z}_{\text{E}} \cap \mathcal{Z}_{\text{P}} \cap \mathcal{Z}_{\text{P2}} \cap \mathcal{Z}_{\text{P3}} \tag{26}$$

To derive the robust counterpart of the constraint, we first introduce the following lemma.

Lemma 1. *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,*

$$\delta^*(\boldsymbol{y} \mid \mathcal{Z}) = \min_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k} \{\sum_{i=1}^k \delta^*(\boldsymbol{y}_i \mid \mathcal{Z}_i) \mid \sum_{i=1}^k \boldsymbol{y}_i = \boldsymbol{y}\},$$

and the constraint becomes

$$\begin{cases} \bar{\boldsymbol{a}}^T\boldsymbol{x} + \sum_{i=1}^k \delta^*(\boldsymbol{y}_i \mid \mathcal{Z}_i) \leq b \\ \sum_{i=1}^k \boldsymbol{y}_i = \boldsymbol{P}^T\boldsymbol{x} \end{cases}$$

where $\delta^*(\cdot \mid \cdot)$ is the support function (i.e., convex conjugate of the indicator function).

The proof of Lemma 1 can be found in Ben-Tal et al. (2015). From Proposition 1, we have $\delta^*(\boldsymbol{y} \mid \mathcal{Z}_{\text{E}}) = \rho_{1-\varepsilon} \|\boldsymbol{y}\|_2$. 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

$$\delta^*(\boldsymbol{y} \mid \mathcal{Z}_{\text{P}}) = \max_{\boldsymbol{z}}\{\boldsymbol{y}^T\boldsymbol{z} \mid \boldsymbol{H}\boldsymbol{z} \leq \boldsymbol{c}\} = \min_{\boldsymbol{u}}\{\boldsymbol{c}^T\boldsymbol{u} \mid \boldsymbol{H}^T\boldsymbol{u} = \boldsymbol{y}, \boldsymbol{u} \geq 0\} \tag{27}$$

where the second equality follows from linear programming duality. Eq. 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. 24, we have

$$\delta^*(\boldsymbol{y}_6 \mid \mathcal{Z}_{\text{P3}}) = \min_{u_3}\{(\Gamma - 1) \cdot (\mathbf{1}^T\bar{\boldsymbol{d}}) \cdot u_3 \mid (\mathbf{1}^T\boldsymbol{D})^T u_3 = \boldsymbol{y}_6, u_3 \geq 0\} \tag{28}$$

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. 29.

[1] $ri(\mathcal{Z}_i)$ indicates the relative interior of the set $\mathcal{Z}_i$.

Based on Lemma 1, the robust counterpart for Eq. 26 is

$$
\begin{aligned}
&(\boldsymbol{A}\bar{\boldsymbol{d}})^T\boldsymbol{p} + \rho_{1-\varepsilon}\left\|\boldsymbol{y}_1\right\|_2 + (\boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}})^T\boldsymbol{u}_1 + (\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}})^T\boldsymbol{u}_2 + (\boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T\boldsymbol{v}_1 + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}})^T\boldsymbol{v}_2 \\
&+ (\Gamma - 1)\cdot(\boldsymbol{1}^T\bar{\boldsymbol{d}})\cdot u_3 \leq b && \text{(29a)}\\
&\boldsymbol{D}^T\boldsymbol{u}_1 = \boldsymbol{y}_2 && \text{(29b)}\\
&-\boldsymbol{D}^T\boldsymbol{u}_2 = \boldsymbol{y}_3 && \text{(29c)}\\
&(\boldsymbol{S}\boldsymbol{D})^T\boldsymbol{v}_1 = \boldsymbol{y}_4 && \text{(29d)}\\
&-(\boldsymbol{S}\boldsymbol{D})^T\boldsymbol{v}_2 = \boldsymbol{y}_5 && \text{(29e)}\\
&(\boldsymbol{1}^T\boldsymbol{D})^T u_3 = \boldsymbol{y}_6 && \text{(29f)}\\
&\sum_{i=1}^{6}\boldsymbol{y}_i = (\boldsymbol{A}\boldsymbol{D})^T\boldsymbol{p} && \text{(29g)}\\
&\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{v}_1, \boldsymbol{v}_2, u_3 \geq 0 && \text{(29h)}
\end{aligned}
$$

Hence, the RO problem can be formulated as

$$
\begin{aligned}
\min_{\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y},t}\quad & t && \text{(30a)}\\
\text{s.t.}\quad & \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr}\cdot d_{hk}\cdot p_{hkr} + \rho_{1-\varepsilon}\left\|\boldsymbol{y}_1\right\|_2 + (\boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}})^T\boldsymbol{u}_1 + (\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}})^T\boldsymbol{u}_2 + (\boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T\boldsymbol{v}_1 \\
& + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}})^T\boldsymbol{v}_2 + (\Gamma - 1)\cdot(\boldsymbol{1}^T\bar{\boldsymbol{d}})\cdot u_3 + Z(\tilde{\boldsymbol{f}}) - \sum_{(h,k,r)\in\mathcal{F}}\beta_{hkr}\tilde{f}_{hkr} \leq t && \text{(30b)}\\
& \text{Constraints } (29b) - (29h) && \text{(30c)}\\
& \text{Other constraints omitted} && \text{(30d)}
\end{aligned}
$$

Eliminate t and put constraint 30b back to the objective function:

$$
\begin{aligned}
\hat{Z}(\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y})^{\mathrm{RC}} = & \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr}\cdot(d_{hk}\cdot p_{hkr} - \tilde{f}_{hkr}) + \rho_{1-\varepsilon}\left\|\boldsymbol{y}_1\right\|_2 + (\boldsymbol{d}^{\mathrm{U}} - \bar{\boldsymbol{d}})^T\boldsymbol{u}_1 + (\bar{\boldsymbol{d}} - \boldsymbol{d}^{\mathrm{L}})^T\boldsymbol{u}_2 \\
& + (\boldsymbol{d}_{\mathcal{H}}^{\mathrm{U}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T\boldsymbol{v}_1 + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}_{\mathcal{H}}^{\mathrm{L}})^T\boldsymbol{v}_2 + (\Gamma - 1)\cdot(\boldsymbol{1}^T\bar{\boldsymbol{d}})\cdot u_3 + Z(\tilde{\boldsymbol{f}}) \qquad \text{(31)}
\end{aligned}
$$

which yields a second-order cone programming (SOCP).

3.5. Solution procedure

After incorporating the demand uncertainty, the final robust counterpart (RC) of the optimal flow problem can be formulated as:

$$[RC(\tilde{\boldsymbol{f}})] \quad \min_{\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y}} \quad \hat{Z}(\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y})^{\text{RC}} = \sum_{(h,k,r)\in\mathcal{F}} \beta_{hkr}(\tilde{\boldsymbol{f}}) \cdot (d_{hk} \cdot p_{hkr} - \tilde{f}_{hkr}) + \rho_{1-\varepsilon} \|\boldsymbol{y}_1\|_2 + (\boldsymbol{d}^{\text{U}} - \bar{\boldsymbol{d}})^T \boldsymbol{u}_1 + (\bar{\boldsymbol{d}} - \boldsymbol{d}^{\text{L}})^T \boldsymbol{u}_2 + (\boldsymbol{d}^{\text{U}}_{\mathcal{H}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T \boldsymbol{v}_1 + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}^{\text{L}}_{\mathcal{H}})^T \boldsymbol{v}_2 + (\Gamma - 1) \cdot (\boldsymbol{1}^T \bar{\boldsymbol{d}}) \cdot u_3 + Z(\tilde{\boldsymbol{f}}) \tag{32a}$$

$$\text{s.t.} \quad \text{Constraints } (29b) - (29h) \tag{32b}$$

$$\sum_{r\in R_k} p_{hkr} = 1 \quad \forall h \in \mathcal{H}, k \in \mathcal{K} \tag{32c}$$

$$0 \le p_{hkr} \le 1 \quad \forall (h,k,r) \in \mathcal{F} \tag{32d}$$

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, we need to run the simulation again to update $\beta_{hkr}(\tilde{\boldsymbol{f}})$. Before that, we need to get the corresponding worst-case demand (WD), which will be used as the new $\tilde{\boldsymbol{f}}$. It can be obtained by solving the worst case $\boldsymbol{z} \in \mathcal{Z}_{\text{E}} \cap \mathcal{Z}_{\text{P1}} \cap \mathcal{Z}_{\text{P2}} \cap \mathcal{Z}_{\text{P3}}$:

$$[WD(\boldsymbol{p}^*)] \quad \max_{\boldsymbol{z}} \quad (\boldsymbol{A}\boldsymbol{D}\boldsymbol{z})^T \boldsymbol{p}^* \tag{33a}$$

$$\text{s.t.} \quad \|\boldsymbol{z}\|_2 \le \rho_{1-\varepsilon} \tag{33b}$$

$$\boldsymbol{d}^{\text{L}} - \bar{\boldsymbol{d}} \le \boldsymbol{D}\boldsymbol{z} \le \boldsymbol{d}^{\text{U}} - \bar{\boldsymbol{d}} \tag{33c}$$

$$\boldsymbol{d}^{\text{L}}_{\mathcal{H}} - \boldsymbol{S}\bar{\boldsymbol{d}} \le \boldsymbol{S}\boldsymbol{D}\boldsymbol{z} \le \boldsymbol{d}^{\text{U}}_{\mathcal{H}} - \boldsymbol{S}\bar{\boldsymbol{d}} \tag{33d}$$

$$\boldsymbol{1}^T(\bar{\boldsymbol{d}} + \boldsymbol{D}\boldsymbol{z}) \le \Gamma \cdot \boldsymbol{1}^T \bar{\boldsymbol{d}} \tag{33e}$$

Denote the solution for Eq. 33 as $\boldsymbol{z}^*$. Then, 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

$$Z(\tilde{\boldsymbol{f}}), \boldsymbol{\beta}(\tilde{\boldsymbol{f}}) = \textsc{Sim-FOA}(\boldsymbol{d}^*, \boldsymbol{p}^*) \tag{34}$$

where $\tilde{\boldsymbol{f}}$ in Eq. 34 indicates $\tilde{f}_{hkr} = d^*_{hk} \cdot p^*_{hkr}$. And SIM-FOA$(\cdot)$ is a pseudo function of simulation plus first-order approximation as described in Section 3.3. The RC, WD, and 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 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 criteria 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 increasing by one or more headways. In fact, we have the following example showing that $\beta_{hkr}(\tilde{\boldsymbol{f}})$ can be arbitrarily large, this may cause the criteria 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 3) and a fixed headway W. Assume every bus has a capacity of 1. And there is one passenger waiting at each station except for the first station (i.e., there are $N-1$ waiting passengers). Now suppose that we add one more passenger to station 1. Since the capacity of buses is 1, the newly added bus 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 new-added passenger can be arbitrarily large depending on the number of stations N.*

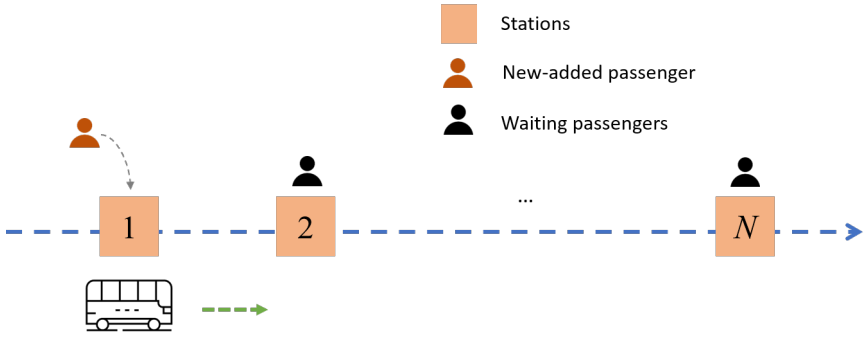


Figure 3: Example for arbitrarily large $\beta_{hkr}(\tilde{\boldsymbol{f}})$

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, the MSA will converge if

$$\left| Z(\tilde{\boldsymbol{f}})^{(t)} - \frac{1}{T^{\text{Cvg}}} \sum_{t'=t-T^{\text{Cvg}}}^{t-1} Z(\tilde{\boldsymbol{f}})^{(t')} \right| \leq \epsilon \tag{35}$$

where $Z(\tilde{\boldsymbol{f}})^{(t)}$ is the system travel time at t-th iteration and ϵ is a predetermined threshold. Eq. 35 means that when the current system travel time is close to its average value of the last T^{Cvg} iterations, the algorithm converges. Taking the average of the last T^{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 1. 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 $T^{\text{Cvg}}+1$ iterations.

Let $\boldsymbol{p}^*$ be the optimal path shares solved from Algorithm 1. To realize the optimal path shares in the real world, the following system design can be used:

- Step 1: Transit operators deploy the recommendation system to smartphone apps, websites, and electrical screens at stations.
- Step 2: Passengers, when using the system, need to input their origins, destinations, and departure times.
- Step 3: For a passenger input OD pair k and departure time h, the system will return a single recommended path r to him/her with probability p^*_{hkr}.

Algorithm 1 Solution procedure of the robust optimal flow problem

Initialize $\boldsymbol{p}^{(0)}$, $\boldsymbol{d}^{(0)}$ and specify T^{Cvg}, ϵ.
Set iteration counter $t = 0$.
do
  $Z(\tilde{\boldsymbol{f}})^{(t)}, \boldsymbol{\beta}(\tilde{\boldsymbol{f}})^{(t)} = \text{SIM-FOA}(\boldsymbol{d}^{(t)}, \boldsymbol{p}^{(t)})$
  Solve the RC problem (Eq. 32) with $Z(\tilde{\boldsymbol{f}})^{(t)}$ and $\boldsymbol{\beta}(\tilde{\boldsymbol{f}})^{(t)}$ as inputs, and return $\hat{\boldsymbol{p}}^{(t+1)}$
  $\boldsymbol{p}^{(t+1)} = \frac{1}{t+1}\hat{\boldsymbol{p}}^{(t+1)} + (1 - \frac{1}{t+1})\boldsymbol{p}^{(t)}$
  Solve the WD problem (Eq. 33) with $\boldsymbol{p}^{(t+1)}$ as input and return $\boldsymbol{d}^{(t+1)}$
  $t = t + 1$
while $t \leq T^{\text{Cvg}}$ or $\left| Z(\tilde{\boldsymbol{f}})^{(t)} - \frac{1}{T^{\text{Cvg}}} \sum_{t'=t-T^{\text{Cvg}}}^{t-1} Z(\tilde{\boldsymbol{f}})^{(t')} \right| > \epsilon$
$t^* = \arg\min_{t'=t-T^{\text{Cvg}},\ldots,t} Z(\tilde{\boldsymbol{f}})^{(t')}$
return $\boldsymbol{p}^{(t^*)}$

In this way, we may expect the final path flows are close to the system optimal path flows if passengers follow the recommendation.

4. Model extension discussion

4.1. Solving the model with rolling horizon

The current model development 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 incident period and after ($[h_0, h_H]$). In a real-world implementation, the model can be solved with a rolling horizon manner.

Specifically, at time interval $\tilde{h}$, we first update the demand and supply information in the system, 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 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 help with better model performance (this is known as adaptive robust optimization). Note that optimal path shares with $h > \tilde{h}$ are solved but not utilized in the rolling horizon manner. Then, when the time goes to $\tilde{h} + 1$, we can set $\tilde{h} = \tilde{h} + 1$ and solve the problem again.

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

[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.

for the RO problem becomes:

$$\mathbb{E}[\hat{Z}(\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y})^{\text{RC}}] = \sum_{\xi\in\Omega}\mathbb{P}(\xi)\left[Z(\tilde{\boldsymbol{f}};\xi) + \sum_{(h,k,r)\in\mathcal{F}}\beta_{hkr}(\tilde{\boldsymbol{f}};\xi)\cdot(d_{hk}\cdot p_{hkr} - \tilde{f}_{hkr})\right] + \rho_{1-\varepsilon}\left\|\boldsymbol{y}_1\right\|_2 + (\boldsymbol{d}^{\text{U}} - \bar{\boldsymbol{d}})^T\boldsymbol{u}_1 + (\bar{\boldsymbol{d}} - \boldsymbol{d}^{\text{L}})^T\boldsymbol{u}_2 + (\boldsymbol{d}^{\text{U}}_{\mathcal{H}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T\boldsymbol{v}_1 + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}^{\text{L}}_{\mathcal{H}})^T\boldsymbol{v}_2 + (\Gamma-1)\cdot(\mathbf{1}^T\bar{\boldsymbol{d}})\cdot u_3 \tag{36}$$

where $\mathbb{P}(\xi)$ is the probability of scenario ξ happens. 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 $\beta_{hkr}(\tilde{\boldsymbol{f}};\Omega) := \sum_{\xi\in\Omega}\mathbb{P}(\xi)\beta_{hkr}(\tilde{\boldsymbol{f}};\xi)$, and subsitute them into, the objective function becomes

$$\mathbb{E}[\hat{Z}(\boldsymbol{p},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y})^{\text{RC}}] = \sum_{(h,k,r)\in\mathcal{F}}\beta_{hkr}(\tilde{\boldsymbol{f}};\Omega)\cdot(d_{hk}\cdot p_{hkr} - \tilde{f}_{hkr}) + \rho_{1-\varepsilon}\left\|\boldsymbol{y}_1\right\|_2 + (\boldsymbol{d}^{\text{U}} - \bar{\boldsymbol{d}})^T\boldsymbol{u}_1 + (\bar{\boldsymbol{d}} - \boldsymbol{d}^{\text{L}})^T\boldsymbol{u}_2 + (\boldsymbol{d}^{\text{U}}_{\mathcal{H}} - \boldsymbol{S}\bar{\boldsymbol{d}})^T\boldsymbol{v}_1 + (\boldsymbol{S}\bar{\boldsymbol{d}} - \boldsymbol{d}^{\text{L}}_{\mathcal{H}})^T\boldsymbol{v}_2 + (\Gamma-1)\cdot(\mathbf{1}^T\bar{\boldsymbol{d}})\cdot u_3 + Z(\tilde{\boldsymbol{f}};\Omega) \tag{37}$$

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 us to change the objective function.

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 4). The incident starts at 8:14 AM and ends at 9:13 AM on Feb 1st, 2019 due to infrastructures issues between Harlem and Jefferson Park stations. And the whole blue line is suspended. During the disruption, the destination for most of the passengers is the "Loop", which is the CBD area in Chicago. Usually, there are four paths leading to the Loop: 1) using blue line (i.e., waiting for the system to recovery), 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 Brown Line. Based on the service structure, we can construct the route sets R_k for each OD pair k.

5.1. Parameter setting

$\mathcal{K}$ is set as all OD pairs with origins at the Blue line and destinations at the Loop. Time interval is set as $\tau = 10$ mins. The largest period with recommendation is $h_H = 10$, corresponding to 9:44 - 9:54 AM (i.e., 50 minutes after the end of the incident). In this study, we assume the incident duration is known or can be reasonably estimated. An alternative way to capture the incident duration uncertainty is stochastic optimization as discussed in Section 4.2. The factor of total demand level, Γ, is set as 1.1, which is the 90% percentile of the total demand distribution.

5.2. Quantification of uncertainty sets

To quantify the demand uncertainty, we need to specify the nominal demand $\bar{\boldsymbol{d}}$, and covariance matrix $\boldsymbol{\Sigma}$ (which can be used to get $\boldsymbol{D}$), and upper and lower bounds for demand (i.e., $\boldsymbol{d}^{\text{U}}$, $\boldsymbol{d}^{\text{L}}$, $\boldsymbol{d}^{\text{U}}_{\mathcal{H}}$, $\boldsymbol{d}^{\text{L}}_{\mathcal{H}}$). These can be estimated from historical demand. However as the demand on the incident day is smaller than usual given some passengers may leave the system, we cannot directly use normal day smart card data as historical demand. One possible solution is to find previous days with similar incidents and use demand in those days

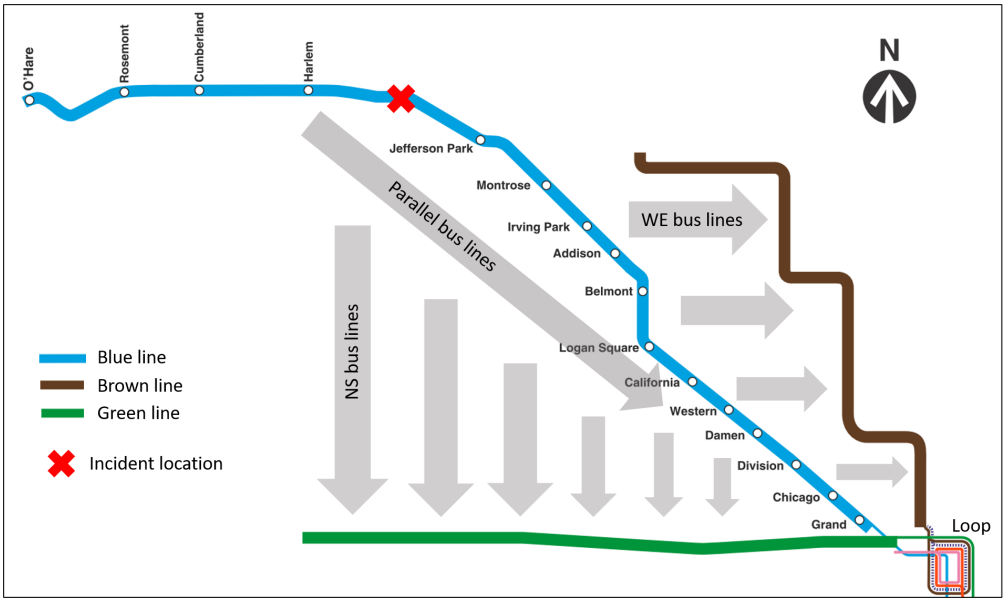


Figure 4: Case study diagram

as samples. But 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.

Recall that the demand uncertainty comes from two parts: 1) the inherent demand variations across different days 2) and the uncertainty in how many passengers left the PT system during the incident. The first part can be captured by historical smart card data (without incidents). And the second part is approximated by the survey results. According to previous survey-based studies, the proportion of the number of passengers leaving the PT system during incidents is around 10%∼30% (Lin et al., 2016; Rahimi et al., 2020). The "synthetic historical demand" is generated as follows:

- Collect smart card data in 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), randomly draw a proportion of leaving passengers from the uniform distribution $\mathcal{U}(0.1, 0.3)$. The demand after removing the leaving passengers is a sample demand vector.

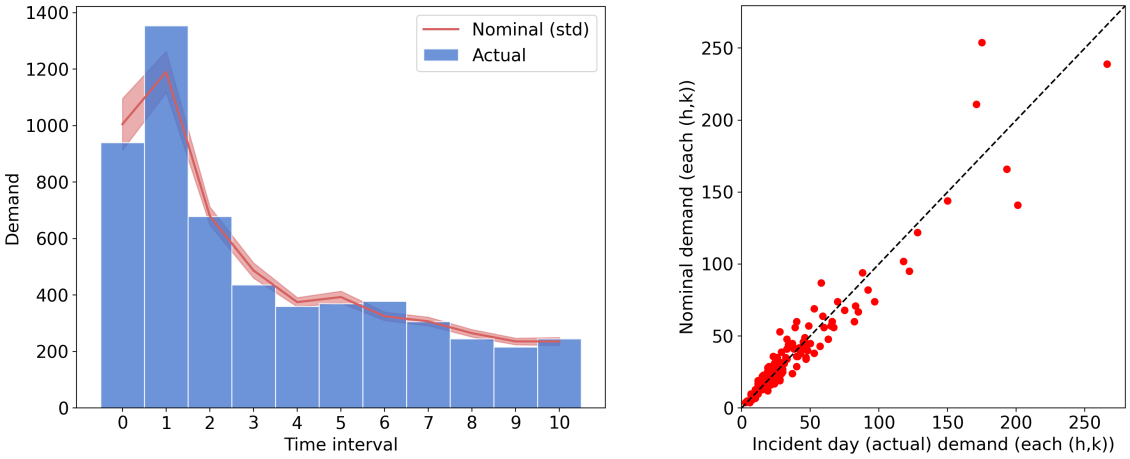
We collected a total of 16 work days from Jan 2019 (the previous month of the incident day) and generated 16 sample demand vectors. The mean value is used as the nominal demand $\bar{\boldsymbol{d}}$ and the co-variance matrix $\boldsymbol{\Sigma}$ are estimated from these samples. The upper and lower bounds for demand (i.e., $\boldsymbol{d}^{\mathrm{U}}$, $\boldsymbol{d}^{\mathrm{L}}$, $\boldsymbol{d}^{\mathrm{U}}_{\mathcal{H}}$, $\boldsymbol{d}^{\mathrm{L}}_{\mathcal{H}}$) 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 $\{50, 60, 70, 80, 90, 95, 99\}$ percentiles of the standard normal distribution. Note that $\rho_{1-\varepsilon} = 0$ equals no uncertainty (i.e., nominal model).

Note that the quantification of demand uncertainty sets does not use any information on the incident day and it is applicable for the real-world incident. Therefore, even if using synthetic historical demand may introduce errors, as long as we show that this approach of uncertainty set definition yields reasonably good results, it is acceptable for real-world applications.

5.3. *Data description*

The nominal and actual (incident day) demands comparison are shown in Figure 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 in 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 disaggregated demand for each (h, k), the differences are more prominent. The discrepancy between nominal and actual demands provides potentials for the RO to perform better.



(a) Total demand for each time interval h

(b) Demand comparison for each (h, k)

Figure 5: Demand patterns

Table 1 shows the results of Mardia test of multivariate normality (Cain et al., 2017) for demand samples. The Mardia test is used to check whether the samples' 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 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, Mardia Skewness testing shows the demand distribution is symmetric. Hence, it is still reasonable to use the ellipsoidal uncertainty set to describe a symmetric distributed uncertain variable. Morever, as mentioned in Remark 1, the distribution of a variable does not affect the definition of uncertainty set (it only affects the calculation of probability guarantees).

Table 1: Mardia test of multivariate normality

Test	p-value	Test	p-value
Mardia Skewness	1.00	Mardia Kurtosis	0.00

Note: Null hypothesis is that the samples are multivariate normally distributed. The smaller p-value indicates we are more likely to reject the Null hypothesis.

5.4. Benchmark models

To compare with the optimization-based path recommendations, the following path shares are used as benchmarks.

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 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 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. For example, for a path consisting of an NS bus route and the Green Line, the path capacity 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 new operation schedules on the incident day (i.e., the service suspension is considered). When no vehicles operate in the Blue line, the path capacity will be zero.

Status-quo path shares. The status-quo path shares are the inferred path choices of passengers on the incident day. Since all passengers using WE, NS and parallel bus lines need to tap in, the demand increases (i.e., the number of incident day tap-ins minus the mean number of normal days tap-ins) in these lines 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 is not directly recognizable because the CTA system does not record the tap-out information. Hence, we approximate the proportion of waiting passengers based on the survey results (Rahimi et al., 2019). Rahimi et al. (2019) 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, obtain the remaining system recovery time if passengers choose to wait, based on which we can derive the proportion of waiting passengers using the results in (Rahimi et al., 2019).
- Step 2: For each OD pair k and time interval h, get the number of passengers using the parallel bus, NS+Green, and WE+Brown paths based on demand increase.
- Step 3: Fix the proportion of waiting passengers obtained from Step 1, other path shares are proportional to the number of passengers using the associated paths obtained in Step 2.

6. Results

In this section, we demonstrate the model's performance in two steps. In the first step, we compare the optimization model without uncertainty (i.e., the nominal model with $\rho_{1-\epsilon} = 0$) with three benchmark path shares. This shows the improvement obtained from the optimization-based path recommendation. And in the second step, we compare the robust model with the nominal model, which shows the improvement from considering uncertainties.

6.1. Model convergence

Figure 6 shows the convergence of the nominal model ($\rho_{1-\epsilon} = 0$) and a robust model ($\rho_{1-\epsilon} = 0.84$). We observe that simulation-based linearization and MSA can successfully decrease the system travel time. And the model converges within 35 iterations. Note that the converged cost for the robust model is higher than the nominal model because the robust model is running with the worst-case demand (by definition with higher system travel time). The performance of all path recommendations will be evaluated on the actual demand (discussed in the next section).

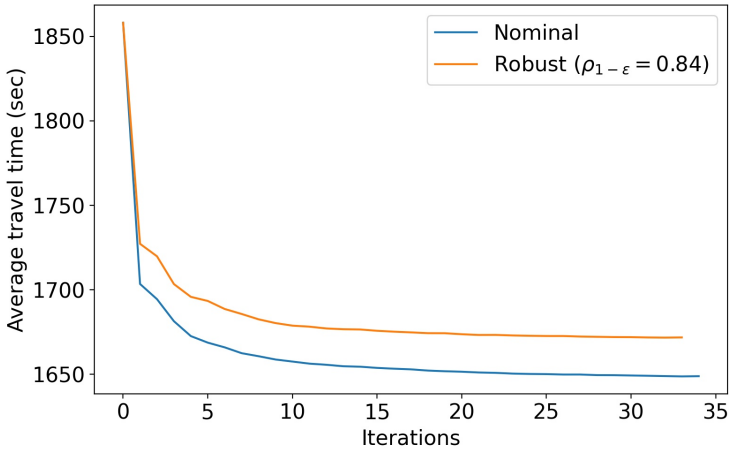


Figure 6: Convergence of optimization models

6.2. Model evaluation

The optimization model only utilizes the information of nominal demand and 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 will be evaluated based on the actual incident day demand using the simulation model. The model can output the travel time of every passenger in the system. And it can be used to compare the performances of different path shares.

6.3. Nominal model vs. Benchmark models

Table 2 compares the results of different path shares. The average travel times are calculated over all passengers (a total of 27,007 passengers) and the passengers who originally want to use the Blue line (i.e., passengers who are provided with recommendations, a total of 5,531 passengers). Results show that the optimization-based path shares can 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 incident-line passengers, the reduction is even higher (20.6%).

We also observe that the uniform path shares are worse than the status quo scenario, meaning that current passengers' choices are not random and show some rationality. The capacity-based path shares can also significantly reduce the system travel time (by 6.9%). However, as the capacity-based path recommendations do not reflect the boarding order of upstream and downstream stations, it is worse than the optimization-based results.

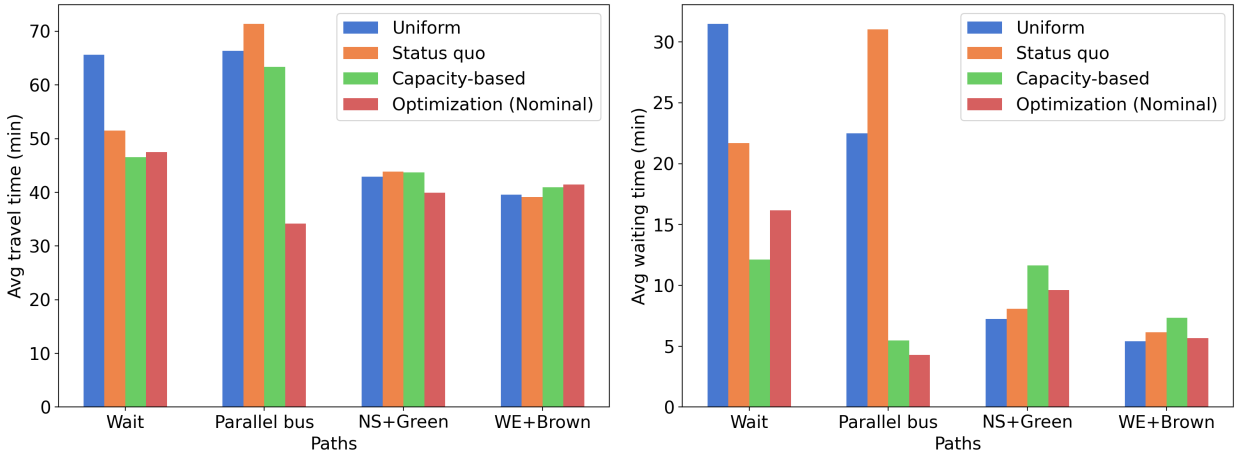
Figure 7 shows the average travel time and waiting time for different paths. We observe that the optimization-based path shares have more evenly travel time across four types of paths, implying a better

Table 2: Average travel time comparison

Path shares	All passengers (# 27,003)		Incident-line passengers (# 5,531)	
	Avg travel time (min)	% change[1]	Avg travel time (min)	% change[1]
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

utilization of the system's capacity. However, for other path shares, passengers using parallel buses have significantly longer travel times than those using other paths. Looking at Figure 7, 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 7: Comparison of average travel time and waiting time

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 8 shows the individual travel time change distribution (optimization-based minus the status quo) for all passengers with non-zero changes. We observe that, though most of the passengers have decreased travel time, some passengers become worse-off after following the path recommendation. This is a typical drawback of system optimal (first-best) assignment (Lawphongpanich and Yin, 2010). 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.

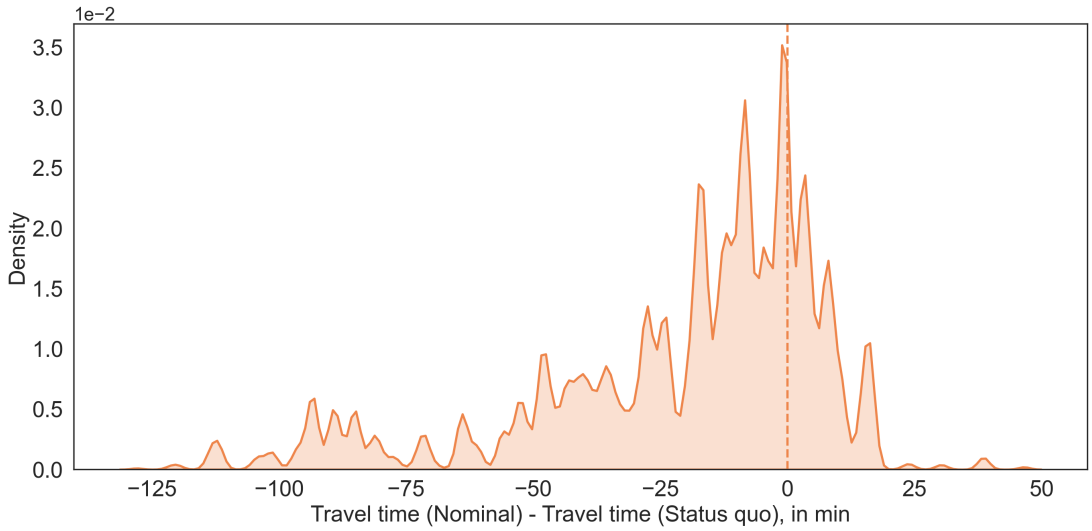


Figure 8: Distribution individual travel time changes (not including passengers without changes as they will distort the distribution with too much density concentrated at zero)

6.4. *Robust models vs. Nominal model*

In this section, we compare the results of RO with different values of $\rho_{1-\epsilon}$. Figure 9 shows the average travel time of robust models. We observe that, except for $\rho_{1-\epsilon} = 2.33$, all other values show better performance than the nominal model. This implies that considering the demand uncertainty can further increase the path recommendation strategies. The best value is $\rho_{1-\epsilon} = 0.84$, where the travel time for 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%.

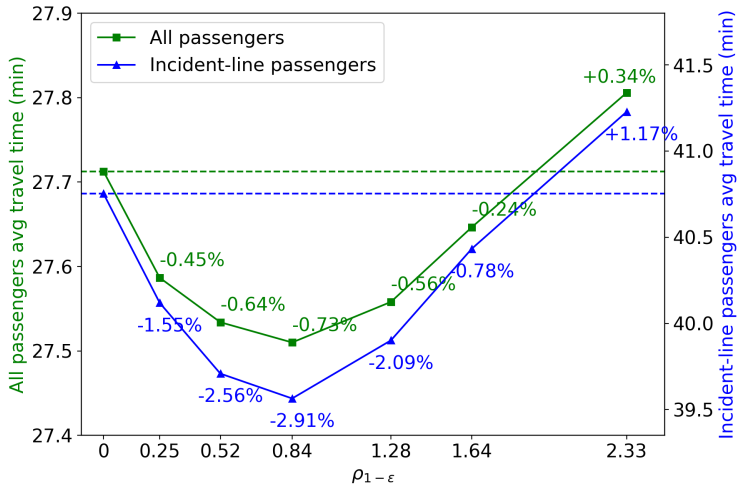


Figure 9: Performance of RO. The percentage changes are compared to the nominal scenario

Note that $\rho_{1-\epsilon} = 2.33$ indicates the largest uncertainty set compared to other values, under which the

worst-case demand patterns may deviate from the actual demand too much, thus performing worse than the nominal model. To validate this, we plot the worst-case demand for different values of $\rho_{1-\epsilon}$ in Figure 10. It is found that the worst-case demands for $\rho_{1-\epsilon} = 0.52, 0.84, 1.28$ scenarios are closer to the actual demand, and $\rho_{1-\epsilon} = 2.33$ overestimate the demands for $h = 0, 1$. These results are consistent with the travel time performance in Figure 9.

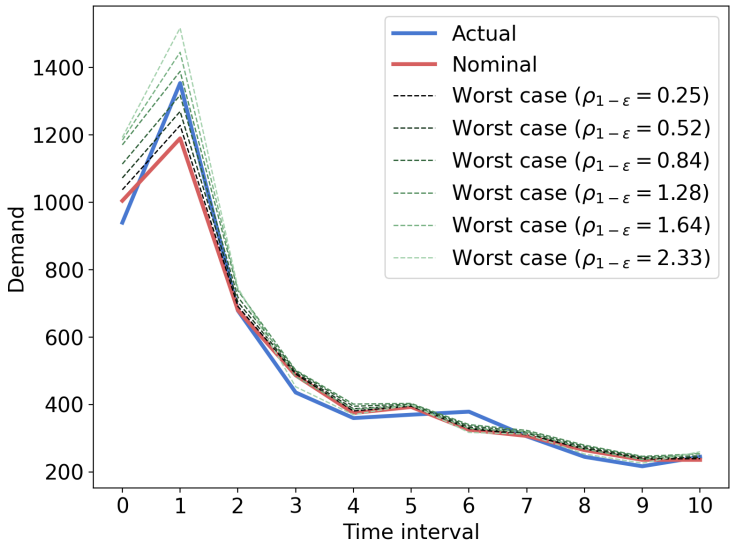


Figure 10: Worst-case demand patterns

7. Conclusion and discussion

In this paper, we propose a path recommendation model to mitigate the congestion during public transit disruptions. Passengers with different ODs and departure times are recommended with different paths such that the 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 linear programming and solve the new problem iteratively with MSA. Uncertainties in demand are modeled with 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 can further reduce the system travel time. The best robust model with $\rho_{1-\epsilon} = 0.84$ can decrease the average travel time of incident-line passengers by 2.91% compared to the nominal model.

The performance increase 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). The path recommendation patterns at nominal and worst-case demand may not be so different as they are both high demand for the system. 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 tasks.

Though we discussed the potential model extension with rolling horizon and incident duration uncertainty, we did not implement these extensions in the case study due to the complexity in updating inputs at the simulation environment. Incorporating real-time information as an adaptive RO would generally increase the model's performance (Bertsimas et al., 2011). This can be done for future related models with easily adjustable inputs.

Future research can be explored from the following directions. 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 in many previous RO problems (Bertsimas et al., 2012; Guo et al., 2021). Future studies may develop some data-driven uncertainty quantification methods to automate the hyperparameter tuning process. 2) As shown in Figure 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 ignore the potential non-compliance problem. That is, passengers may not follow the recommendation and the actual path flows may deviate from the expectation. Future research may develop new modeling techniques for path recommendations to 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.

8. Acknowledgement

The authors would like to thank Chicago Transit Authority (CTA) for their support and data availability for this research.

References

Abdelgawad, H., Abdulhai, B., 2012. Large-scale evacuation using subway and bus transit: approach and application in city of toronto. Journal of Transportation Engineering 138, 1215–1232.

Ben-Tal, A., Den Hertog, D., Vial, J.P., 2015. Deriving robust counterparts of nonlinear uncertain inequalities. Mathematical programming 149, 265–299.

Ben-Tal, A., El Ghaoui, L., Nemirovski, A., 2009. Robust Optimization. Princeton Series in Applied Mathematics, Princeton University Press.

Ben-Tal, A., Nemirovski, A., 1998. Robust convex optimization. Mathematics of Operations Research 23, 769–805.

Ben-Tal, A., Nemirovski, A., 1999. Robust solutions of uncertain linear programs. Operations Research Letters 25, 1–13.

Bertsimas, D., Brown, D.B., Caramanis, C., 2011. Theory and applications of robust optimization. SIAM Review 53, 464–501.

Bertsimas, D., den Hertog, D., 2020. Robust and adaptive optimization. Dynamic Ideas LLC, Belmont, Massachusetts.

Bertsimas, D., Litvinov, E., Sun, X.A., Zhao, J., Zheng, T., 2012. Adaptive robust optimization for the security constrained unit commitment problem. IEEE transactions on power systems 28, 52–63.

Bertsimas, D., Sim, M., 2004. The price of robustness. Operations Research 52, 35–53.

Böhmová, K., Mihalák, M., Pröger, T., Srámek, R., Widmayer, P., 2013. 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, Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik. pp. 27–41.

Bruglieri, M., Bruschi, F., Colorni, A., Luè, A., Nocerino, R., Rana, V., 2015. A real-time information system for public transport in case of delays and service disruptions. Transportation Research Procedia 10, 493–502.

Cain, M.K., Zhang, Z., Yuan, K.H., 2017. Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. Behavior research methods 49, 1716–1735.

Chen, S., Di, Y., Liu, S., Wang, B., 2017. Modelling and analysis on emergency evacuation from metro stations. Mathematical Problems in Engineering 2017.

Cox, A., Prager, F., Rose, A., 2011. Transportation security and the role of resilience: A foundation for operational metrics. Transport policy 18, 307–317.

De Cea, J., Fernández, E., 1993. Transit assignment for congested public transport systems: an equilibrium model. Transportation science 27, 133–147.

Guo, X., Caros, N.S., Zhao, J., 2021. Robust matching-integrated vehicle rebalancing in ride-hailing system with uncertain demand. Transportation Research Part B: Methodological 150, 161–189.

Hamdouch, Y., Lawphongpanich, S., 2008. Schedule-based transit assignment model with travel strategies and capacity constraints. Transportation Research Part B: Methodological 42, 663–684.

Hamdouch, Y., Szeto, W., Jiang, Y., 2014. A new schedule-based transit assignment model with travel strategies and supply uncertainties. Transportation Research Part B: Methodological 67, 35–67.

Hassannayebi, E., Memarpour, M., Mardani, S., Shakibayifar, M., Bakhshayeshi, I., Espahbod, S., 2020. A hybrid simulation model of passenger emergency evacuation under disruption scenarios: A case study of a large transfer railway station. Journal of Simulation 14, 204–228.

Lawphongpanich, S., Yin, Y., 2010. Solving the pareto-improving toll problem via manifold suboptimization. Transportation Research Part C: Emerging Technologies 18, 234–246.

Lin, T., Shalaby, A., Miller, E., 2016. 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.

Ma, C., Hao, W., He, R., Jia, X., Pan, F., Fan, J., Xiong, R., 2018. Distribution path robust optimization of electric vehicle with multiple distribution centers. PloS One 13.

Mo, B., Franque, M.Y.v., Koutsopoulos, H.N., Zhao, J., 2021. Impact of unplanned rail disruption on urban transit systems, in: Transportation Rsearch Board 100th Annual Meeting.

Mo, B., Ma, Z., Koutsopoulos, H.N., Zhao, J., 2020. Capacity-constrained network performance model for urban rail systems. Transportation Research Record , 0361198120914309.

Nguyen, S., Pallottino, S., Malucelli, F., 2001. A modeling framework for passenger assignment on a transport network with timetables. Transportation Science 35, 238–249.

Nielsen, O.A., 2000. A stochastic transit assignment model considering differences in passengers utility functions. Transportation Research Part B: Methodological 34, 377–402.

Rahimi, E., Shamshiripour, A., Shabanpour, R., Mohammadian, A., Auld, J., 2019. Analysis of transit users' waiting tolerance in response to unplanned service disruptions. Transportation Research Part D: Transport and Environment 77, 639–653.

Rahimi, E., Shamshiripour, A., Shabanpour, R., Mohammadian, A., Auld, J., 2020. Analysis of transit users' response behavior in case of unplanned service disruptions. Transportation Research Record , 0361198120911921.

Roelofsen, D., Cats, O., van Oort, N., Hoogendoorn, S., 2018. 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.

Schmöcker, J.D., Bell, M.G., Kurauchi, F., 2008. A quasi-dynamic capacity constrained frequency-based transit assignment model. Transportation Research Part B: Methodological 42, 925–945.

Schmöcker, J.D., Fonzone, A., Shimamoto, H., Kurauchi, F., Bell, M.G., 2011. Frequency-based transit assignment considering seat capacities. Transportation Research Part B: Methodological 45, 392–408.

Tan, Z., Xu, M., Meng, Q., Li, Z.C., 2020. 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.

Wang, J., Yuan, Z., Yin, Y., 2019a. Optimization of bus bridging service under unexpected metro disruptions with dynamic passenger flows. Journal of Advanced Transportation 2019.

Wang, X., Chen, S., Zhou, Y., Peng, H., Cui, Y., 2013. Simulation on passenger evacuation under fire emergency in metro station, in: 2013 IEEE International Conference on Intelligent Rail Transportation Proceedings, IEEE. pp. 259–262.

Wang, Y., Zhang, Y., Tang, J., 2019b. A distributionally robust optimization approach for surgery block allocation. European Journal of Operational Research 273, 740–753.

Wu, J.H., Florian, M., Marcotte, P., 1994. Transit equilibrium assignment: a model and solution algorithms. Transportation Science 28, 193–203.

Xiong, P., Jirutitijaroen, P., Singh, C., 2016. A distributionally robust optimization model for unit commitment considering uncertain wind power generation. IEEE Transactions on Power Systems 32, 39–49.

Zhou, M., Dong, H., Zhao, Y., Ioannou, P.A., Wang, F.Y., 2019. Optimization of crowd evacuation with leaders in urban rail transit stations. IEEE transactions on intelligent transportation systems 20, 4476–4487.