IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 May 2024, accepted 19 June 2024, date of publication 1 July 2024, date of current version 27 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3421312

RESEARCH ARTICLE

A Generalized Network Level Optimized Disruption Strategy Selection Model for Urban Zone Transport Systems

QI LIU[1,2], WEI WANG[1], YUMIAO LIU[1], AND YANZHAO SU[3]

[1]China Academy of Building Research, Beijing 100013, China
[2]Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China
[3]School of Vehicle and Mobility, Tsinghua University, Beijing 100190, China

Corresponding authors: Wei Wang (weiwang89@hotmail.com) and Yanzhao Su (yanzhaosu66@163.com)

This work was supported in part by the Fundamental Research Funds of China Academy of Building Research under Project 20221802330730004, and in part by the NSFC Program under Grant 52102439.

ABSTRACT A fast recovery from disruptions is of vital importance for the reliability and sustainability of urban zone transit systems. It's a complex task to coordinate multiple transit departments to mitigate disruption. There are many ways to response but it's not always obvious how to combine them in an optimized manner. This study presents a new attempt to tackle this problem in a comprehensive and hierarchical way. At phase (i), a network-scale strategy selection optimization model is formulated as a joint routing and resource allocation (nJRRA) problem. This model produces solutions for efficient allocation of network resources to facilitate inter-department coordination. By constraining the problem further into an ϵ-constrained nJRRA problem, classic solution algorithms can be applied to solve the quadratically constrained quadratic program (QCQP). On top of this ''basic model'', we propose adding a decision to delay the resource allocation decisions up to a maximum initiation time when the incident duration is stochastic. To test the models, a quasi-dynamic evaluation program with a given incident duration distribution is constructed using discretized time steps and discrete distributions. Five different demand patterns and four different disruption duration distributions (20 combinations) are tested on a small transit network. The results show that the two models outperform benchmark strategies such as using only line level adjustment or only bus bridging. They also highlight conditions when delaying the decision is preferred.

INDEX TERMS Disruption mitigation, public transportation, urban zone transport, user path recommendation.

I. INTRODUCTION

A. BACKGROUND

In the face of global climate change, reducing carbon emissions from transportation systems has a significant impact on the overall achievement of energy conservation and carbon reduction targets. To reduce transportation carbon emissions in urban zone such as industrial parks, developing intelligent traffic management technology can effectively promote the development of urban transportation systems towards a more environmentally friendly and sustainable direction. Daily transit operations encounter various types of disruptions, like track failure, rolling stock failure, intrusions, medical emergencies, weather/nature disasters, etc. (see **Figure 1**). Serious service degeneration may propagate through the network and last for hours. If disruptions occur in the urban zone such as industrial park, it would significantly impact the commuting of employees and the production supply. Given the importance of transit service reliability, the application of recovery models and algorithms for real-time disturbance and disruption management is considered a key element for improving the service and reliability of transit systems [1]. This is true for urban zone transport in general. It's a complex task to coordinate multiple transit departments to mitigate

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

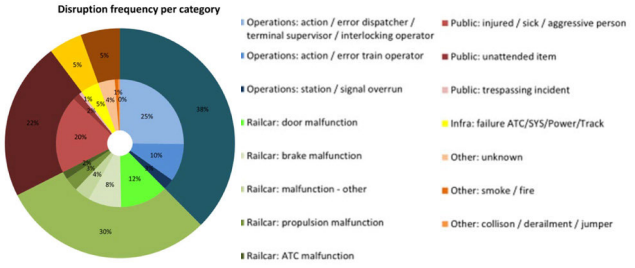


FIGURE 1. Disruption types from Washington metro network (WMATA) (Yap and Cats [7]).

disruption. There are many strategies in use today for a typical transit system; however, it is not always obvious how to find the optimal combination of strategies in real-time.

The exact set of feasible disruption mitigation strategies may differ from system to system or even from line to line because of the availability of crossings, parallel tracks, backup vehicles and staff, user acceptance, etc. Ceder [2] gives a comprehensive list of real-time control strategies, including holding the vehicle, stop-skipping, adding a reserve vehicle, changing speed, interlining, deadheading, short-turning, etc. Other strategies include bus bridging for metro [3], emergency lines [4], service network redesign [5], and cancellation/addition of tasks [6].

B. RELATED STUDIES

There are many studies on disruption mitigation strategies. Reviews of disruption management for passenger railway transportation can be found in Jespersen-Groth et al. [8] and Cacchiani et al. [1].

1) MINOR SERVICE ADJUSTMENTS PROBLEM

Headway adjustment is one of the most commonly used ways to adjust the service in a minor way (Hickman [9], Berrebi et al. [10]). Joint optimization models involving multiple strategies like holding, stop-skipping, expressing, short-turning, and deadheading, are usually formulated as mixed integer programming problems; studies include Su et al. [11], Hassannayebi et al. [12], Zhu et al. [13], etc.

2) SERVICE RUN ADJUSTMENT PROBLEM

A ''service run'' means a task on the timetable, like ''R train: Bay Ridge-95 St to Forest Hills-71 Av, starting at 7:03AM'' (MTA, NYC). Run addition or removal changes the headways resulting in larger consequences than holding strategies. If a run gets canceled, the current vehicle or crew plan may become infeasible. The model is formulated as an integer linear programs (ILP) by Thengvall et al. [6], Jarrah et al. [14], Zhan et al. [15]. Veelenturf et al. [16] proposed a model for the joint rescheduling of timetable and rolling stock for a railway system, solved by heuristic algorithm. Yuan et al. [17] proposed a model to jointly optimize the assignment of users and transit schedules. Yuan et al. [17] proposed integrated optimization approach for passenger flow control and metro scheduling considering skip-stop patterns.

3) SERVICE LINE REDESIGN PROBLEM

There are only a few studies on real-time service line redesign. Kiefer et al. [5] proposed a mixed integer programming model to respond to serious disruptions by redesigning the lines in a particular region around the disruption and adjusting the frequencies. In Cadarso et al. [4], lines can be canceled and emergency lines added. The rolling stock is jointly optimized.

4) SUBSTITUTION SERVICE DESIGN PROBLEM

Substituting a service by another mode may occur when a disruption disables the service locally. The bus is the most popular choice for substituting other modes (i.e. bus bridging). The bus bridging problem is similar to the transit route network design problem [3], [18], [19]. Gu et al. [18] developed a two-stage integer linear programming model to flexibly allocate and schedule buses to a set of shuttle bus routes during unexpected metro disruptions. Zhang and Lo [20] investigated the optimal initiation time for substitute bus service. Chen and An [21] studied integrated optimization of bus bridging routes and train timetables under rail disruption.

5) VEHICLE/CREW RESCHEDULING PROBLEM

Recovering from serious disruptions may require changes to the timetable, the rolling stock, as well as the crew duties. The vehicle and crew rescheduling problems are very similar. They are both about finding paths to cover a set of tasks. They are usually formulated as multi-commodity minimum cost flow problems [22], [23], [24]. Alternatively, they can be formulated as set partition/covering problems if trajectories are enumerated [25], [26], [27]. The set of possible routes of a realistic network is too large to enumerate. Hence, column generation is often used to solve the vehicle/crew rescheduling problem. Visentini et al. [28] reviewed the vehicle rescheduling problem for road traffic, railway, and airlines.

C. RESEARCH GAPS AND CONTRIBUTIONS

Firstly, disruption durations are typically unknown in advance. User demands are usually stochastic and only partially observable. Disruption mitigation considering all these stochastic factors has not been fully investigated. Secondly, most previous studies on urban transport are line level models to limit the size of the problem. However, passengers re-route on the whole network, and resources (like crews and vehicles) are distributed across the network. Models for intercity trains or airlines are indeed network level, but the disruption mitigation strategies for these systems are not as rich as urban public transport. For example, urban public transport needs more precise controls when it comes to dwell times and headways; urban public transport has bus-bridging options, etc. There

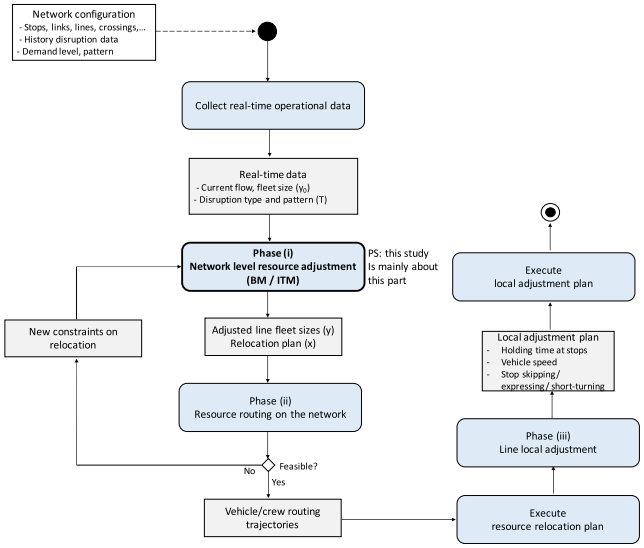


FIGURE 2. Activity diagram of hierarchical disruption mitigation.

is a need for an efficient disruption strategy selection model for urban transport that incorporates most of the commonly seen strategies at a network level and accounts for all the uncertainties. The main contributions of this paper are:

- We propose a new network-level disruption mitigation framework (**Figure 2.**); phase (i) strategy making is cast as a resource allocation problem. Strategies like bus bridging, inter-lining, short-turning, service line redesigning, service run adjustment, are all considered; this is done by mapping the strategies into equivalent service line fleet allocation decisions. This approach is comprehensive while still maintains tractability.
- Two phase (i) strategy making models (BM and ITM) are proposed, and tested over 20 different combinations of demand and disruption duration patterns. Comparisons with two other benchmark strategies are made. Insights are gained from analytics.

The paper is organized as follows: Section II presents the hierarchical framework and two strategy making models. Section III discusses the numerical tests on a small network example. Section IV concludes.

II. METHODOLOGY

A. FRAMEWORK

In this study, we focus on metro systems. In the case of disruptions, efficient use of available resources is desired. Previous studies formulate this strategy selection problem for urban public transport the same way we handle intercity trains or airline systems: a service run is a basic unit of task; train trajectories on the network are sought to cover these tasks. However, urban public transport system users typically do *not* buy tickets for a specific run or even a specific line. Instead, users pay for recurrent network services. The service level of a transit line is usually characterized by its average headway. And it's determined by the available resources allocated and

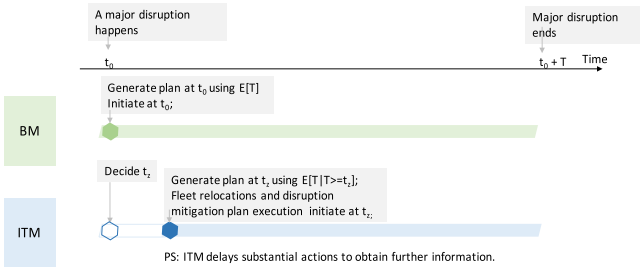


FIGURE 3. Comparison of BM and ITM.

physical constraints imposed. We argue that the basic task unit to be adjusted for urban public transport resource allocation is best at a *line service level* instead of service run. Operators respond to disruptions by changing line service levels through diverting vehicles and crews between lines, including some newly setup *emergency lines*. Vehicles and crews may come from a high-cost backup depot. The service line level approach allows us to evaluate network-wide resources in a tractable manner while still accounting for user delays over a finite time horizon.

The disruption mitigation is naturally separated into three phases: (i) Network level resource adjustment (i.e. the strategy selection); (ii) resource routing on the network; and (iii) local line adjustment. See Figure 2 for the activity diagram of hierarchical disruption mitigation. Modules corresponding to three phases run sequentially. The proposed models are implemented in Phase (i), where the strategy selection problem becomes a network resource allocation problem. Phases (i), (ii) and (iii) are run at network-level, regional-level, and line-level, respectively. That means multiple regions (or multiple lines) will run phase (ii) (or phase (iii)) concurrently.

Two models are proposed in increasing levels of complexity that both include the line service level basic structure for network resource allocation. In Section II-B, a novel "basic model" (BM) is first proposed to study the simplest deterministic disruption duration case. In Section II-C, the BM is extended to a random model called "Initiation Time Model" (ITM). The two models differ in the way they handle random disruption duration (see Figure 3 for an illustration of the differences). BM uses expected values and treats duration as deterministic. ITM delays substantial actions to obtain further information over a single time horizon.

B. BASIC MODEL (BM)

1) NOTATIONS

$c_{ll'}$:	average one-way cost of diverting vehicle from line l to line l' (constant);
$\mathrm{IE}[\boldsymbol{T}\|\boldsymbol{T} \geq z]$:	expected duration conditioning on event $\{\boldsymbol{T} \geq z\}$;
f_l :	the frequency of line l;
F^{BM} :	the objective of basic model (BM);
F^{ITM} :	the objective of initiation time model (ITM);
$g(T)$:	probability density function (pdf) of $\boldsymbol{T}$;
G :	transit network graph;

H_w : the set of paths between OD pair w;
I^{ITM} : ITM interval;
L : the set of transit lines;
M : set of transport modes;
$p_{w,h}$: the proportion of users of OD pair w on path h suggested by transit operator during disruption;
$p^N_{w,h}$: the path choices when system is undisrupted (''normal'');
$p^D_{w,h}$: the path choices when system is disrupted and with no relocation (''disrupted'');
$q_w(\tau)$: the user demand density for OD pair w at time τ;
$Q_w(t_1, t_2)$: the number of users belonging to OD pair w during time interval $[t1, t2]$;
Q_w : the number of users belonging to OD pair w during $[0, T]$; $Q_w := Q_w(0, T)$;
R_{l_s} : round-trip time of line l that is incident to segment S;
S : the set of transit line segments;
S_h : the set of segments on path h;
S^B_h : the set of boarding segments on path h;
t^R_s : running (traversing) time of transit segment S (constant);
t^P_h : path h cost during disruption after relocation finished;
$t^{P,N}_h$: path cost when system is undisrupted (''normal''), a constant;
$t^{P,D}_h$: path cost when system is disrupted and with no relocation (''disrupted''), a constant;
t^P_h : path cost when system is disrupted and with relocation;
T : disruption duration (a fix number) used in BM;
$\boldsymbol{T}$: disruption duration (a random variable) used in ITM;
$\overline{T}$: the upper bound of $\boldsymbol{T}$;
V : the set of transit stops;
W : the set of all OD pairs;
$x_{ll'}$: the number of vehicles relocated from line l to l' where l and l' are lines of the same mode $m \in M$;
y_l : adjusted fleet size for transit line l;
K : the capacity of the vehicle;
y^0_l : original line fleet size for line l;
Y_l : the upper bound of fleet size for line l;
G_s : left hand side of Eq. (2);
H_l : left hand side of Eq. (3);
I : left hand side of Eq. (4);
J_w : left hand side of Eq. (5);
K_l : left hand side of Eq. (6);
α : weighing coefficient for operator cost;
β : user value of time (VOT) per minute;
γ : wait time penalty coefficient;
μ_s : Lagrange multiplier for Eq. (2);
ϑ_l : Lagrange multiplier for Eq. (3);
η : Lagrange multiplier for Eq. (4);
π_w : Lagrange multiplier for Eq. (5);
θ_l : Lagrange multiplier for Eq. (6);
$\delta_{h,s}$: path h and segment S incidence;
$\delta_{h,l}$: path-line incidence;

Remarks:

1) Notation that appear only once are explained in text in place and not listed here;
2) A subscript is used for indexing, like 'w' for OD pair, 'h' for path, 'S' for segment, 'l' for line; superscript is used for differentiating, like 'B' for 'boarding', 'R' for running (traversing), 'D' for diverting, 'P' for path, '0' for naught.

The basic model (BM) is a phases (i) model (Figure 2). BM make high-level network resource allocation (equivalently, line service level adjustment) decisions. This network-scale model is quite general while still maintains tractability. Multiple transit department (metro, bus) may need to cooperate to fulfill the plan. The transit network is represented by a graph $G = (V, S)$ where V is the set of transit stops and S is the set of transit line segments. Initially, let disruption duration T be a fixed real number. We assume that users follow paths. A *path* is composed of a sequence of transit line segments. This is a simplified version of user assignment under a disruption setting. There is no consensus about how users react to disruptions. In some studies, it is assumed that users are rational, self-interested, and always choose the shortest path [4]. However, as argued by Xu and Ng [29], under unforeseen disruptions, commuters may need to react with limited information. Instead, users can be guided toward alternative contingency routes by operators. As such, we let the user path choices be decision variables of the model (i.e. decisions of the operator). It is possible to assume that a proportion of the users comply with operator orders and the rest of them act on their own with full information. We leave this option for future work.

BM is shown in Eq. (1) – (9). The decision variables are the fleet size for each line (y), assignment of users to paths (p), and the fleet relocation decisions among lines (x) needed to achieve y. Networks of different modes, like metro, bus, are jointly considered, where lines operate vehicles that belong to different non-interchangeable classes. In other words, $L = \cup_{m \in M} L_m$, where vehicles belonging to lines in class L_m cannot be exchanged with vehicles in a line belonging to a difference class $L_{m'}$, $m \neq m'$. The variable $x_{ll'}$ is applied only to $l, l' \in L_m$.

2) FORMULATION

$$\min_{p,y,x} F^{BM}(p, y, x)$$

$$= \sum_{w \in W, h \in H_w} Q_w p_{w,h} \left(\sum_{s \in S^B_h} \frac{\gamma R_{l_s}}{2 y_{l_s}} + \sum_{s \in S_h} t^R_s \right)$$

$$+ 2\alpha \sum_{m \in M} \sum_{l, l'' \in L_m} c_{ll'} x_{ll'} \quad (1)$$

Subject to: (Segment capacity constraint)

$$\mathrm{G}_s := \sum_{w \in W, h \in H_w} Q_w p_{w,h} \delta_{h,s} - \frac{KT}{R_{l_s}} y_{l_s} \leq 0, \quad (\mu_s)\,, \ \forall s \in S \quad (2)$$

(Fleet size adjustments)

$$\mathrm{H}_l := \sum_{l'} x_{ll'} - \sum_{l'} x_{l'l} + y_l = y_l^0, (\vartheta_l)\,, \quad \forall l \in L \quad (3)$$

$$I := \sum\nolimits_l y_l = \sum\nolimits_l y_l^0, (\eta) \quad (4)$$

(User path choices)

$$\mathrm{J}_w := \sum\nolimits_{h \in H_w} p_{w,h} = 1, (\pi_w)\,, \ \forall w \in W \quad (5)$$

(Fleet size bounds)

$$\mathrm{K}_l := y_l - Y_l \leq 0, (\theta_l)\,, \ \forall l \in L \quad (6)$$

(Non-negativity)

$$p_{w,h} \geq 0, \quad \forall w \in W, h \in H_w \quad (7)$$

$$y_l \geq 0, \quad \forall l \in L \quad (8)$$

$$x_{ll'} \geq 0, \quad \forall l, l' \in L_m, m \in M \quad (9)$$

The objective is to minimize the weighted sum of costs to transit users and the operator (Eq. (1)). User cost is the trip time multiplied by value of time (VOT) β as shown by the first term. Let $Q_w(t_1, t_2)$ be the number of users belonging to OD pair w during time interval $[t1\,, t2]$. $Q_w(t_1, t_2) = \int_{t_1}^{t_2} q_w(\tau)\, d_\tau$, where $q_w(\tau)$ is the user demand density for OD pair w at time τ. Let $Q_w := Q_w(0, T)$. For those passengers that enter the system before the horizon begins, their location in the system at time 0 is regarded as their origins. We add $Q_w^0 \delta_0(t)$ to the density $q_w(\tau)$ to take account of these demands, where Q_w^0 with $w = (O, D)$ means the number of users queueing at O heading to D at time 0 and $\delta_0(t)$ is the Dirac delta function with a peak at $t = 0$. The complex process of transit system state transition during resource relocation is not modeled in this phase to avoid dynamic transit assignment modeling. The problem in reality is much more complex, as discussed in the literature on dynamic transit assignment (e.g. see Hamdouch and Lawphongpanich [30]; Jin et al. [31]). Time-varying travel times and flows mean that paths may not be easily categorized into pre-initiation/ recovery/ recovered stages. There could be passengers crossing the boundaries. Keeping track of these passengers will require the use of dynamic transit assignment with time-expanded networks (TE-network). The problem with adopting such frameworks is that they are not very scalable, which prevents the use of a strategy selection model at a network level. Instead, we try to keep things simple by assuming that the time horizon of the incident is small enough between those three stages that paths can be pre-identified for OD pairs. After selecting strategies, more detailed dynamic models may be deployed to aid implementation of the strategies in Phases ii and iii as shown in Figure 1.

The average user waiting cost of a boarding segment is computed by $R_{l_s}/2y_{l_s}$ where l_s refers to the line of segment S and R_{l_s} is the roundtrip time of this line. The path h has average cost $\sum_{s \in S_h^B} \gamma R_{l_s}/2y_{l_s} + \sum_{s \in S_h} t_s^R$ where γ is the wait penalty coefficient and t_s^R is the segment travel cost. The second term, operator cost, is the spending on resource relocation. Operator cost is weighted by a parameter α as shown in the objective. We assume that all fleet sizes restore to normality after disruption. $c_{ll'}$ is a unit one-way relocation cost. We do not restrict fleet size change variables x and y to be integral. The rounded values are typically good enough at a strategy selection level in phase i and can provide informative results for deploying strategies in phases ii and iii. We also allow p to be fractional, which means the operator can control the exact proportion of users on a path. There are more sophisticated ways to estimate the passenger delay, like Sun et al. [32]. The use of fractions for passenger paths is even less of an issue than for frequencies, as passenger volumes tend to be high enough (e.g. rounding 287.8 to 288), just as all transit assignment models in the literature do not assume integer values.

Eq. (2) requires that the total demand to cross a *line segment* during T be no larger than the expected capacity provided during T, which again depends on the average headway. Eq. (3) and Eq. (4) are about the fleet conservation constraints. Eq. (5) are the path flow conservation constraints and Eq. (6) are the fleet size bounding constraints. Eq. (7) are the non-negativity constraints. $\mathrm{G}_l, \mathrm{H}_l, I, \mathrm{J}_w, \mathrm{K}_l$ are functions representing left-hand side (LHS) of constraints; μ_l, ϑ_l, η_l, π_l, θ_l are the corresponding Lagrange multipliers of the constraints. Paths are enumerated under this formulation. For convenience, k-shortest paths are used to approximate the true set of paths (see Bekhor et al. [33]). While the appropriate number of k paths to be chosen can vary with the size of the network [34], for simple networks Cascetta et al. [35] showed that 4-7 paths may suffice.

Parameter Y_l is the maximum fleet size of line l. Y_l is determined by the throughput capability of line l. If multiple lines share some segments, the maximum fleet sizes of these lines are related; constraints like $\sum_{l_i=l_1}^{l_n} Y_{l_i} \leq c_s$ should be imposed. We leave this out in the formulation for simplicity. The relocation cost is defined in Eq. (10). The diverted fleets cannot provide regularly scheduled service during the diversion. This cost is captured by term $\gamma_{ll'}^D t_{ij}^D$. The costs of using backup vehicles and crews are represented by the term $\overline{c}_{ll'}$ when l is a backup depot. c^0 represents the minimum costs associated with making diversions.

$$C_{ll'} = c^0 + \overline{C}_{ll'} + \gamma_{ll'}^D t_{ij}^D \quad (10)$$

where

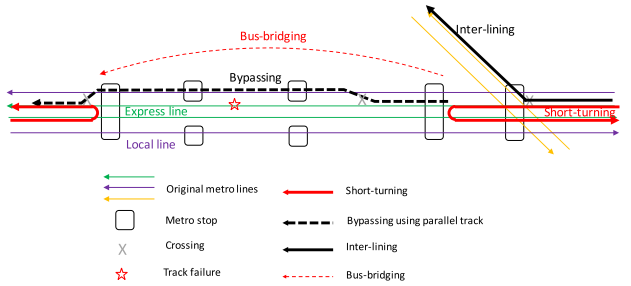


FIGURE 4. Strategies considered in this study.

$\gamma^D_{ll'}$: user cost for unit time spent on diverting unit vehicle from l to l' (which includes lost service on l and unavailability on line l' until its arrival);
t^D_{ij}: time that it takes to divert unit vehicle from line l to line l';
$\overline{c}_{ll'}$: vehicle and crew cost for diverting unit vehicle from l to l';
c^0 : penalty for making changes.

BM is *generalized* in the sense that several other commonly seen models can be regarded as special cases. We use "fixing a network" to refer to the network topology being fixed but service level being subject to change; and use "fixing a service" to refer to both the network topology and the line service levels being fixed.

- Special case 1): If we fix the bus service, and allow metro network redesign as well as metro resource relocation, this is the *service line redesign problem*;
- Special case 2): If we fix bus service and fix the metro network, but can relocate the resources possibly across metro lines, this is the *service run adjustment problem*;
- Special case 3): If we fix the metro service but can adjust the bus services by adding bus-bridging lines and adjust bus line frequencies, this is the *substitution service design problem (bus-bridging problem)*;
- Special case 4): If we fix metro and bus network but can adjust the service levels of both original metro and original bus lines; this is the *multi-modal joint optimization problem*.

Figure 4 gives an illustration of the strategies considered in this study. When disruption happens, we can adjust current metro and bus line services, as well as set up new emergency metro and bus lines. The needed fleet could come from a backup depot or from existing lines. Not all strategies are needed at the same time. The best combinations are sought. It's in fact challenging for multiple departments to coordinate in real life. Bus-bridging using backup buses is one of the most common ways to handle disruption; coordination more that is rare. However, we see this as opportunity rather than limitation. The hierarchical framework and line average-service-level adjustment approach (other than individual run adjustments) simplify the problem of system-level strategic decision making and facilitate coordination.

3) PARAMETER ESTIMATION

BM is parametrized by OD demands, cost coefficients, value of time, and the expected disruption duration. The strategy selection model in phase i assumes that the transit system knows at the start of the disruption which service lines are available, which are impacted, such that immediately available emergency lines can be determined instantaneously. Similarly, OD demand is assumed to be known. These assumptions are similar to the state of the art, as summarized in the literature review (see Section 1.4). For example, most modern transit systems have Advanced Vehicle Location (AVL) systems to keep track of their vehicle fleets at any time and can pinpoint the exact line segment or track section that is disrupted. Similarly, transit systems have historical data and Automated Passenger Counters (APC). Combined with state-of-the-art origin-destination inference methods (see Liu and Chow [36]), transit systems can infer expected passenger OD flows over a time horizon. For example, NYC Transit keeps track of passenger arrivals through turnstile data and wifi detection using the Transit Wireless system. They also have a transit control center that keeps track of the status of all rail segments in the subway system. These systems help provide a picture of passenger ODs and paths. Readers are recommended to follow studies on OD flow estimation (Castillo et al. [37]), network design problem for building set of emergency lines (Jin et al. [31]), and survival analysis for disruption duration (Tinguely et al. [38]) among others.

4) OPTIMALITY CONDITIONS

Eqs. (1) – (9) have a nonlinear objective with linear constraints. There are two weight coefficients α and β in the objective. Without loss of generality, we may assume $\beta = 1$ after the transformation $\alpha = \alpha/\beta$. The KKT conditions are shown in Eq. (11) along with primal constraints.

$$\begin{aligned}
&\nabla \mathrm{L} := \nabla F^{BM} + \sum_s \mu_s \nabla \mathrm{G}_s + \sum_l \vartheta_l \nabla \mathrm{H}_l \\
&\qquad + \eta \nabla I + \sum_w \pi_w \nabla \mathrm{J}_w + \sum_l \theta_l \nabla \mathrm{K}_l \geq 0 \\
&\nabla_{p_{w,h}} \mathrm{L} \cdot p_{w,h} = 0, \quad \forall w \in W, h \in H_w \\
&\nabla_{y_l} \mathrm{L} \cdot y_l = 0, \quad \forall l \in L \\
&\nabla_{x_{ll'}} \mathrm{L} \cdot x_{ll'} = 0, \quad \forall l, l' \in L \\
&\mathrm{G}_s \cdot \mu_s = 0, \quad \forall s \in S \\
&\mathrm{K}_l \cdot \theta_l = 0 \quad \forall l \in L \\
&\mu_s \geq 0, \ \forall s \in S \\
&\theta_l \geq 0, \quad \forall l \in L
\end{aligned} \tag{11}$$

From the KKT conditions, we have the following observations.

Observation 1: Condition for path h belonging to OD w to be in use is:

$$\begin{aligned}
&p_{w,h} > 0 \Rightarrow \nabla_{p_{w,h}} \mathrm{L} = 0 \\
&\nabla_{p_{w,h}} \mathrm{L} = Q_w t^P_h(y) + \sum_s \mu_s Q_w \delta_{h,s} + \pi_w = 0
\end{aligned}$$

where path cost $t_h^P(y) = \sum_{s\in S_h^B} \frac{\gamma R_{l_s}}{2y_{l_s}} + \sum_{s\in S_h} t_s^R$; or, equivalently, as in Eq. (12).

$$t_h^P(y) + \sum_s \mu_s \delta_{h,s} = -\frac{1}{Q_w}\pi_w \tag{12}$$

The first term on the LHS is the user cost of path h; the second term on the LHS are segment-capacity shadow prices. The RHS can be interpreted as the cost of the marginal shortest path for OD w: the cost of sending marginal flow along the shortest path under the optimally loaded flow. Note π_w is unrestricted. This condition says that, if a path h for OD w is used, then its cost plus the segment-capacity shadow prices equals the marginal shortest path length. This type of condition is common for a multi-commodity flow problem.

Observation 2: Condition for emergency line l to be in use is:

$$\begin{aligned} y_l > 0 \Rightarrow \nabla_{y_l} \mathrm{L} = 0 \\ \nabla_{y_l} \mathrm{L} = \sum_{w\in W, h\in H_w} \sum_{s\in S_h^B, l_s=l} \left(-\frac{\gamma R_l}{2y_l^2}\right) \\ - \mu_l \frac{KT}{R_l} + \vartheta_l + \eta + \theta_l = 0 \end{aligned}$$

Moving some negative terms to the RHS, we get:

$$\begin{aligned} \vartheta_l + \eta + \theta_l = \left(\sum_{w\in W, h\in H_w} Q_w p_{w,h} \sum_{s\in S_h^B, l_s=l} \frac{\gamma R_l}{2y_l^2}\right) \\ + \sum_{s\in S_h^B, l_s=l} \mu_s \frac{KT}{R_{l_s}} \end{aligned} \tag{13}$$

ϑ_l is the multiplier associated with relocation flow x conservation; it is the node potential in the transportation problem. It could be interpreted as the marginal cost of diverting vehicles to line l. η is the shadow price of fleet resource. θ_l is the price associated with upper bound Y_l which could be positive if line l is operated at capacity. The first term on the RHS is the (positive) waiting time savings of users with respect to unit y_l increase. The second term on the RHS is the marginal benefit of improving line l capacity which could be nonzero if some segment belonging to l operates at capacity. Hence the equation means marginal cost of diverting plus fleet shadow price and fleet upper bound shadow price are equal to the marginal savings of user wait time plus marginal benefits of expanding capacity. Conversely, if the following condition (Eq. (14)) is satisfied, then we must have $y_l = 0$; namely, this emergency line is not in use.

$$\begin{aligned} \vartheta_l + \eta > \left(\sum_{w\in W, h\in H_w} Q_w p_{w,h} \sum_{s\in S_h^B, l_s=l} \frac{\gamma R_l}{2y_l^2}\right) \\ + \sum_{s\in S_h^B, l_s=l} \mu_s \frac{KT}{R_l} \end{aligned} \tag{14}$$

Observation 3: Condition for fleets being diverted from line l to line l' is:

$$\begin{aligned} x_{ll'} > 0 \Rightarrow \nabla_{x_{ll'}} \mathrm{L} = 0 \\ \nabla_{x_{ll'}} \mathrm{L} = 2\alpha c_{ll'} + \vartheta_l - \vartheta_{l'} = 0 \end{aligned} \tag{15}$$

where $2\alpha c_{ll'}$ is the cost of $x_{ll'}$; ϑ_l is the node potential as we mentioned. This is exactly the optimality condition fora transportation problem.

5) SOLUTION METHOD

The BM formulation can be generalized to (P0), which has potential for broader applications. Variable p is the user demand assignment decision; y is the servicelevel decision for lines (or any other type of service entities); x is the resource diversion decision among lines. The objective is composed of the diversion cost and user cost. It has nonlinear terms like $p_{w,h}/y_l$ in the objective which represent delay from a deterministic queueing system. This objective is not convex. For other types of queueing systems, the exact formulation may be different, but what is in common is non-convexity. Take M/M/1 for example; average delay has the form of $v/(c-v)$ where v is link flow and c is capacity; when v and resource c are decision variables, this delay function is also not convex. Here the constraints are Eq. (2) to (9) as before, although other types of systems may call for changes. We call (P0) the nonconvex *Joint Routing and Resource Allocation* (nJRRA) problem. This problem shares similar properties with the multi-commodity capacitated network design problem, which differs in the use of binary variables to allocate link investment resources while subject to optimal passenger flows (see Gendron et al. [39]).

$$\text{(P0)}\ \min\ F(p,y,x) = c'x + c''p + \sum_{w,h,l} c''_{w,h,l} \frac{p_{w,h}}{y_l}$$
$$\text{subject to linear constraints (2)} - \text{(9)}.$$

Convex or nonconvex JRRA problems arise when studying many different types of networks, like transit networks, computer networks, or power grids. Operators (or ISPs for internet, utility companies for power grids) plan the resource relocation and can control how flows are distributed on the network at the same time. Xiao et al. [40] studied the JRRA problem (called "simultaneous routing and resource allocation (SRRA)" there) for a wireless network. They assumed the objective to be convex for minimization and concave for maximization, like the utility function. The problem is solved through Lagrange duality. Capacity multipliers are introduced, then the resulting Lagrange dual problem can be decomposed. A subgradient method is used to update capacity multipliers. El-Sherif and Mohamed [41] studied JRRA minimizing delay for cognitive radio based wireless mesh networks. The objective is similar to (P0). Their model is formulated as mixed integer programming. Similar studies include Rasekh et al. [42]. In this section, we discuss global solution algorithms for nJRRA. The domain is compact. Note $y_l = 0$ is within the domain. We define p_{wl}/y_l to be 0 if p_{wl}

and y_l are both zero. Namely, if a line has no vehicle, and if no user is diverted to this line, then the user cost accumulated on this line is zero. If $y_l = 0$ for l and $p_{w,h} > 0$ for some w, h and path h uses this line l, then objective F of (P0) becomes infinite. So, the dependence of F on p and y is discontinuous at $y_l = 0$. The logical relation in Eq.(16) holds at optimality but the reverse is wrong in general.

$$\left(y_l^* = 0 \wedge \delta_{h,l} = 1\right) \Rightarrow p_{w,h}^* = 0, \quad \forall w \tag{16}$$

The essential singularity point at the boundary caused by p/y terms may bring trouble to the convergence of iterative algorithms. Hence, we define a more constrained version of P0 that additionally requires y_l to be no less than a small positive number ε, say 0.01. If we find that the algorithm outputs $y_l^* = \varepsilon$, then we can safely regard y_l^* as zero for practical purpose. Let $u_l := 1/y_l$, then u_l is bounded above by $1/\varepsilon$. In this way, our problem has a compact domain and the objective is smooth on this domain. With new variable and new constraints added, we have problem (P1) reflecting an "ε-constrained nJRRA". As $\varepsilon \to 0$, P1 approaches P0.

$$\text{(P1)} \ \min \ cx + c'p + \sum_{w,h,l} c'''_{w,h,l} p_{w,h} u_l$$

subject to linear constraints in (P0), and:

$$u_l y_l = 1, \quad \forall l$$
$$\varepsilon \le y_l \le \bar{y}_l, \quad \forall l$$
$$\frac{1}{\bar{y}_l} \le u_l \le \frac{1}{\varepsilon}, \quad \forall l$$

The ε -constrained JRRA problem is a special case of a Quadratically Constrained Quadratic Program (QCQP), although not every QCQP is of type (P1) and there may exist more efficient algorithms dedicated to the JRRA problems. This is reserved for future research. QCQP with a nonconvex objective is generally NP-hard [43]. QCQP is a fundamental problem that has been studied extensively in the global optimization literature (Al-Khayyal et al. [44], Audet et al. [45]). Two limit cases are discussed below to draw insights on the BM strategy.

For the first case (Figure 5 (a)), assume that there are K lines connecting K different OD pairs and there is no user interaction of any kind. Also, assume that we can ignore the relocation cost, namely x's have coefficient zero. This simplified version of the problem can be written as:

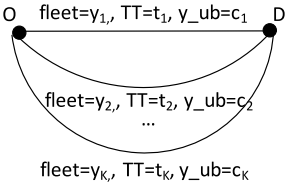
$$\min_y \sum_k \frac{q_k}{y_k}$$
$$\sum_k y_k = c$$
$$y_k \ge 0$$

where

q_k : demand of line k;
y_k : feet size of line k;
c : total number of vehicles.

(a) Resource relocation between independent lines; The optimal solution is *square-root rule*:

$$y_1 : y_2 : \ldots : y_K = \sqrt{q_1} : \sqrt{q_2} : \ldots : \sqrt{q_K}$$



(b) Resource relocation between parallel lines; The optimal solution is *shortest path first rule*.

FIGURE 5. Two special cases.

We can easily solve by using the first order conditions to find that at optimality, Eq. (17) holds:

$$y_1 : y_2 : \ldots : y_K = \sqrt{q_1} : \sqrt{q_2} : \ldots : \sqrt{q_K} \tag{17}$$

This corresponds to the *square root rule* [46]– the fleet sizes should be proportional to the square roots of number of passengers. As for the second case (Figure 5(b)), suppose there are K lines connecting one single OD pair. Each line k has travel time t_k and fleet size upper bound c_k. Also, suppose $t_1 < t_2 < \ldots < t_K$. The optimal solution is obvious: first assign fleets of size $y_1 = \min\{c_1, c\}$ to line 1; If there are vehicles left, then assign $\min\{c_2, c - y_1\}$ to line 2; continue until we run out of vehicles. We also give a name to this simple strategy – *shortest path first rule*.

C. INITIATION TIME MODEL (ITM)

When a disruption happens, it may be advantageous for an operator to wait for some time before taking any costly actions like bus bridging or inter-line vehicle diversion. Ideally, disruption mitigation should be modeled as a continuous decision-making process. For simplicity, the ITM model assumes that vehicle relocation initiates only once in the horizon. The exact time to start such are location is up to the operator. If the disruption recovers while waiting, then there is no need to make any relocations. Delaying actions reflects the principle that there is a tradeoff between the user cost and operator cost. This idea can be found in Zhang and Lo [20], although they only focus on a single disrupted metro line and a single strategy - bus bridging.

Let $\boldsymbol{T}$ now be the random disruption duration. Suppose $\boldsymbol{T}$ is bounded above by $\overline{T}$ and suppose that it is continuously distributed with probability density function (pdf) g. We add a new variable z, the relocation initiation time. The other variables are the same as before and the problem is labeled (P2). The objective (Eq. (18)) has three terms. The first term

corresponds to the user cost when $\boldsymbol{T} < z$, the case that the relocation has never been initiated. The second term corresponds to the user cost when $\boldsymbol{T} \geq z$. The second term can again be decomposed into three sub-terms, corresponding to pre-initiation, recovery, and recovered periods. The expected operator cost is captured by the third term. Eq. (19) means that the capacity in the interval $[z, \mathrm{E}[\boldsymbol{T}|\boldsymbol{T} \geq z]]$ can satisfy the demand in that period where $\mathrm{E}[\boldsymbol{T}|\boldsymbol{T} \geq z]$ means $\boldsymbol{T}$ is conditioned on the event that the disruption has not ended at time z. Eq. (20) is the upper bound on z and ρ is its corresponding Lagrange multiplier.

1) FORMULATION

$$
\begin{aligned}
\text{(P2)} \quad & \min_{z,p,y,x} F^{ITM}(z,p,y,x) \\
&= \underbrace{\sum_{w\in W,h\in H_w} \int_0^z \left(Q_w(0,T)\, p^D_{w,h} t^{P,D}_h + Q_w(T,\overline{T})\, p^N_{w,h} t^{P,N}_h\right) g(T)\, d_T}_{case \boldsymbol{T} < z: user\ cost\ in [0,\overline{T}]} \\
&\quad + \underbrace{\sum_{w\in W,h\in H_w} \int_z^{\overline{T}} \left(\underbrace{Q_w(0,z)\, p^D_{w,h} t^{P,D}_h}_{user\ cost\ in\ [0,z]} + \underbrace{Q_w(z,T)\, p_{w,h} t^P_h(y)}_{user\ cost\ in [z,T]} \right) g(T)\, d_T}_{case\ \boldsymbol{T} \geq z: user\ cost\ in [0,\overline{T}]} \\
&\quad + \underbrace{\sum_{w\in W,h\in H_w} \int_z^{\overline{T}} \left(\underbrace{Q_w(T,\overline{T})\, p^N_{w,h} t^{P,N}_h}_{user\ cost\ in [T,\overline{T}]} \right) g(T)\, d_T}_{case\ \boldsymbol{T} \geq z: user\ \ cost\ in [0,\overline{T}]} \\
&\quad + \underbrace{2\alpha \sum_{l,l'} C_{ll'} \chi_{ll'} \int_z^{\bar{T}} g(T) d_T}_{\mathrm{E}\ [fleet\ \ size\ adjust\ \ cost]} \qquad (18)
\end{aligned}
$$

(Segment capacity constraint)

$$
\begin{aligned}
\mathrm{G}_s := & \sum_{w\in W,h\in H_w} Q_w(z, \mathrm{E}[\boldsymbol{T}|\boldsymbol{T} \geq z])\, p_{w,h} \delta_{h,s} \\
& - \frac{K}{R_{l_s}} (\mathrm{E}[\boldsymbol{T}|\boldsymbol{T} \geq z] - z)\, y_{l_s} \\
& \leq 0, (u_s) \quad \forall s \in S \qquad (19) \\
\mathrm{R} := & z - \bar{T} \leq 0, \quad (\rho)
\end{aligned}
$$

Subject to Eq.(3) to (9). (20)

2) SOLUTION METHOD

The capacity constraints are nonlinear now. The rest of the constraints (Eqns. (3) - (9)) are linear as before. The objective is nonlinear and more complex than that of BM. We note that if we fix z, the solution algorithm previously discussed still applies with some minor changes. It is usually more convenient to discretize the time for application. Here we describe a practical way to speed up discrete ITM; we call it *early-break-ITM*. First, we discretize $\overline{T}$ and restrict the candidate initiation time to be a multiple of an ITM interval, I^{ITM}.The idea is to start with $z = 0$ and increase z until no successive improvement can be made; then break out of the iteration and return the last z. We call the sub-problem of ITM with fixed variable z by the name (P3).

Problem (P3) has the same complexity as BM. From our experiences, delay time z is mostly within 30 minutes. If I_{ITM} is set to be 10 minutes, four iterations would suffice in most cases. Hence, the complexity of Algorithm 2 would typically be about four times that of the BM solution algorithm.

Algorithm: Early-Break-ITM

Start with $z = 0$, $z^{opt} = \varnothing$, $F^{opt} = \infty$;
While $F^*(z) < F^{opt}$:
  $z^{opt} = z$;
  $F^{opt} = F^*(z)$;
  $z = z + I^{ITM}$;
  Solve (P3) to get $F^*(z)$;
Output z^{opt} and F^{opt}.

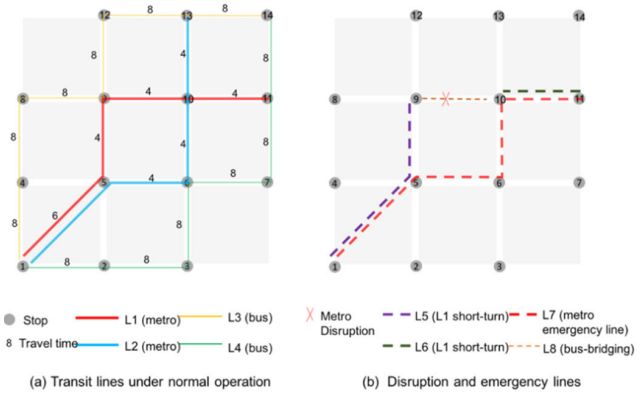


FIGURE 6. Example network and disruption.

III. NUMERICAL TESTS

We test the proposed models on a small transit network for reproducibility (Figure 6). The network has two metro lines and two bus lines (Figure 6 (a)).

A. DETERMINISTIC DISRUPTION DURATION CASE

Eight representative OD pairs are considered: 1-10, 5-14, 3-13, 8-11, 11-2, 13-4, 14-1, and 10-5. The time-dependent demands are assumed to be deterministic and concave, represented by parabolic functions. We will use two parameters q_{min} and q_{max} to specify the curve:

$$q(t) = -\frac{4}{T^2}(q_{max} - q_{min})\, t^2 + \frac{4}{T}(q_{max} - q_{min})\, t + q_{min}$$

The disruption to be considered is the failure of bi-directional link N9-N10 on line L1, say, due to tunnel power failure. The disruption lasts for $T = 60$ minutes for certainty. Four emergency lines—L5, L6, L7, and L8—are generated manually for this example as shown in Figure 6 (b). For real networks, the generation of these candidate lines should be automated and reserved for future research. L5 and L6 are short-turned metro lines for L1. L7 is a detour of the broken link in L1 using the L2 track. L8 is a bus-bridging line connecting metro stops 9 and 10.

For this deterministic case, three models are considered: line-level adjustment (LLA), bus-bridging (BB) and basic model (BM). LLA includes line level strategies, like

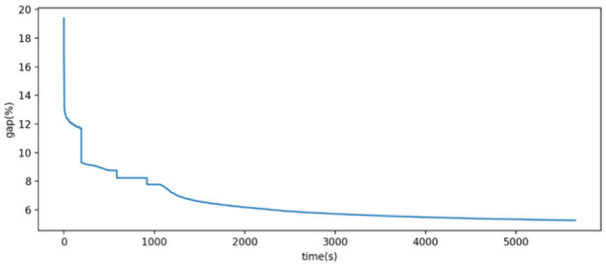


FIGURE 7. Gap changes over iterations (BM model solved by Gurobi).

short-turn and diverting users, but there is no inter-line fleet exchange. BB allows any strategies in LLA, and also allows the operator to allocate buses from a backup depot or existing lines to bridge the disputed links. BM allows any strategy in LLA and BB, and allows fleet exchange among different lines. We can see that BB is an extended model of LLA and BM is extended model of BB.

The convergence rate of BM is illustrated in Figure 7. The gap is defined as $(UB\text{-}LB)/UB$, measuring how far the current best solution's objective is from the current relaxation, not the distance to the optimum. As such, even an optimal solution may have a nonzero gap. Within 2 minutes, the gap drops from an initial value of 20% to 10%. In 10 minutes, it drops to 8%. Afterwards, the trajectory becomes quite slow as there are many leaves on the branching tree. This convergence behavior is typical in such algorithms for QCQPs.

The testing results are shown in Table 1. BM runs for 5 minutes, timing out before converging. LLA and BB run much faster (solved within 1 minute) and their results represent global optima (since each problem has a different objective, we do not expect the values to be equal). BM, even with a suboptimal solution, has almost the same level of service for users compared with an optimal BB, but at a much smaller operator cost. The improvement may not seem impressive at first sight, just about 4 percent compared to BB; but note that the "total costs" include the costs of all users on the whole network, disrupted or not. In summary:

- Introducing more strategies can significantly reduce operational cost without compromising user service levels.

This result supports the use of comprehensive strategy selection models instead of models that focus on single strategy optimization.

B. STOCHASTIC DISRUPTION DURATION CASE

Next, we test BM and ITM with stochastic disruption duration $\boldsymbol{T}$. The maximum duration of disruption, $\overline{T}$, is set to be 4 hours. Five different demand patterns over time are used: uniform, increasing, decreasing, concave and convex (Figure 8 (a)). They are represented by zero, first and second order polynomials. The concave and convex functions are similar to what we used in the deterministic case. All these patterns have two range parameters q_{max} and q_{min}. When the demand pattern is uniform, the density is $1/2\,(q_{min} + q_{max})$.

TABLE 1. Performance comparisons under deterministic disruption duration case.

Model	User cost ($)	Operator cost ($)	Total cost ($)
Line-level adjust (LLA)	16607.5 (100%)	0	16607.5 (**100%**)
Bus-bridging (BB)	15237.5 (91.2%)	1200	16437.5 (**98.9%**)
Basic model (BM)	15260.9 (91.9%)	353	15614.2 (**94.0%**)

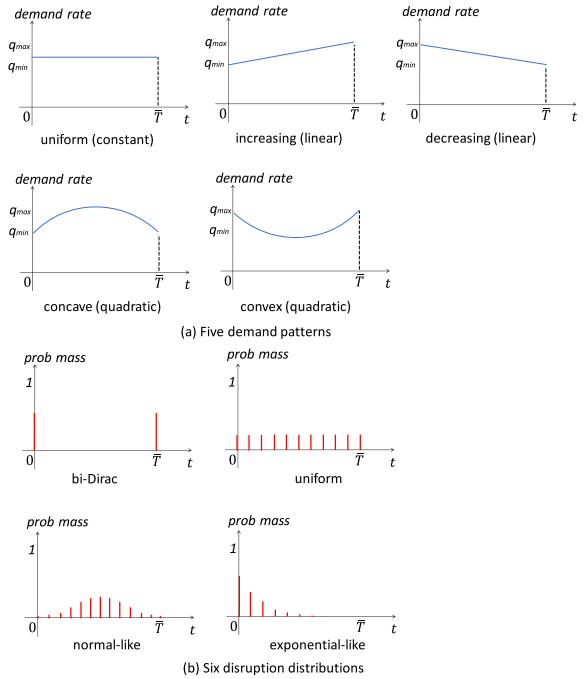


FIGURE 8. Demand and disruption distributions.

Probability mass functions (pmfs) are used to reflect the probability of the disruption duration $\boldsymbol{T}$. Six disruption duration distributions are illustrated in Figure 8 (b): Dirac at time zero (δ_0, "Dirac-0"), Dirac at $\overline{T}$ ($\delta_{\overline{T}}$, "Dirac-Tub"), weighted sum of Diracs at time 0 and $\overline{T}(1/2\,(\delta_0 + \delta_{\overline{T}})$, "bi-Dirac"), uniform, "normal-like", and "exponential-like". Note that Dirac-0 means that the disruption lasts for one time interval (10 min). The first three distributions are simple and interesting for theoretical purposes. The first two are even deterministic. They can be treated as limiting cases of more complicated patterns to be considered. We list these three simple distributions since they can help us to understand the problem.

1) TEST RESULTS

We test five demand patterns and four disruption duration distributions, 20 combinations in total. The detailed results are shared at https://github.com/BUILTNYU/transit-disruption-mitigation (see Table A1-A6). "Operator cost" is part of the output of the mitigation plan generation model. "# BU bus" in the table means the number of back-up buses used;

"z" means the initiation time found (min). Note the models are not solved with the same objective: LLA, BB and BM uses Eq. (1); ITM uses Eq. (18). Here we let algorithms to time out after 5 minutes. It's better to survey transit operators on how long they can tolerate to make a decision. This is left for future work. The algorithms do not necessarily converge to the global optimum under the 5-minute running time constraint (runs that time out at 5 minutes are noted in appendix tables). However, they are all evaluated using the same metric, "Expected User cost". Both the objective values and the performance metrics are reported in the tables A1-A5. The operator cost weighting factor alpha is set to be 2. ITM interval is set to be 10 minutes. Demand level is set at $q_{min} = 10$ and $q_{max} = 20$. All possibilities of disruption duration are tested and the resulting expected costs are determined under uncertain disruption duration. Based on the test results, some observations are made below. We label cases by "demand pattern, duration distribution", like "uniform, normal-like".

Remark 1: The overall performance of ITM is the best compared to LLA, BB, and BM among the instances tested.

ITM outperforms BM significantly when the duration distribution is bi-Dirac; the overall performance of BM is otherwise very close to ITM. When initiation time z is zero, the ITM model performance is slightly behind BM, like the case "uniform, uniform" in Table A1. This is because ITM is more difficult to solve, and its gap is larger for the 5-minute running time. For cases in which z is positive, ITM is significantly better than BM, like the case "uniform, bi-Dirac" in Table A1.

Remark 2: The overall performances of LLA and BB are significantly worse than that of BM and ITM.

Remark 3: When the demand pattern is concave or the duration distribution is bi-Dirac/exponential-like, it is advantageous to postpone the resource relocation decision.

For example, for the case "increasing, exponential" in Table A2, ITM generates $z = 20$.

Remark 4: When the demand pattern tends to be uniform or even decreasing, which means most of the users will arrive in the near future, or when disruption is likely to last for a long time, it makes less sense to delay actions.

As we can see from Table A1 and Table A3, as long as the duration distribution is not bi-Dirac, ITM delay z's are zero.

Remark 5: When the disruption distribution is exponential-like, a backup bus is not used in BB.

For cases with "exponential-like" disruption distribution in Table A1 to Table A5, the number of backup buses used by the BB model is 0. Backup buses are heavily used when the demand pattern is concave or decreasing. For cases where the demand pattern is "concave" and the disruption distribution is not "exponential-like", the number of backup buses used is 2.

Remark 6: ITM will not delay for more than an upper limit due to the penalties of user delay.

In the instances tested, ITM never delays more than 30 minutes.

TABLE 2. Model outputs under varying alpha settings.

Id	Case: (demand, disruption)	Decision variables	Alpha		
			10	20	30
1	uniform, exponential-like	# BKP bus	0	0	0
		z	0	0	30
2	increasing, normal-like	# BKP bus	2	1	0
		z	0	0	0
3	decreasing, bi-Dirac	# BKP bus	2	2	0
		z	0	10	10
4	convex, bi-Dirac*	# BKP bus	2	1	0
		z	30	10	10
5	increasing, uniform*	# BKP bus	2	1	0
		z	10	0	0
6	concave, normal-like*	# BKP bus	2	2	1
		z	10	30	0

* Seemingly counter-intuitive cases: ITM–z decreases as alpha increases.

The effect of weighting parameter α is also investigated. The overall effect is that as α gets larger, the operating cost becomes more important, the number of relocations decreases, and the adjustments are initiated later. These effects match our expectations. This is illustrated by the case 1 to 3 in Table 2. We just listed two indices - the number of backup buses used by BB and initiation time of ITM; they are enough to provide insights on the effects of alpha. We also notice that there are many cases in which the relationship between the number of relocations and alpha is not so obvious. Cases 4 to 6 in Table 2 illustrate this situation. Take case 4 for example. When alpha increases from 10 to 20, ITM initiation time decreases from 30 to 10. This is counter-intuitive: why would operator initiate relocations earlier when they care more about their own costs? Looking closer at the results we find that the number of relocations is less - the number of backup buses used is less than before.

Remark 7: Optimal decision variables don't necessarily depend monotonically on alpha; good decisions are hard to guess and best found by optimization.

Although a large number of scenarios have been investigated; there is still a need to extract the factors of variation underplaying these scenarios and to explain their relationships with model performances. Figure 9 represents such an attempt. We collapse demand pattern into a single variable - demand expected arrival time; and disruption duration distribution is also represented by single variable – its variance. From Figure 9(a) we can see that the number of backup buses used decreases as the demand expected arrival time becomes larger. And its trend w.r.t. disruption duration variance is unimodal. It's interesting to see that initiation time has the

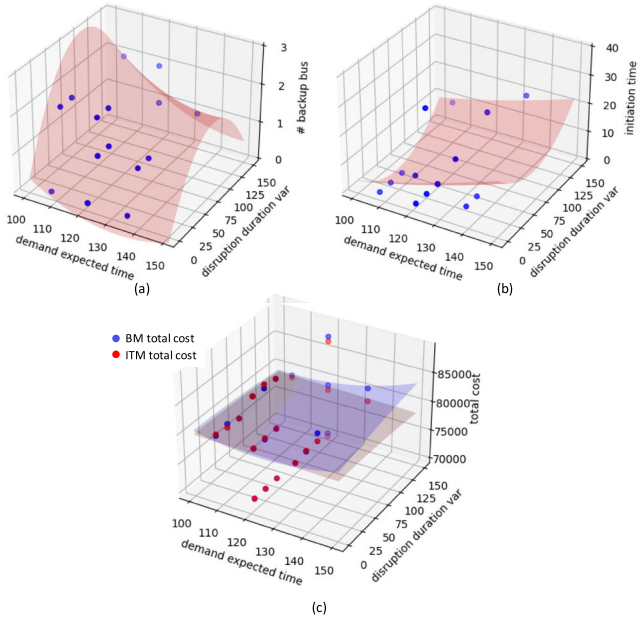


FIGURE 9. (a), (b) Test results illustrating how action variables are affected by demand expected arrival time and disruption duration variance; z-axis in (a) is the number of backup buses used; z-axis in (b) is the initiation time. (c) total cost comparisons between BM and ITM.

opposite relationship (Figure 9 (b)). BM and ITM total costs are compared in Figure 9 (c). As the disruption duration becomes more uncertain and demand arrivals more late, ITM performs the better.

IV. CONCLUSION

Typical urban zone transit systems are so complex and occupies an important component of urban carbon emissions, any attempt to find the optimal utilization of all resources and reduce energy consumption in a short period of time would encounter great difficulties. We choose to simplify the unit resource from run level to line level so that strategy selection can be optimized at a network level. Two models following this idea are proposed. They differ in the way they handle the uncertainty in disruption duration. When strategies are mapped into resource allocation, the resulting problem is classified as a nonconvex joint routing and resource allocation (nJRRA) problem. We propose a more constrained form that can be solved as a quadratic constrained quadratic programming (QCQP) problem. The assumptions and main ideas of the methodology in this study are summarized below:

- Disruption mitigation decision making is multi-leveled;
- The basic task unit of resource relocation model is average line service level;
- There is a trade-off between the user cost and operator cost;
- Disruption mitigation is a dynamic decision-making process.

To test the models, a quasi-dynamic evaluation program with a given incident duration distribution is constructed using discretized time steps and discrete distributions. FIFO conditions for users are incorporated with dynamic capacity assumptions to determine expected user costs under different strategies. Five different demand patterns and four different disruption distributions are tested on a small network. The optimal strategies for different combinations of demand pattern and disruption duration distributions are also obtained. Key insights include:

- The overall performance of ITM is the best compared to LLA, BB, and BM among the instances tested, although BM is not far behind and in some cases better.
- When the demand pattern is concave or the duration distribution is bi-Dirac/exponential-like, it is advantageous to postpone the resource relocation decision.
- When users tend to arrive in the near future, or when a disruption is likely to last for a long time, it makes less sense to delay actions.
- When the disruption distribution is exponential-like, a backup bus is not used in BB.
- ITM will not delay for more than an upper limit due to the penalties of user delay.

A network level strategy selection optimization model is formulated to tackle the transit disruption mitigation problem in a comprehensive and hierarchical way, which effectively reduces the disruption duration of the transportation system and improves transportation efficiency, and has a positive effect on reducing energy and carbon emissions. For future work, system states can be extended to be stochastic and partially observable, and multistage Markov decision processes can be modeled. Overlapping incidents may also be considered, such that resources allocated become unavailable for subsequent disruptions. User responses to mitigation plan could be modeled in a more complex way. User compliance ratios with respect to operator suggestions can be introduced to make the model more realistic. And test on larger networks, even real city networks, is also left for future work.

REFERENCES

[1] V. Cacchiani, D. Huisman, M. Kidd, L. Kroon, P. Toth, L. Veelenturf, and J. Wagenaar, "An overview of recovery models and algorithms for real-time railway rescheduling," *Transp. Res. B, Methodol.*, vol. 63, pp. 15–37, May 2014.

[2] A. Ceder, *Public Transit Planning and Operation: Modeling, Practice and Behavior*. Boca Raton, FL, USA: CRC Press, 2016.

[3] K. Kepaptsoglou and M. G. Karlaftis, "The bus bridging problem in metro operations: Conceptual framework, models and algorithms," *Public Transp.*, vol. 1, no. 4, pp. 275–297, Nov. 2009.

[4] L. Cadarso, Á. Marín, and G. Maróti, "Recovery of disruptions in rapid transit networks," *Transp. Res. E, Logistics Transp. Rev.*, vol. 53, pp. 15–33, Jul. 2013.

[5] A. Kiefer, S. Kritzinger, and K. F. Doerner, "Disruption management for the viennese public transport provider," *Public Transp.*, vol. 8, no. 2, pp. 161–183, Sep. 2016.

[6] B. G. Thengvall, J. F. Bard, and G. Yu, "Balancing user preferences for aircraft schedule recovery during irregular operations," *IIE Trans.*, vol. 32, no. 3, pp. 181–193, Mar. 2000.

[7] M. Yap and O. Cats, "Analysis and prediction of disruptions in metro networks," in *Proc. 6th Int. Conf. Models Technol. Intell. Transp. Syst. (MT-ITS)*, Jun. 2019, pp. 1–7.

[8] J. Jespersen-Groth, D. Potthoff, J. Clausen, D. Huisman, L. Kroon, G. Maróti, and M. N. Nielsen, "Disruption management in passenger railway transportation," in *Robust and Online Large-Scale Optimization*. Berlin, Germany: Springer, 2009, pp. 399–421.

[9] M. D. Hickman, "An analytic stochastic model for the transit vehicle holding problem," *Transp. Sci.*, vol. 35, no. 3, pp. 215–237, Aug. 2001.

[10] S. J. Berrebi, E. Hans, N. Chiabaut, J. A. Laval, L. Leclercq, and K. E. Watkins, "Comparing bus holding methods with and without real-time predictions," *Transp. Res. C, Emerg. Technol.*, vol. 87, pp. 197–211, Feb. 2018.

[11] B. Su, Z. Wang, S. Su, and T. Tang, "Metro train timetable rescheduling based on Q-learning approach," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Rhodes, Greece, Sep. 2020, pp. 1–6.

[12] E. Hassannayebi, A. Sajedinejad, A. Kardannia, M. Shakibayifar, H. Jafari, and E. Mansouri, "Simulation-optimization framework for train rescheduling in rapid rail transit," *Transportmetrica B, Transp. Dyn.*, vol. 9, no. 1, pp. 343–375, Jan. 2021.

[13] L. Zhu, S. Li, Y. Hu, and B. Jia, "Robust collaborative optimization for train timetabling and short-turning strategy in urban rail transit systems," *Transportmetrica B, Transp. Dyn.*, vol. 11, no. 1, pp. 147–173, Dec. 2023.

[14] A. I. Z. Jarrah, G. Yu, N. Krishnamurthy, and A. Rakshit, "A decision support framework for airline flight cancellations and delays," *Transp. Sci.*, vol. 27, no. 3, pp. 266–280, Aug. 1993.

[15] S. Zhan, L. G. Kroon, L. P. Veelenturf, and J. C. Wagenaar, "Real-time high-speed train rescheduling in case of a complete blockage," *Transp. Res. B, Methodol.*, vol. 78, pp. 182–201, Aug. 2015.

[16] L. P. Veelenturf, L. G. Kroon, and G. Maróti, "Passenger oriented railway disruption management by adapting timetables and rolling stock schedules," *Transp. Res. C, Emerg. Technol.*, vol. 80, pp. 133–147, Jul. 2017.

[17] F. Yuan, H. Sun, L. Kang, and S. Zhang, "Joint optimization of train scheduling and dynamic passenger flow control strategy with headway-dependent demand," *Transportmetrica B, Transp. Dyn.*, vol. 10, no. 1, pp. 627–651, Dec. 2022.

[18] W. Gu, J. Yu, Y. Ji, Y. Zheng, and H. M. Zhang, "Plan-based flexible bus bridging operation strategy," *Transp. Res. C, Emerg. Technol.*, vol. 91, pp. 209–229, Jun. 2018.

[19] L. Kang, X. Zhu, H. Sun, J. Wu, Z. Gao, and B. Hu, "Last train timetabling optimization and bus bridging service management in urban railway transit networks," *Omega*, vol. 84, pp. 31–44, Apr. 2019.

[20] S. Zhang and H. K. Lo, "Metro disruption management: Optimal initiation time of substitute bus services under uncertain system recovery time," *Transp. Res. C, Emerg. Technol.*, vol. 97, pp. 409–427, Dec. 2018.

[21] Y. Chen and K. An, "Integrated optimization of bus bridging routes and timetables for rail disruptions," *Eur. J. Oper. Res.*, vol. 295, no. 2, pp. 484–498, Dec. 2021.

[22] A. Löbel, "Optimale vehicle scheduling in public transit," Doctoral dissertation, Technische Universität Berlin, Berlin, Germany, 1997.

[23] M. Mesquita and J. Paixão, "Exact algorithms for the multi-depot vehicle scheduling problem based on multicommodity network flow type formulations," in *Computer-Aided Transit Scheduling*. Berlin, Germany: Springer, 1999, pp. 221–243.

[24] D. Huisman, R. Freling, and A. P. M. Wagelmans, "A robust solution approach to the dynamic vehicle scheduling problem," *Transp. Sci.*, vol. 38, no. 4, pp. 447–458, Nov. 2004.

[25] A. Mingozzi, M. A. Boschetti, S. Ricciardelli, and L. Bianco, "A set partitioning approach to the crew scheduling problem," *Oper. Res.*, vol. 47, no. 6, pp. 873–888, Dec. 1999.

[26] G. Yu, M. Argüello, G. Song, S. M. McCowan, and A. White, "A new era for crew recovery at continental airlines," *Interfaces*, vol. 33, no. 1, pp. 5–22, Feb. 2003.

[27] M. Mesquita and A. Paias, "Set partitioning/covering-based approaches for the integrated vehicle and crew scheduling problem," *Comput. Oper. Res.*, vol. 35, no. 5, pp. 1562–1575, May 2008.

[28] M. S. Visentini, D. Borenstein, J.-Q. Li, and P. B. Mirchandani, "Review of real-time vehicle schedule recovery methods in transportation services," *J. Scheduling*, vol. 17, no. 6, pp. 541–567, Dec. 2014.

[29] L. Xu and T. S. A. Ng, "A robust mixed-integer linear programming model for mitigating rail transit disruptions under uncertainty," *Transp. Sci.*, vol. 54, no. 5, pp. 1388–1407, Sep. 2020.

[30] Y. Hamdouch and S. Lawphongpanich, "Schedule-based transit assignment model with travel strategies and capacity constraints," *Transp. Res. B, Methodol.*, vol. 42, nos. 7–8, pp. 663–684, Aug. 2008.

[31] J. G. Jin, K. M. Teo, and A. R. Odoni, "Optimizing bus bridging services in response to disruptions of urban transit rail networks," *Transp. Sci.*, vol. 50, no. 3, pp. 790–804, Aug. 2016.

[32] H. Sun, J. Wu, L. Wu, X. Yan, and Z. Gao, "Estimating the influence of common disruptions on urban rail transit networks," *Transp. Res. A, Policy Pract.*, vol. 94, pp. 62–75, Dec. 2016.

[33] S. Bekhor, M. E. Ben-Akiva, and M. S. Ramming, "Evaluation of choice set generation algorithms for route choice models," *Ann. Oper. Res.*, vol. 144, no. 1, pp. 235–247, Apr. 2006.

[34] S. Bekhor, T. Toledo, and J. N. Prashker, "Effects of choice set size and route choice models on path-based traffic assignment," *Transportmetrica*, vol. 4, no. 2, pp. 117–133, Jan. 2008.

[35] E. Cascetta, F. Russo, and A. Vitetta, "Stochastic user equilibrium assignment with explicit path enumeration: Comparison of models and algorithms," *IFAC Proc. Volumes*, vol. 30, no. 8, pp. 1031–1037, Jun. 1997.

[36] Q. Liu and J. Y. J. Chow, "A congested schedule-based dynamic transit passenger flow estimator using stop count data," *Transportmetrica B, Transp. Dyn.*, vol. 11, no. 1, pp. 231–256, Dec. 2023.

[37] E. Castillo, Z. Grande, A. Calviño, W. Y. Szeto, and H. K. Lo, "A state-of-the-art review of the sensor location, flow observability, estimation, and prediction problems in traffic networks," *J. Sensors*, vol. 2015, no. 1, pp. 1–26, 2015.

[38] R. A. Tinguely, K. J. Montes, C. Rea, R. Sweeney, and R. S. Granetz, "An application of survival analysis to disruption prediction via random forests," *Plasma Phys. Controlled Fusion*, vol. 61, no. 9, Sep. 2019, Art. no. 095009.

[39] B. Gendron, T. G. Crainic, and A. Frangioni, "Multicommodity capacitated network design," in *Telecommunications Network Planning*. Boston, MA, USA: Springer, 1999, pp. 1–19.

[40] L. Xiao, M. Johansson, and S. P. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Trans. Commun.*, vol. 52, no. 7, pp. 1136–1144, Jul. 2004.

[41] A. A. El-Sherif and A. Mohamed, "Joint routing and resource allocation for delay minimization in cognitive radio based mesh networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 186–197, Jan. 2014.

[42] M. Eslami Rasekh, D. Guo, and U. Madhow, "Joint routing and resource allocation for millimeter wave picocellular backhaul," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 783–794, Feb. 2020.

[43] P. M. Pardalos and S. A. Vavasis, "Quadratic programming with one negative eigenvalue is NP-hard," *J. Global Optim.*, vol. 1, no. 1, pp. 15–22, 1991.

[44] F. A. Al-Khayyal, C. Larsen, and T. Van Voorhis, "A relaxation method for nonconvex quadratically constrained quadratic programs," *J. Global Optim.*, vol. 6, no. 3, pp. 215–230, Apr. 1995.

[45] C. Audet, P. Hansen, B. Jaumard, and G. Savard, "A branch and cut algorithm for nonconvex quadratically constrained quadratic programming," *Math. Program.*, vol. 87, no. 1, pp. 131–152, Jan. 2000.

[46] P. G. Furth and N. H. Wilson, "Setting frequencies on bus routes: Theory and practice," *Transp. Res. Rec.*, vol. 818, no. 1, pp. 1–7, 1981.

QI LIU was born in Zhongxiang, Hubei, China, in 1988. He received the B.S. degree in transportation engineering from Huazhong University of Science and Technology, China, in 2006, the M.S. degree in transportation informatics and control from Tongji University, Shanghai, China, in 2011, and the Ph.D. degree in transportation planning and engineering from New York University. He is currently a Postdoctoral Researcher with the College of Transportation Engineering, Tongji University. His research interests include transportation system modeling and optimization, connected vehicles, and intelligent transportation systems.

WEI WANG was born in Yantai, Shandong, China. He received the master's degree from Cornell University and the Ph.D. degree in civil engineering from Columbia University, USA. He is a Chief Engineer with the China Building Technique Group, China Academy of Building Research. His research interests include smart city, low/zero carbon emission infrastructures, operation, and maintenance management in urban infrastructures. He has published more than 20 articles in his field of interest. He is the author of the books *Low Carbon and Smart Operation and Maintenance for Urban Zone Infrastructure*. He is a Committee Member of China Engineering Construction Standardization Association (CECS), Chinese Society for Urban Studies Digital Twin and Future Cities Specialty Committee, and China Association for Science and Technology Low Carbon and Smart City Specialty Committee.

YUMIAO LIU was born in Shijiazhuang, Hebei, China. She received the master's degree from Hebei University of Engineering, Handan, China.

She participates in the National Key Research and Development Project of China. Her research interests include concrete structures and prestressed concrete structures, building industrialization, and low-carbon and smart cities. She has published five articles and applied for seven invention patents in her field of interest.

YANZHAO SU was born in Zhengzhou, Henan, China. He was a Postdoctoral Fellow with Tsinghua University, Beijing, China. He was a joint Ph.D. Scholar with the University of Michigan, USA, in 2018. He hosts and participates at the National Natural Science Foundation of China and the National Key Research and Development Project. His research interests include autonomous driving technology, software-defined vehicles, and new energy vehicle design. He has published more than 20 articles and applied for more than 50 national invention patents in his field of interest. He is the co-author of the books *Automotive Scale Chip Technology* and *Automotive Autonomous Driving*.