Transportation Research Part C 182 (2026) 105441

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Incorporating reservation strategies into demand management and train scheduling in metro systems

Yahan Lu [a,b], Lixing Yang [a,c,*], Dongyang Xia [b,*], Fanting Meng [d], Shadi Sharif Azadeh [b]

[a] School of Systems Science, Beijing Jiaotong University, Beijing, 100044, China
[b] Department of Transport & Planning, Delft University of Technology, the Netherlands
[c] Hebei Key Laboratory of Future Urban Intelligent Traffic Management, Beijing Jiaotong University, Beijing, 100044, China
[d] College of urban rail transit and logistics, Beijing Union University, Beijing, 100101, China

ARTICLE INFO

Keywords:
Urban rail transit
Reservation-based travel
Time-varying reservation slot allocation
Passenger flow control
Adaptive large neighborhood search

ABSTRACT

Emerging reservation-based travel technologies offer a promising solution to mitigate supply-demand mismatches in metro systems. This paper presents a framework to support metro operators by optimizing time-varying reservation slot allocation plans, passenger flow control strategies, and train schedules. The proposed approach ensures that passengers with reservations can directly access platforms and board the first available train services, while those without reservations are managed through effective passenger flow control strategies to optimize train capacity utilization. To address this, an integer nonlinear programming model is formulated, incorporating constraints that capture interactions between passengers with and without reservations, with the objective of minimizing passengers' waiting time and line congestion. A hybrid algorithm is developed to improve computational efficiency, combining the adaptive large neighborhood search method with a commercial solver and incorporating valid inequalities tailored to the properties of the model. The effectiveness of the proposed approaches is demonstrated through numerical experiments using real-world operational data from the Beijing metro Batong line. Computational results indicate that the integrated optimization approach reduces the objective value by 6.19 % compared to a step-by-step optimization method, achieving better alignment of capacity with dynamic passenger flows. In addition, the extreme unfairness between reserved and unreserved passengers, where passengers with reservations have a 100 % service ratio compared to less than 20 % for unreserved passengers, is mitigated by increasing passenger waiting times by 3.51 % and line congestion by 0.51 %. Furthermore, the proposed algorithm efficiently solves large-scale and real-world instances, outperforming the state-of-the-art commercial solver.

1. Introduction

For a long time, effectively addressing the highly concentrated commuting demand in both time and space has posed significant challenges to the operations and management of urban rail transit systems. Initially, the main solutions involved increasing transportation capacity by constructing new lines or reducing train departure headways. However, urban physical space utilization

* Corresponding authors.
E-mail addresses: yahanlu@tudelft.nl (Y. Lu), lxyang@bjtu.edu.cn (L. Yang), d.xia@tudelft.nl (D. Xia), cgtfanting@buu.edu.cn (F. Meng), s.sharifazadeh@tudelft.nl (S. Sharif Azadeh).

https://doi.org/10.1016/j.trc.2025.105441
Received 20 February 2025; Received in revised form 3 November 2025; Accepted 4 November 2025
Available online 13 November 2025

in megacities worldwide is nearing saturation, and the minimum train headway is approaching the two-minute limit. As a result, passenger flow control strategies have become widely adopted as demand management measures in urban rail transit systems in cities such as Beijing, Shanghai, and Guangzhou (see, Beijing Daily, 2021; Guangzhou Municipal People's Government, 2023; Shentong Metro Group, 2024). These strategies manage platform crowding by controlling the number of passengers waiting here, thereby reducing the risks of train delays and operational disruptions.

With rapid advancements in technological innovations, traditional metro systems are undergoing accelerated transformation and upgrades. A notable example is the Beijing metro's introduction of a travel reservation system in March 2020 (People's Daily Beijing, 2020). This system allows passengers to reserve entry slots for the following day via an official mobile app, with booking windows opening twice daily at 12:00 and 20:00 on a first-come, first-served basis. Reserved passengers can access stations through a fast-track lane, while non-reserved passengers continue to use regular entry channels. During its one-year pilot phase, this system facilitated 1.17 million trips, helped passengers collectively save 40,000 hours of queuing time, and reduced average waiting time by 3.5 minutes per person per day (Beijing Municipal Commission of Transport, 2020). In addition to Beijing, Chongqing has also implemented a similar reservation-based entry system since March 2020 to manage peak-hour passenger flow, adopting comparable mechanisms of voluntary booking and fast-track access (Ministry of Transport of the People's Republic of China, 2021). These outcomes demonstrate the initiative's effectiveness in improving operational efficiency and service quality, offering a promising solution for the management of intelligent urban rail transit systems.

Reservation-based travel modes provide high-quality services but also introduce significant complexity to the operation and management of urban rail transit systems. In these systems, reserved and unreserved passengers must share limited station space and train capacity, creating strong interdependencies. An unreasonable allocation of reservation slots would result in considerable stranding of passengers without reservations, reducing their travel experience. Moreover, the core issue of passenger flow oversaturation stems from a mismatch between supply and demand, which requires simultaneous adjustment of supply-side resource allocation and demand management strategies to address effectively. As demonstrated by related literature, addressing these challenges requires an integrated approach that optimizes both supply-side resource allocation and demand management strategies jointly (e.g., Shi et al., 2018; Lu et al., 2022). However, in practical operations, demand management strategies and train schedules are often designed subjectively by operators, step by step, which limits their effectiveness in resolving such mismatches.

To this end, jointly optimizing demand management and train scheduling with reservation strategies is crucial. Such collaboration ensures a better alignment of capacity and demand while balancing the needs of both reserved and unreserved passengers. However, most existing research focuses on isolated aspects of the problem, primarily falling into two categories: dynamic demand-oriented schedule optimization (e.g., Cacchiani et al., 2020; Binder et al., 2021; Gkiotsalitis et al., 2022) and integrated optimization of train schedules and passenger flow control strategies (e.g., Yuan et al., 2022; Lu et al., 2023; Yuan et al., 2023). More recently, Tang et al. (2024) investigated the joint optimization of reservation slot allocation at a single station with dynamic passenger demand. This study excludes train scheduling and system-wide demand management across entire lines from its scope and assumes the given train schedule has evenly distributed headway. Yang et al. (2025) studied the integrated demand-side management and timetabling problem, focusing on incentivizing non-reserved passengers to shift their departure time based on a given number of reservation slots. To sum up, these studies either restrict their scope to a single station or treat reservation slot allocations as fixed inputs. Consequently, there remains a gap in integrating train scheduling, reservation slot allocation, and passenger flow control across an entire line, which is essential for achieving system optimality.

To address the aforementioned challenges, this paper presents an integrated optimization framework for the demand management and train scheduling problem. The framework generates system-optimal dynamic reservation slot allocations, passenger flow control strategies, and train schedules for an urban rail transit line on a daily basis. We formulate the problem as an integer nonlinear program, aiming to minimize passengers' waiting times and *line congestion*, defined as the maximum number of waiting passengers across stations when a train arrives. In this study, we define *fairness* between passengers with and without reservations as providing both groups with equitable opportunities to access train services without systematic disadvantage to either group. An allocation is considered unfair if one group is consistently denied boarding opportunities while capacity remains available for the other group. In the proposed framework, passengers with reservations are allowed to access the platform upon arrival and board the first available train. In contrast, non-reserved passengers must wait and are admitted based on a passenger flow control strategy, which introduces the risk of delayed boarding. To operationalize the above fairness concept, we impose constraints in our model that guarantee a minimum boarding proportion for non-reserved passengers, ensuring that both groups can access train services in accordance with commonly accepted principles of fairness in public transportation. To efficiently solve large-scale instances, we develop a hybrid algorithm that integrates an adaptive neighborhood search with tailored destroy operators and a commercial solver This approach enables the computation of high-quality solutions within the acceptable computational time.

In summary, the main contributions of this paper are as follows:

(i) Formalizing the demand management and train scheduling problem with reservation strategies. We consider the interactions between passengers and trains, passengers with and without reservations, as well as capacity limitations.

(ii) Developing an integer non-linear formulation aimed at minimizing passengers' waiting time and line congestion. The proposed model is linearized by introducing additional binary variables and big-M constraints.

(iii) Designing a framework to decompose the problem into a train scheduling subproblem and a demand management subproblem. To efficiently solve the problem, we develop a solution algorithm combining the adaptive large neighborhood search (ALNS) and a commercial solver. In the ALNS framework, we develop destroy operators tailored to our formulation. This approach eliminates the need for additional variables and constraints required for linearization while incorporating valid inequalities to enhance solution performance.

Table 1
Overview of relevant studies.

Publications	Domains	Train scheduling	Demand management strategies	
			Passenger flow control	Reservation slot allocation
Bertsimas and De Boer (2005)	Airline			✔
Shi et al. (2018)	Metro lines	✔	✔	
Cacchiani et al. (2020)	Railway lines	✔		
Molnar and De Almeida Correia (2019)	Carpooling			✔
Polinder et al. (2021)	Railway networks	✔		
An et al. (2021)	Airline			✔
Binder et al. (2021)	Railway networks	✔		
Zhu and Goverde (2021)	Railway networks	✔		
Lu et al. (2022)	Metro lines	✔	✔	
Yin et al. (2021)	Metro networks	✔	✔	
Li et al. (2023)	Road traffic			✔
Lu et al. (2023)	Metro lines	✔	✔	
Xia et al. (2023)	Bus lines	✔		
Wang et al. (2024)	Metro lines	✔		
Xia et al. (2024b)	Bus networks	✔		
Tang et al. (2024)	Metro stations		✔	At a station
Yang et al. (2025)	Metro lines	✔	✔	
This paper	Metro lines	✔	✔	All stations

(iv) Showing that the proposed algorithm outperforms a state-of-the-art commercial solver in terms of computational efficiency, while still providing metro operators with high-quality strategies for reservation slot allocation, passenger flow control, and matched train schedules.

(v) Demonstrating that neglecting the feedback between train scheduling and demand management subproblems leads to inefficient timetables with longer passenger waiting times and higher line congestion, and highlighting the strong time-varying nature of optimal reservation slot allocation plans.

The remainder of this paper is organized as follows. Section 2 reviews the literature relevant to this research. A detailed description of the problem addressed in this paper is provided in Section 3. Section 4 introduces the formulated models and presents the analysis of their complexity. The proposed model decomposition framework and solution algorithm are discussed in Section 5. Numerical experiments, including sensitivity analyses, are presented in Section 6. Lastly, we conclude in Section 7.

2. Literature review

The train scheduling problem is an important topic in the transportation planning field (e.g., Zhu and Goverde, 2019, 2021; Polinder et al., 2021; Binder et al., 2021; Yang et al., 2021; Schettini et al., 2023; Gong et al., 2024; Li et al., 2024). For an overview of various models and solution methodologies under regular operational conditions, we refer to Scheepmaker et al. (2017). For a comprehensive review of train rescheduling, see Zhan et al. (2024). To position our work within the broader field, we highlight relevant studies from three interconnected research streams: train scheduling and demand management strategies in metro systems, reservation-based demand management strategies in other transportation contexts, and dynamic capacity allocation strategies in public transit systems. An overview of the relevant literature is presented in Table 1.

2.1. Train scheduling and demand management strategies in metro systems

Demand management strategies in metro systems include passenger flow control and reservation-based travel mechanisms. In the following, we review studies on train scheduling combined with passenger flow control, as well as train scheduling integrated with reservation-based travel strategies in metro systems.

(i) **Train scheduling and passenger flow control in metro systems.** Most studies on train scheduling and demand management focus on optimizing train timetables and passenger flow control, with the objectives of minimizing passenger waiting times. In this context, one line of related research focuses on timetabling strategies where each train stops at every station it passes. For example, Shi et al. (2018) proposed an integer linear programming model with the time-dependent passenger demand on a single-direction line, which can be solved by a hybrid algorithm combining local search and CPLEX. Liu et al. (2020) developed a collaborative optimization model with headway-based passenger demand on a bi-directional line and designed a Lagrangian relaxation-based heuristic approach. Lu et al. (2022) proposed a two-stage distributionally robust optimization model to address this problem under uncertain demand and developed a tailored decomposition framework to solve the formulation. Yuan et al. (2022) formulated a mixed-integer nonlinear programming (MINLP) model for this problem on urban rail transit networks. Lu et al. (2023) formulated three robust models to balance passengers' waiting time and service fairness. Among these, the scenario-based formulation is solved using an algorithm that combines an improved local search with GUROBI. Recently, Liang et al. (2024) proposed an online optimization policy to address the collaborative train timetable and passenger flow control problem under stochastic demand, aiming to generate timetables and passenger flow control plans for each stochastic scenario.

Another stream of research on the train scheduling and passenger flow control problem explores scheduling with skip-stop operational patterns, where trains skip certain stations to enhance the efficiency of capacity utilization. Applications include robust optimization under demand uncertainty on a metro line (Hu et al., 2023), dynamic programming approaches in metro networks (Yuan et al., 2023), local search algorithms addressing elastic demand (Shi et al., 2023), and discrete Markov decision models integrating energy-saving strategies (Zhang et al., 2024b).

Our study contributes to this field by integrating reservation strategies into demand management and train scheduling, enhancing the alignment between passenger demand and capacity resources. This approach equips metro managers with high-quality reservation slot allocation and passenger flow control plans. Furthermore, we incorporate service fairness among passengers across different stations to ensure more equitable service delivery.

(ii) **Train scheduling and reservation-based travel strategy in metro systems.** Reservation strategies are an emerging concept in the context of metro systems. To the best of our knowledge, only two studies have addressed a similar topic. Tang et al. (2024) investigated the reservation slot allocation problem at a single station on a metro line, assuming a fixed train schedule with evenly spaced headways. The authors proposed a multi-objective optimization model and developed an iterative sequential search algorithm combined with GUROBI. Computational results based on instances from a Beijing metro line demonstrate that the joint optimization model with reservation strategies reduces the number of stranded passengers by 88.46 % compared to the original passenger flow. Yang et al. (2025) explored train scheduling, trip-shifting, and passenger flow control strategies on metro lines in the context of trip reservations. They designed an exact method to solve the formulated model and validated the effectiveness of the proposed approaches in the Beijing metro system.

Our study builds on these prior works by jointly optimizing train scheduling and demand management at the metro line level rather than focusing on an individual station, thereby achieving system-optimal solutions. Compared to Tang et al. (2024), our contribution lies in the integrated optimization of train schedules and reservation-based demand management across all stations, rather than at a single station. Additionally, we ensure that the allocation of reservation slots is socially beneficial by granting reserved passengers direct access to stations while still accommodating a portion of unreserved passengers, preventing them from being entirely stranded due to a lack of reservations. Compared to Yang et al. (2025), we advance the field by jointly optimizing reservation strategies with train schedules and passenger flow control, rather than treating reservation slots as fixed parameters. This provides a proof of concept for the appropriate allocation of reservation slots across both time and stations. In the following experiments, we show the importance of jointly optimizing time-varying reservation slots along with train schedules and passenger flow control strategies.

2.2. Reservation-based travel strategies in other transportation systems

Reservation-based travel strategies are well-established in the airline industry and have also been explored in road traffic systems. In the airline industry, such strategies typically involve allocating identical seats at different price levels based on booking classes to improve revenue (e.g., Lee and Hersh, 1993; Robinson, 1995; Bertsimas and De Boer, 2005; Jiang and Barnhart, 2009). For example, An et al. (2021) studied the multi-fare, network revenue management problem in the airline industry, aiming to generate an optimal booking-limit policy.

In road traffic, reservation-based travel strategies have been proposed to mitigate congestion during rush hours and are often combined with congestion pricing strategies for private cars, see, Liu et al. (2015), Lamotte et al. (2017), Li et al. (2023). For reservation-based carpooling services, reservation-based travel strategies are designed to increase the percent of requests matched (e.g., Molnar and De Almeida Correia, 2019). For example, Ouyang et al. (2021) examined many-to-many carpooling services with advance reservations, and constraints on waits and detours.

Our study contributes to this field by extending reservation-based travel strategies to metro systems, proposing a mathematical framework, designing an algorithm for this problem, and showing benefits in real-life operations.

2.3. Dynamic capacity allocation strategies in public transit systems

With the increasing variability of passenger flows, dynamic capacity allocation based on time-dependent passenger demand has become an effective strategy to improve the alignment between demand and supply. There are two main approaches to dynamic resource allocation. The first is to adjust train schedules by optimizing non-uniform departure intervals to better match fluctuating passenger flows, as reviewed in Section 2.1. The second approach involves dynamically optimizing train timetables and vehicle compositions to achieve cost reductions and efficiency gains, leveraging advanced vehicle technologies such as virtual coupling in metro systems and modular vehicles in bus systems. In this line of research, the demand management approach is usually neglected and most of the related research focuses on supply-side resource allocation. For example, Wang et al. (2024) proposed a real-time optimization framework to jointly optimize train timetables and rolling stock schedules under a virtual coupling strategy at the line level, accounting for dynamic passenger demand. Xia et al. (2023) and Xia et al. (2024a) explored timetabling and dynamic capacity allocation for a bus line and for an intermodal system that integrates a bus line with demand-responsive services, where modular vehicles can be decoupled and coupled at different locations and times. Furthermore, Xia et al. (2024b) proposed an integrated optimization framework for timetabling, vehicle scheduling, and dynamic capacity allocation of modular vehicles at the network level, allowing vehicles to be decoupled and coupled at depots and transfer stations and flexibly dispatched across multiple lines.

We contribute to this field by integrating demand-side management strategies with supply-side resource allocation, considering dynamic passenger demand. By jointly optimizing both passenger dynamics and train schedules, our approach enables a more effective alignment between demand and capacity, and facilitates system-wide optimization.

3. Problem description

This paper addresses a tactical-level train scheduling problem with integrated demand management strategies, including reservation slot allocations and passenger flow control. It incorporates range constraints on train operations, train capacity limitations, service fairness, and passenger dynamics. This section provides a detailed discussion of these elements.

3.1. Infrastructure and train scheduling

We consider a metro line that provides infrastructure information about the stations and the tracks connecting them. The set of stations is represented as $\mathcal{S}$. For every pair of adjacent stations $s \in \mathcal{S}$ and $v \in \mathcal{S} \setminus \{s\}$, tracks exist to facilitate train travel.

During the study time horizon, a fixed number of train services, denoted by the set $\mathcal{I}$, are operated with a uniform capacity C. Each train service is required to stop at every station, with the dwell time of train service $i \in \mathcal{I}$ at station $s \in \mathcal{S}$ represented as ε_{is}. The running time of train service i from station s to station $s+1$ is denoted by η_{is} for all $i \in \mathcal{I}$ and $s \in \mathcal{S} \setminus \{|\mathcal{S}|\}$. The headway of two consecutive trains departing from the same station has to follow maximum and minimum boundaries (denoted as $\underline{h}$ and $\overline{h}$, respectively) for operational safety and efficiency. To improve the alignment between capacity and demand, we determine the departure and arrival times of train service i at station s, denoted by d_{is} and a_{is}, respectively, for all $i \in \mathcal{I}$ and $s \in \mathcal{S}$.

3.2. Demand management strategies

To model the time-dependent passenger demand within metro systems, the study time horizon is discretized into a finite number of timestamps, represented by the set $\mathcal{T}$. The number of passengers arriving at station $s \in \mathcal{S}$ at timestamp $t \in \mathcal{T}$ and heading to destination $v \in \mathcal{S}$, where $v \geq s+1$, is denoted by D_{svt}.

We propose two demand management strategies: time-varying reservation slot allocation and train-based passenger flow control. The time-varying reservation slot allocation ensures that passengers with reservations for each origin-destination (OD) pair from station s to v at timestamp t can access platforms immediately upon arrival at the origin station and depart on the first available train service. This strategy guarantees priority boarding for reserved passengers, removing the risk of being stranded in their travel experience. We define an integer variable ϱ_{svt} for all $s, v \in \mathcal{S}, v > s, t \in \mathcal{T}$ to model this strategy. In contrast, passenger flow control manages passengers without reservations by requiring them to queue outside the gates until entry permission is granted. Only the number of passengers that can board the arriving train according to the passenger flow control strategies are allowed to pass the gates, enter the platform, and wait to board. As a result, passengers without reservations may face the risk of being stranded. The train-based passenger flow control is defined as an integer variable b_{isv} for all $i \in \mathcal{I}, s, v \in \mathcal{S}, v > s$.

Considering that reserved and unreserved passengers share the same train capacity, these two demand management strategies must be optimized jointly. The goal is to ensure that reserved passengers can board their designated trains while also preventing all unreserved passengers from being stranded. This balance aims to achieve fairness between passengers with and without reservations by prioritizing the needs of reserved passengers while still providing opportunities for unreserved passengers to access the service. Additionally, we incorporate fairness among unreserved passengers at different stations into the demand management. To achieve this, we formulate constraints ensuring that train i serves at least κ_{isv} (%) of the passengers waiting at station s with a destination of station v. This ensures that unreserved passengers at different stations are treated equitably, preventing disproportionate stranding at upstream stations.

In summary, this paper develops a mathematical formulation and an algorithmic framework to integrate reservation strategies into the demand management and train scheduling problem. The objective is to minimize the total waiting time of passengers and the line congestion experienced by the most overcrowded train service. The total waiting time of passengers consists of two components: the time spent waiting for a train and the additional time incurred due to being stranded and waiting for the next available service. Our formulation explicitly accounts for the boarding prioritization of passengers with reservations, the interaction between passengers with and without reservations, and service fairness among unreserved passengers at different stations. The mathematical formulation is built upon the following four key assumptions.

Assumption 1. We assume that historical and time-varying passenger demand data are known and deterministic. The train schedules and demand management strategies are optimized based on these data and are intended to be implemented repeatedly on typical weekdays without daily re-optimization.

Assumption 2. As we consider the system-optimal solution, passengers are assumed to be homogeneous, and demand is treated as exogenous.

Assumption 3. Reservation slots are released via a mobile application on the day before operation, allowing passengers to make reservations independently. All reservation slots are assumed to be fully booked, based on evidence from the pilot implementation in Beijing.

Assumption 4. We assume that passengers with reservations will show up, and their specific behavior is beyond the scope of this study.

Assumption 1 reflects common practice in metro operations, where tactical plans are developed based on stable demand patterns and applied consistently across similar days to ensure operational stability and ease of implementation. This assumption is widely adopted in metro scheduling and demand management studies (e.g., Yin et al., 2023; Lu et al., 2023; Tang et al., 2024). **Assumption 2** is standard in system-optimal models and has been widely adopted in related studies (e.g., Robenek et al., 2018). **Assumption 3** is aligned with the operational procedures observed in the Beijing metro. **Assumption 4** is reasonable from an operational perspective,

as reservation slots are mainly used during weekday peak hours when passengers typically have rigid commuting needs, ensuring both their participation and their preference for a more comfortable journey.

4. Model formulation

In Section 4.1, we introduce the notations and decision variables. The constraints associated with timetabling, passenger dynamics, and interactions between passengers with and without reservations are formulated in Section 4.3. Subsequently, we present the objective function and formulate an INLP model to address the demand management and train scheduling problem incorporating reservation strategies in Section 4.2. In Setion 4.4, we introduce the big-M constraints to equivalently transform the INLP model into a linear formulation, derive the big-M values, and obtain an ILP model with a tighter upper bound. Lastly, the complexity of the proposed integer nonlinear programming (INLP) and linear programming models is analyzed in Section 4.5.

4.1. Notations

To formally define the problem of interest, we begin by listing all the notations used in the modeling process, as summarized in Table 2. This study focuses on simultaneously optimizing the train timetable, passenger flow control strategies, and reservation slot plans. Accordingly, three types of decision variables are introduced and detailed below.

(i) x_{ist}: Binary variable. If train service i has departed from station s at time t, $x_{ist} = 0$; otherwise, $x_{ist} = 1$.

(ii) ϱ_{st}: Integer variable. This variable represents the number of reservation slots at station s and time t.

(iii) b_{isv}: Integer variable. This variable indicates the number of passengers without reservations who are allowed to board train service i at station s and go to station v.

4.2. Objective function

As the backbone of urban transportation, urban rail transit systems have to balance efficiency and safety in their operations. In this context, this paper aims to effectively mitigate operational risks while ensuring service quality by minimizing the weighted sum of passenger waiting time and line congestion, i.e.,

$$\min \quad \lambda^e F^e + \lambda^c F^c \tag{1}$$

$$F^e = \Delta \left[\sum_{i\in I} \sum_{s\in S} \sum_{t\in \mathcal{T}} \left(y_{ist} \sum_{t'\in \mathcal{T}, t'\le t} y_{ist'} \sum_{v\in S, v>s} D_{svt} \right) + \sum_{i\in I} \sum_{s\in S} \sum_{t\in \mathcal{T}} \left(y_{ist} \sum_{v\in S, v>s} r_{isv} \right) \right], \tag{2}$$

$$F^c = \sum_{i\in I} z_i, \tag{3}$$

where λ^e and λ^c represent the weighting coefficients. F^e denotes the total waiting time of passengers, while F^c represents the total line congestion defined as the total maximum number of waiting passengers across stations when trains arrive, which will be introduced in the following subsection in detail.

Constraints (2) calculate the total waiting time, including both the waiting time of newly arriving passengers and that of stranded passengers within each headway. Here, $y_{ist} = 1$ if time t falls within the headway between train services $i-1$ and i. Example 1 illustrates the interpretation of $\mathbf{y}$, while Example 2 clarifies the first term of constraints (2), specifically the waiting time of newly arriving passengers. Constraints (3) define the level of congestion experienced by trains along the metro line, ensuring a more balanced distribution of passenger loads across stations.

Example 1. The 0–1 variables representing the departures of trains $i-1$ and i from station s, along with their corresponding headway, are illustrated in Fig. 1. In this example, trains $i-1$ and i depart from station s at timestamps 2 and 4, respectively, making the second and third timestamps part of the headway duration.

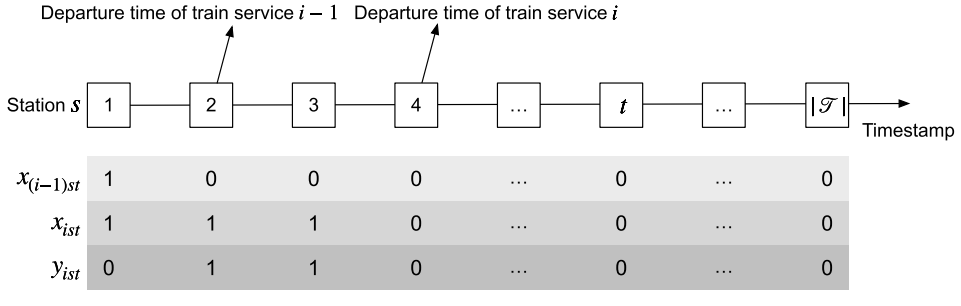

Fig. 1. Illustration of 0–1 binary variables related to the departure time and headway.

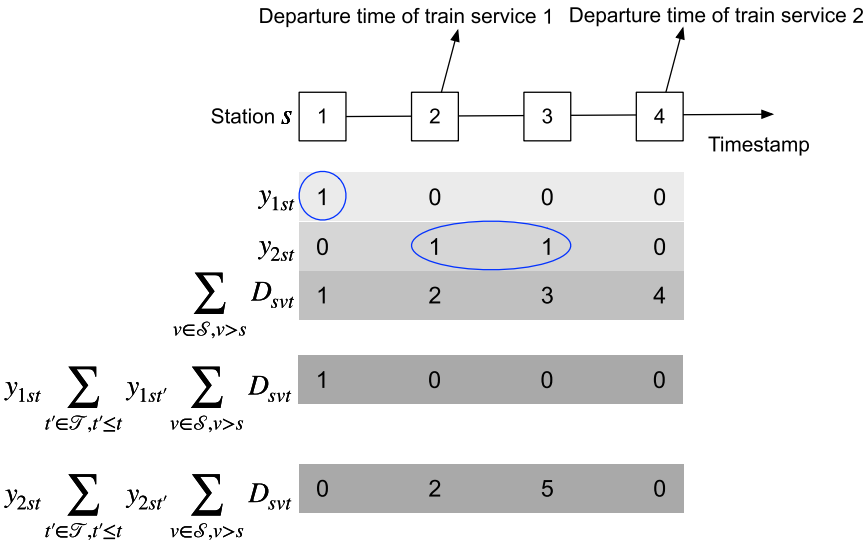

Fig. 2. Illustration of newly arriving passengers' waiting time.

Example 2. Fig. 2 illustrates an example of computing the waiting time for newly arriving passengers within each headway, corresponding to the first term in constraints (2). Trains 1 and 2 depart from station s at timestamps 2 and 4, respectively. As a result, the first timestamp corresponds to the first headway, while the second and third timestamps belong to the second headway. At each timestamp, 1, 2, 3, and 4 passengers arrive sequentially. Consequently, one passenger is waiting before train 1 departs, while two and five newly arriving passengers are waiting at timestamps 2 and 3, respectively, before train 2 departs.

Thus, the waiting time of newly arriving passengers during the first headway is calculated as:

$$\Delta \sum_{s\in S}\sum_{t\in\mathcal{T}}\left(y_{1st}\sum_{t'\in\mathcal{T},t'\le t} y_{1st'}\sum_{v\in S,v>s} D_{svt}\right)=1,$$

and during the second headway is

$$\Delta \sum_{s\in S}\sum_{t\in\mathcal{T}}\left(y_{2st}\sum_{t'\in\mathcal{T},t'\le t} y_{2st'}\sum_{v\in S,v>s} D_{svt}\right)=2+5=7.$$

4.3. Constraints

The constraints are categorized into scheduling-related constraints, reservation slot allocation plan constraints, constraints related to the dynamic evolution of passengers without reservations, and constraints of interactions between passengers with and without reservations.

(i) **Scheduling-related constraints.** To facilitate the linear modeling of interactions between trains and passengers, motivated by Xia et al. (2024b), binary variables **x** and **y** are introduced to represent the train operating dynamics. Furthermore, train operations must adhere to headway limitations to ensure operational safety, i.e.,

$$x_{is(t+1)} \le x_{ist} \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, t\in\mathcal{T}\setminus\{|\mathcal{T}|\}, \tag{4}$$

$$x_{is|\mathcal{T}|}=0 \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{5}$$

$$d_{is}=\sum_{t\in\mathcal{T}\setminus\{1\}}[t(x_{is(t-1)}-x_{ist})] \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{6}$$

$$\underline{h}\le d_{is}-d_{(i-1)s}\le\overline{h} \qquad \forall i\in\mathcal{I}\setminus\{1\}, s\in\mathcal{S}, \tag{7}$$

$$d_{is}=a_{is}+\varepsilon_{is} \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{8}$$

$$a_{is}=d_{i(s-1)}+\eta_{i(s-1)} \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}\setminus\{1\}, \tag{9}$$

$$y_{ist}=\begin{cases} x_{ist} & i=1\\ x_{ist}-x_{(i-1)st} & i\in\mathcal{I}\setminus\{1\}\end{cases} \qquad \forall s\in\mathcal{S}, t\in\mathcal{T}, \tag{10}$$

$$x_{ist}\in\{0,1\} \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, t\in\mathcal{T}. \tag{11}$$

Table 2
Parameters and variables used in the formulation.

Notations	Definition
$\mathcal{S}$	Set of stations
$\mathcal{I}$	Set of train services
$\mathcal{T}$	Set of discrete timestamp
s	Station index, $s \in \mathcal{S}$
i	Train service index, $i \in \mathcal{I}$
t	Time index, $t \in \mathcal{T}$
$\underline{h}, \overline{h}$	Minimum and maximum headway
ε_{is}	Dwell time of train service i at station s
η_{is}	Running time of train service i from stations s to $s+1$
D_{svt}	Number of passengers arriving at station s at time t who are heading to destination v
κ_{isv}	Minimum percentage of waiting passengers with no reservations served by train service i
C	Train capacity
a_{is}	Arrival time of train service i at station s
d_{is}	Departure time of train service i from station s
y_{ist}	0–1 variable indicating whether time t falls within the headway between train services $i-1$ and i
$\hat{b}_{isv}$	Number of reserved passengers boarding train service i departing from station s towards destination v
w_{isv}	Number of passengers waiting at station s for train service i heading towards destination v
r_{isv}	Number of passengers detained by train service i at station s
$\hat{o}_{is}$	Number of in-vehicle passengers with reservations when train service i departs from station s
o_{is}	Number of in-vehicle passengers without reservations when train service i departs from station s
l_{is}	Number of unreserved passengers alighting from train service i at station s
z_i	Line congestion experienced by train service i during operations
F^e	Total waiting time of passengers
F^c	Total line congestion

Constraints (4) enforce the non-increasing nature of variable **x**. Here, if train service i has departed from station s at time t, $x_{ist} = 0$; otherwise, $x_{ist} = 1$. Constraints (5) ensure that all trains operate and reach the terminal station within the study time domain. Constraints (6) calculate the real-valued departure time of train service i from station s based on the train operating variable **x**. Constraints (7) define the upper and lower bounds of the headway. Constraints (8) and (9) link the arrival and departure times of each train at every station, capturing the dynamics of train movements. Constraints (10) track the 0–1 variable associated with the headway to facilitate the calculation of extra passenger waiting time due to being stranded. Lastly, constraints (11) define the domain of the decision variables.

(ii) **Constraints related to the time-varying reservation slot allocation plan.**

Passengers with reservations can enter the station upon arrival, wait on the platform, and board the next available train. This dynamic process is modeled as follows:

$$\hat{b}_{isv} = \sum_{t \in \mathcal{T}} y_{ist} \varrho_{st} \qquad \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s, \tag{12}$$

$$\sum_{i \in \mathcal{I}} \hat{b}_{isv} = \sum_{t \in \mathcal{T}} \varrho_{svt} \qquad \forall s, v \in \mathcal{S}, v > s, \tag{13}$$

$$0 \le \varrho_{svt} \le D_{svt} \qquad \forall s, v \in \mathcal{S}, v > s, t \in \mathcal{T}, \tag{14}$$

$$\varrho_{svt} \in \mathbb{Z}_{\ge 0} \qquad \forall s, v \in \mathcal{S}, v > s, t \in \mathcal{T}. \tag{15}$$

Constraints (12) ensure that all passengers with reservations arriving during the headway between trains $i-1$ and i are able to board train service i. Constraints (13) guarantee that all passengers with reservations are served within the study time horizon. Constraints (14) define the domain of the decision variable ϱ_{svt}, ensuring that the number of reservation slots does not exceed the total demand. Constraints (15) specify that ϱ_{svt} is an integer variable.

(iii) **Constraints related to the dynamic evolution of passengers without reservations.**

Subject to the passenger flow control strategy, passengers without reservations must queue outside the station upon arrival and wait for an entry permit. Some passengers receive the permit and proceed to the platform to await the next train, while others are stranded and have to wait in the queue outside the station. This dynamic evolution process is formulated as follows:

$$w_{isv} = \begin{cases} \sum_{t \in \mathcal{T}} x_{ist}(D_{svt} - \varrho_{svt}) i = 1 \\ \sum_{t \in \mathcal{T}} x_{ist}(D_{svt} - \varrho_{svt}) - \sum_{j \in \mathcal{I}, j < i} b_{jsv} i \in \mathcal{I} \setminus \{1\} \end{cases} \qquad \forall s, v \in \mathcal{S}, v > s, \tag{16}$$

$$\kappa_{isv} w_{isv} \le b_{isv} \le w_{isv} \qquad \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s, \tag{17}$$

$$r_{isv} = w_{isv} - b_{isv} \qquad \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s, \tag{18}$$

$$\sum_{i \in \mathcal{I}} b_{isv} = \sum_{t \in \mathcal{T}} (D_{svt} - \varrho_{svt}) \qquad \forall s, v \in \mathcal{S}, v > s, \tag{19}$$

$$b_{isv} \in \mathbb{Z}_{\ge 0} \qquad \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s. \tag{20}$$

Constraints (16) calculate the number of passengers waiting at station s for train service i who are traveling to station v. For the first train, the number of waiting passengers equals the number of arriving passengers without reservations. For subsequent trains, this number is determined as the difference between the arriving passengers and those who have already boarded.

Constraints (17) specify that train service i must serve at least κ_{isv} (%) of the passengers waiting at station s with a destination of station v. These constraints ensure service fairness across stations by preventing excessive numbers of passengers from being stranded at upstream stations to reserve capacity for downstream passengers. Constraints (18) compute the number of passengers with a destination of station v who are stranded by train service i at station s. Constraints (19) ensure that all passengers without reservations are served during the study time domain. Constraints (20) define the domain of the decision variable related to the passenger flow control strategy.

(iv) **Constraints of interactions between passengers with and without reservations.** There is a certain interconnection between the passengers with reservations and those without reservations in sharing the capacity resources of the urban rail transit system. Specifically, when a train arrives at a station, both the reserved passengers waiting on the platform and the unreserved passengers who have been permitted to enter the platform are allowed to board the train. However, to ensure operational safety, the total number of passengers in the vehicle cannot exceed the maximum capacity of the train. These interactions can be formulated as follows:

$$o_{is} + \hat{o}_{is} \le C \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{21}$$

$$\hat{o}_{is} = \sum_{m \in \mathcal{S}, m \le s-1} \sum_{v \in \mathcal{S}, v \ge s+1} \hat{b}_{imv} \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{22}$$

$$o_{is} = \begin{cases} \sum\limits_{v \in \mathcal{S}, v > s} b_{isv} & s = 1 \\ o_{i(s-1)} - l_{is} + \sum\limits_{v \in \mathcal{S}, v > s} b_{isv} & s \in \mathcal{S} \setminus \{1, |\mathcal{S}|\} \\ 0 & s = |\mathcal{S}| \end{cases} \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{23}$$

$$l_{is} = \begin{cases} 0 & s = 1 \\ \sum\limits_{m \in \mathcal{S}, m \le s-1} b_{ims} & s \in \mathcal{S} \setminus \{1\} \end{cases} \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{24}$$

$$z_i \ge \sum_{s, v \in \mathcal{S}, v > s} (w_{isv} + \hat{b}_{isv}) \qquad \forall i \in \mathcal{I}. \tag{25}$$

Constraints (21) are hard capacity constraints, ensuring that the number of in-vehicle passengers does not exceed the train's capacity. Constraints (22) calculate the number of passengers with reservations on board when train service i departs station s, which includes passengers with reservations who boarded the train at upstream stations of s and are destined for downstream stations. Constraints (23) compute the number of in-vehicle passengers without reservations. Constraints (24) calculate the number of unreserved passengers disembarking from train service i at station s. Lastly, constraints (25) calculate the maximum number of waiting passengers across stations when train service i arrives.

Based on the aforementioned formulations, we now present the INLP model that integrates reservation strategies into demand management and train scheduling for metro systems, which reads as follows:

$$\begin{aligned} \min \quad & \lambda^e F^e + \lambda^c F^c \\ s.t. \quad & (2) - (25). \end{aligned} \tag{26}$$

4.4. Linearization

Since the train departure 0–1 decision variables $\mathbf{x}$ and $\mathbf{y}$ (related to timetabling), the reservation slot allocation decision variable $\mathbf{q}$, and the stranded passenger variable $\mathbf{r}$ are all variables, constraints (2), (12), and (16) are nonlinear. To enable solving by commercial solvers, these constraints are linearized as follows.

To linearize constraints (2), we first define a binary variable $\theta_{istt'} = y_{ist} y_{ist'}$ for all $i \in \mathcal{I}$, $s \in \mathcal{S}$, and $t, t' \in \mathcal{T}$ with $t' \le t$. The linear form of this variable is expressed as:

$$\begin{cases} \theta_{istt'} \le y_{ist} \\ \theta_{istt'} \le y_{ist'} \\ \theta_{istt'} \ge y_{ist} + y_{ist'} - 1 \\ \theta_{istt'} \in [0, 1] \end{cases} \qquad \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s, t, t' \in \mathcal{T}, t' \le t. \tag{27}$$

In addition, an auxiliary variable $\gamma_{ist} = y_{ist} \sum_{v>s, v\in S} r_{isv}$ is introduced. Its linear form is expressed as:

$$\begin{cases} \gamma_{ist} \le M_{is} y_{ist} \\ \gamma_{ist} \le \sum_{v>s, v\in S} r_{isv} \\ \gamma_{ist} \ge \sum_{v>s, v\in S} r_{isv} - M_{is}(1 - y_{ist}) \\ \gamma_{ist} \in [0, M_{is}] \end{cases} \quad \forall i \in \mathcal{I}, s \in \mathcal{S}, t \in \mathcal{T}, \tag{28}$$

where M_{is} is set as $\sum_{v>s, v\in S} (D_{svt} - \varrho_{svt}), \forall i \in \mathcal{I}, s \in \mathcal{S}$ to give a tight upper bound without cutting the optimal solutions.

Therefore, constraints (2) can be reformulated in their linearized form as:

$$F^e = \Delta \left[\sum_{i\in\mathcal{I}} \sum_{s\in\mathcal{S}} \sum_{t\in\mathcal{T}} \sum_{t'\in\mathcal{T}, t'\le t} \left(\theta_{istt'} \sum_{v\in\mathcal{S}, v>s} D_{svt'} \right) + \sum_{i\in\mathcal{I}} \sum_{s\in\mathcal{S}} \sum_{t\in\mathcal{T}} \gamma_{ist} \right]. \tag{29}$$

By defining an auxiliary variable $\alpha_{isvt} = y_{ist}\varrho_{svt}$ for all $i \in \mathcal{I}, s, v \in \mathcal{S}, t \in \mathcal{T}$, constraints (12) can be linearized as

$$\begin{cases} \hat{b}_{isv} = \sum_{t\in\mathcal{T}} \alpha_{isvt} & \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s, \\ \alpha_{isvt} \le M_{svt} y_{ist} & \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s, t \in \mathcal{T}, \\ \alpha_{isvt} \le \varrho_{svt} & \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s, t \in \mathcal{T}, \\ \alpha_{isvt} \ge \varrho_{svt} - M_{svt}(1 - y_{ist}) & \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s, t \in \mathcal{T}, \\ \alpha_{isvt} \in [0, M_{svt}] & \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s, t \in \mathcal{T}. \end{cases} \tag{30}$$

where $M_{svt} = D_{svt}, \forall s, v \in \mathcal{S}, v > s, t \in \mathcal{T}$.

By introducing an auxiliary variable $\beta_{isvt} = x_{ist}\varrho_{svt}$ for all $i \in \mathcal{I}$, $s, v \in \mathcal{S}$, and $t \in \mathcal{T}$, constraints (16) can be reformulated as follows:

$$w_{isv} = \begin{cases} \sum_{t\in\mathcal{T}} (x_{ist} D_{svt} - \beta_{isvt}) & i = 1 \\ \sum_{t\in\mathcal{T}} (x_{ist} D_{svt} - \beta_{isvt}) - \sum_{j\in\mathcal{I}, j<i} b_{jsv} & i \in \mathcal{I} \setminus \{1\} \end{cases} \quad \forall s, v \in \mathcal{S}, v > s, \tag{31}$$

where the linear form of β_{isvt} can be formulated as

$$\begin{cases} \beta_{isvt} \le M_{svt} x_{ist} \\ \beta_{isvt} \le \varrho_{svt} \\ \beta_{isvt} \ge \varrho_{svt} - M_{svt}(1 - x_{ist}) \\ \beta_{isvt} \in [0, M_{svt}] \end{cases} \quad \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s, t \in \mathcal{T}, \tag{32}$$

where $M_{svt} = D_{svt}, \forall s, v \in \mathcal{S}, v > s, t \in \mathcal{T}$.

Based on the aforementioned reformulations, Formulation (26) can be equivalently transformed into the following linear programming model:

$$\begin{aligned} \min \quad & \lambda^e F^e + \lambda^c F^c \\ s.t. \quad & (3) - (11), (13) - (15), (17) - (25), (27) - (32). \end{aligned} \tag{33}$$

Remark 1. While the proposed framework is developed for a single metro line, it can be extended to a network setting by incorporating transfers and schedule coordination across multiple lines. In the following, we outline the key conceptual ideas for extending the proposed framework to the network level. The detailed modeling and solution methods for such an extension are left for future research. The core idea behind this extension is to preserve the joint optimization of demand management and train scheduling while introducing line-level dimensions and transfer-related constraints and variables.

Specifically, a limited number of reservation slots can still be allocated for each OD pair. Passengers with reservations are allowed to enter the platform upon arrival at their origin station and, if necessary, transfer directly to connecting lines through dedicated corridors. For non-reserved passengers, we assume the operator applies passenger flow control strategies. These passengers must wait for access permission at their origin station according to the passenger control plan. The principles for network-level passenger flow control follow existing literature, such as Lu et al. (2022) and Yuan et al. (2022).

4.5. Complexity analysis

Formulation (26) includes three types of decision variables: train departure binary variables, reservation slot allocation variables, and passenger flow control variables. In contrast, Formulation (33) incorporates an additional fourth type: auxiliary variables introduced for linearization. The size of these two models depends on the number of stations, trains, and timestamps. Table 3 summarizes the corresponding number of decision variables and selected constraints for each formulation.

Notably, transforming the nonlinear model into its linear form significantly increases the number of decision variables and constraints due to the introduction of auxiliary variables and additional constraints. For example, Formulation (26) does not include

Table 3
The numbers of decision variables and constraints in the models.

Decision variables & Constraints	Number	Formulation (26)	Formulation (33)
x_{ist}	$\|I\| \times \|S\| \times \|T\|$	✓	✓
o_{svt}	$\|S\| \times \|S\| \times \|T\|/2$	✓	✓
b_{isv}	$\|I\| \times \|S\| \times \|S\|/2$	✓	✓
$\alpha_{isvt}, \beta_{isvt}$	$\|I\| \times \|S\| \times \|S\| \times \|T\|$		✓
$\theta_{istt'}$	$\|I\| \times \|S\| \times \|T\| \times \|T\|/2$		✓
γ_{ist}	$\|I\| \times \|S\| \times \|T\|$		✓
Constraints (27)	$2 \times \|I\| \times \|S\| \times \|T\| \times \|T\|$		✓
Constraints (28)	$4 \times \|I\| \times \|S\| \times \|T\|$		✓
Constraints (30)	$2 \times \|I\| \times \|S\| \times \|S\| \times \|T\| + \|I\| \times \|S\| \times \|S\|/2$		✓
Constraints (32)	$2 \times \|I\| \times \|S\| \times \|S\| \times \|T\|$		✓

auxiliary variables, whereas Formulation (33) introduces $|I| \times |S| \times |S| \times |T| + |I| \times |S| \times |T| \times |T|/2 + |I| \times |S| \times |T|$ auxiliary variables. Similarly, Formulation (26) excludes constraints (27), (28), (30), and (32), while Formulation (33) includes an additional $2 \times |I| \times |S| \times |S| \times |T| + |I| \times |S| \times |T|/2 + 2 \times |I| \times |S| \times |S| \times |T| + 2 \times |I| \times |S| \times |T| \times |T| + 4 \times |I| \times |S| \times |T|$ constraints.

5. Solution methodologies

Based on the analysis in Section 4.5, Formulation (33) represents a typical integer linear programming (ILP) formulation, which can theoretically be solved directly using commercial solvers such as GUROBI. However, due to the large number of constraints and integer decision variables, solving the problem at a practical scale with GUROBI becomes challenging. The introduction of auxiliary variables and additional constraints during the model linearization process further increases its complexity. If the decision variable related to train scheduling (i.e., $\mathbf{x}$) is treated as an input parameter, the nonlinear constraints (2), (12), and (16) can be reformulated into linear forms. This would decouple the hard constraints between train scheduling and demand management, avoiding the need for linearization techniques and greatly reducing the solving complexity.

To address these challenges, this paper proposes a model decomposition framework that separates the problem into train scheduling and demand management subproblems, armed with tailored valid inequalities. Additionally, a heuristic algorithm combining Adaptive Large Neighborhood Search (ALNS) and GUROBI is developed to efficiently solve large-scale real-world problems. As indicated in Fig. 3, the algorithm iterates between the train scheduling and demand management subproblems, using ALNS to generate feasible timetables and evaluate their quality by solving the demand management subproblem with GUROBI. Within the framework of the ALNS algorithm, two customized destroy operators are designed to adhere to the properties of the proposed model.

5.1. Decomposition framework and valid inequalities

The decomposition framework and valid inequalities are introduced in detail in this section. For clarity, we express Formulation (26) as follows:

$$(O) \quad \min \{ \lambda^e F^e + \lambda^c F^c \mid f(x) \geq 0,\ g(x, \boldsymbol{\rho}, \boldsymbol{b}) \geq 0,\ x \in \chi,\ (\boldsymbol{\rho}, \boldsymbol{b}) \in \Theta \},$$

where (χ, Θ) represents the domain of the decision variables.

To analyze the theoretical properties of Formulation (26), we introduce the following definition associated with problem O.

Definition 1. Let $\hat{\mathbf{x}}$ be an arbitrary vector of binary values satisfying constraints (4), (5), and (11). The reduced problem associated with $\hat{\mathbf{x}}$ is defined as:

$$O(\hat{\mathbf{x}}) \quad \min \{ \lambda^e F^e + \lambda^c F^c \mid g(\hat{\mathbf{x}}, \boldsymbol{\rho}, \boldsymbol{b}) \geq 0,\ (\boldsymbol{\rho}, \boldsymbol{b}) \in \Theta \}.$$

Since the binary variables $\hat{\mathbf{x}}$ are fixed, $O(\hat{\mathbf{x}})$ becomes a linear programming model that assigns reservation slots at each timestamp and allocates passengers to each train service, accounting for the movement of passenger flows in the rail transit system. This reduced problem $O(\hat{\mathbf{x}})$ can be efficiently solved using GUROBI.

However, problem $O(\hat{\mathbf{x}})$ is not always feasible due to constraints (19) and (21). To effectively evaluate the quality of generated timetable $\hat{\mathbf{x}}$, the evaluation function $\phi(\hat{x})$ is finally formulated as follows

$$\phi(\hat{x}) = \lambda^e F^e + \lambda^c F^c + \sum_{i \in I} \sum_{s \in S} M_{is} \max\{o_{is} + \hat{o}_{is} - C, 0\}.$$

As a result, Formulation (26) is decomposed into the following subproblems, i.e., train scheduling (TS) (35) and demand management (DM) (36). Based on these definitions, we first introduce Lemma 1, which establishes the feasibility of the proposed decomposition method with respect to TS (35) and DM (36).

Lemma 1. *DM (36) is always feasible for any $\hat{\mathbf{x}} \in \chi$ obtained by TS (35).*

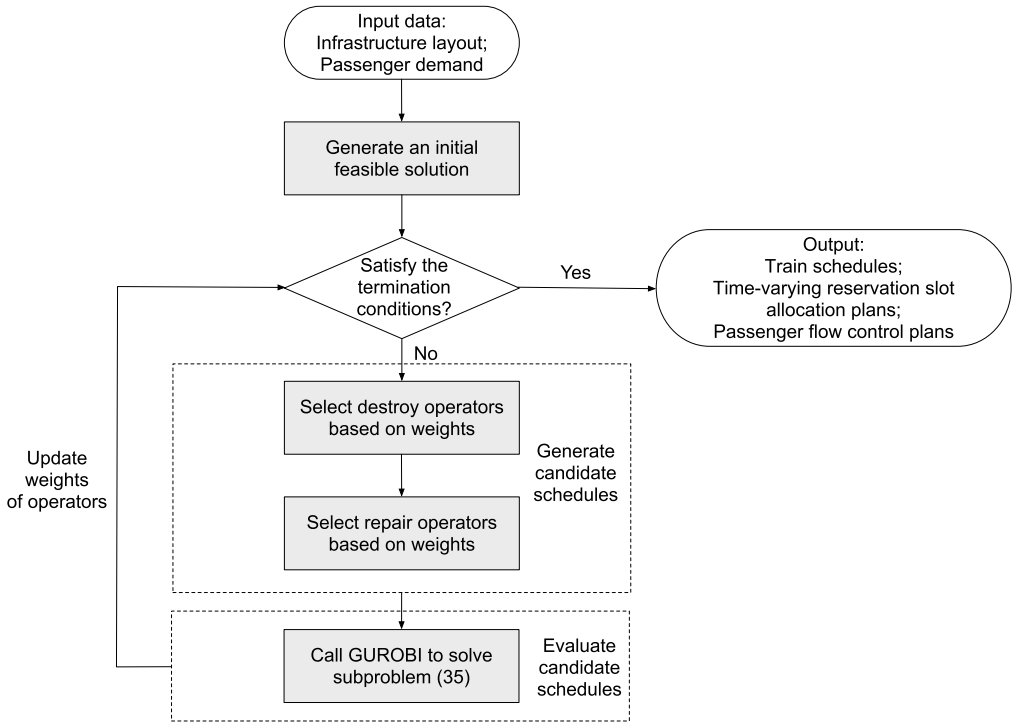

Fig. 3. Framework of the ALNS + GUROBI algorithm.

Proof.

We prove this by explicitly constructing a feasible solution to DM (36) for any given $\hat{\mathbf{x}} \in \chi$.

Consider the following assignments:

- $\varrho_{svt} = 0$ for all $s, v \in \mathcal{S}$, $v > s$, and $t \in \mathcal{T}$ (no reservation slots are allocated).
- $b_{isv} = w_{isv}$ for all $i \in \mathcal{I}$, $s, v \in \mathcal{S}$, $v > s$ (boarding passengers match those waiting at stations).

Under these assignments, we verify that all constraints in DM are satisfied:

(i) Constraints (14) ensure that the number of reservation slots does not exceed demand. Since $\varrho_{svt} = 0$, this condition is trivially satisfied for all $s, v \in \mathcal{S}$ and $t \in \mathcal{T}$.

(ii) Constraints (17) enforce that boarding passengers at station s with destination v do not exceed available capacity. Setting $b_{isv} = w_{isv}$ ensures that passengers boarding at each station directly match those waiting, satisfying this constraint.

(iii) Constraints (19) guarantee that all passengers are served. Since $b_{isv} = w_{isv}$ for all $i \in \mathcal{I}$ and $s, v \in \mathcal{S}$, all passengers are appropriately assigned to trains, fulfilling this requirement.

The above construction demonstrates that all inequalities and hard constraints in DM (36) are satisfied. Therefore, DM is always feasible for any $\hat{\mathbf{x}} \in \chi$ obtained by TS (35).

The proof is complete. □

To enhance computational efficiency and ensure feasibility in train scheduling, a set of valid inequalities can be introduced to tighten the solution space. By leveraging the structure of train departure and running times, we derive a valid inequality that strengthens the scheduling formulation in Lemma 2.

Lemma 2. *To further accelerate the solution, the following valid inequality is added to TS (35):*

$$\sum_{i \in \mathcal{I} \setminus \{1\}} (d_{i1} - d_{(i-1)1}) \leq |\mathcal{T}| - d_{11} - \sum_{s \in \mathcal{S} \setminus \{|\mathcal{S}|\}} \eta_{|\mathcal{I}|s}, \tag{34}$$

where the number of timestamps $|\mathcal{T}|$, the departure time of train 1 from station 1 d_{11}, and the running time between sections $\eta_{|\mathcal{I}|s}$ are all predefined parameters. ***Proof.*** *Recall that constraints (5) require all trains arrive at the terminal station within the study time domain. By combining constraints (5) and (6), it follows that:*

$$d_{|\mathcal{I}||\mathcal{S}|} \leq |\mathcal{T}|.$$

Furthermore, combing constraints (8) and (9), we have

$$\begin{aligned} & d_{|\mathcal{I}||\mathcal{S}|} \leq |\mathcal{T}| \\ \Rightarrow & d_{11} + \sum_{i \in \mathcal{I} \setminus \{1\}} (d_{i1} - d_{(i-1)1}) + \sum_{s \in \mathcal{S} \setminus \{|\mathcal{S}|\}} \eta_{|\mathcal{I}|s} \leq |\mathcal{T}| \end{aligned}$$

$$\Rightarrow \sum_{i\in\mathcal{I}\setminus\{1\}} (d_{i1} - d_{(i-1)1}) \le |\mathcal{T}| - d_{11} - \sum_{s\in\mathcal{S}\setminus\{|\mathcal{S}|\}} \eta_{|\mathcal{I}|s}.$$

Thus, the inequality holds, and the proof is complete. □

The TS subproblem can be expressed as

$$\text{[TS]}\quad \begin{cases} x_{is(t+1)} \le x_{ist} & \forall i\in\mathcal{I}, s\in\mathcal{S}, t\in\mathcal{T}\setminus\{|\mathcal{T}|\}, \\ x_{is|\mathcal{T}|} = 0 & \forall i\in\mathcal{I}, s\in\mathcal{S}, \\ d_{is} = \sum_{t\in\mathcal{T}\setminus\{1\}}[t(x_{is(t-1)} - x_{ist})] & \forall i\in\mathcal{I}, s\in\mathcal{S}, \\ \underline{h} \le d_{is} - d_{(i-1)s} \le \overline{h} & \forall i\in\mathcal{I}\setminus\{1\}, s\in\mathcal{S}, \\ d_{is} = a_{is} + \varepsilon_{is} & \forall i\in\mathcal{I}, s\in\mathcal{S}, \\ a_{is} = d_{i(s-1)} + \eta_{i(s-1)} & \forall i\in\mathcal{I}, s\in\mathcal{S}\setminus\{1\}, \\ y_{ist} = \begin{cases} x_{ist} & i = 1 \\ x_{ist} - x_{(i-1)st} & i\in\mathcal{I}\setminus\{1\} \end{cases} & \forall s\in\mathcal{S}, t\in\mathcal{T}, \\ x_{ist}\in\{0,1\} & \forall i\in\mathcal{I}, s\in\mathcal{S}, t\in\mathcal{T}. \end{cases} \tag{35}$$

The DM subproblem is

$$\text{[DM]}\quad \begin{cases} \min \quad \lambda^e F^e + \lambda^c F^c + \sum_{i\in\mathcal{I}}\sum_{s\in\mathcal{S}} M_{is}\max\{o_{is} + \hat{o}_{is} - C, 0\} & \\ \text{s.t.} \quad F^e = \Delta\left[\sum_{i\in\mathcal{I}}\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}}\left(\hat{y}_{ist}\sum_{t'\in\mathcal{T},t'\le t}\hat{y}_{ist'}\sum_{v\in\mathcal{S},v>s} D_{svt}\right) + \sum_{i\in\mathcal{I}}\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}}\left(\hat{y}_{ist}\sum_{v\in\mathcal{S},v>s} r_{isv}\right)\right], & \\ \quad F^c = \sum_{i\in\mathcal{I}} z_i, & \\ \quad \hat{b}_{isv} = \sum_{t\in\mathcal{T}} \varrho_{st}\hat{y}_{ist} & \forall i\in\mathcal{I}, s, v\in\mathcal{S}, v > s, \\ \quad \sum_{i\in\mathcal{I}}\hat{b}_{isv} = \sum_{t\in\mathcal{T}}\varrho_{svt} & \forall s, v\in\mathcal{S}, v > s, \\ \quad 0 \le \varrho_{svt} \le D_{svt} & \forall s, v\in\mathcal{S}, v > s, t\in\mathcal{T}, \\ \quad w_{isv} = \begin{cases} \sum_{t\in\mathcal{T}} x_{ist}(D_{svt} - \varrho_{svt}) & i = 1 \\ \sum_{t\in\mathcal{T}} x_{ist}(D_{svt} - \varrho_{svt}) - \sum_{j\in\mathcal{I},j<i} b_{jsv} & i\in\mathcal{I}\setminus\{1\} \end{cases} & \forall s, v\in\mathcal{S}, v > s, \\ \quad \kappa_{isv} w_{isv} \le b_{isv} \le w_{isv} & \forall i\in\mathcal{I}, s, v\in\mathcal{S}, v > s, \\ \quad r_{isv} = w_{isv} - b_{isv} & \forall i\in\mathcal{I}, s, v\in\mathcal{S}, v > s, \\ \quad \sum_{i\in\mathcal{I}} b_{isv} = \sum_{t\in\mathcal{T}}(D_{svt} - \varrho_{svt}) & \forall s, v\in\mathcal{S}, v > s, \\ \quad \hat{o}_{is} = \sum_{m\in\mathcal{S},m\le s-1}\sum_{v\in\mathcal{S},v\ge s+1}\hat{b}_{imv} & \forall i\in\mathcal{I}, s\in\mathcal{S}, \\ \quad o_{is} = \begin{cases} \sum_{v\in\mathcal{S},v>s} b_{isv} & s = 1 \\ o_{i(s-1)} - l_{is} + \sum_{v\in\mathcal{S},v>s} b_{isv} & s\in\mathcal{S}\setminus\{1,|\mathcal{S}|\} \\ 0 & s = |\mathcal{S}| \end{cases} & \forall i\in\mathcal{I}, s\in\mathcal{S}, \\ \quad l_{is} = \begin{cases} 0 & s = 1 \\ \sum_{m\in\mathcal{S},m\le s-1} b_{ims} & s\in\mathcal{S}\setminus\{1\} \end{cases} & \forall i\in\mathcal{I}, s\in\mathcal{S}, \\ \quad z_i = \sum_{s,v\in\mathcal{S},v>s}(w_{isv} + \hat{b}_{isv}) & \forall i\in\mathcal{I}, \\ \quad \varrho_{svt}\in\mathbb{Z}_{\ge 0} & \forall s, v\in\mathcal{S}, v > s, t\in\mathcal{T}, \\ \quad b_{isv}\in\mathbb{Z}_{\ge 0} & \forall s, v\in\mathcal{S}, v > s. \end{cases} \tag{36}$$

5.2. Adaptive large neighborhood search

To address the tightly coupled and combinatorial nature of the joint optimization problem of demand management and train scheduling, we incorporate the Adaptive Large Neighborhood Search (ALNS) algorithm into our solution framework. Compared to other metaheuristics or decomposition-based methods, ALNS offers greater flexibility and performs well when handling large-scale problems characterized by complex decision interdependencies and a mix of binary and integer variables (e.g., Yin et al., 2021; Zhang et al., 2024a; Tundulyasaree et al., 2025).

ALNS is an efficient heuristic method, distinguished by its ability to integrate multiple destroy and repair operators within a single search process. Its effectiveness lies in the adaptive optimization of the search strategy, leveraging solution information to

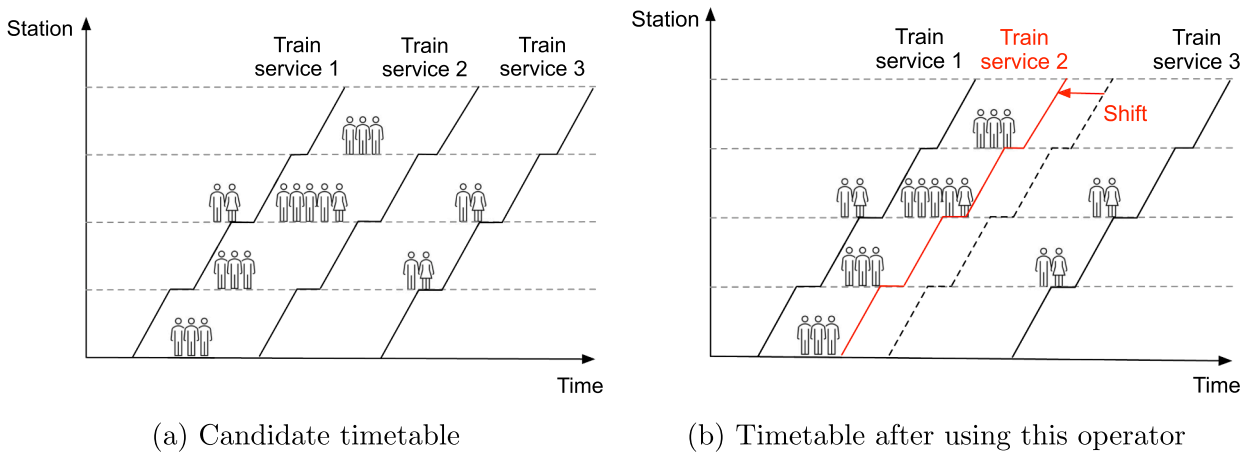

Fig. 4. Illustration of the crowding-based destroy operator.

dynamically select the most suitable operators and enhance search performance. Initially, the algorithm assigns weights to each operator, which are then continuously updated throughout the search based on their historical performance and frequency of use. This section introduces the key components of the ALNS algorithm, including the construction of the initial solution, the design of destroy and repair operators tailored to the characteristics of our model, the adaptive operator selection mechanism, and the termination criteria of the algorithm.

Initial solution generation. The first step in designing an efficient ANLS algorithm is to find a good initial solution as the starting point of the search. Considering that an entire train schedule can be obtained when the headway is determined, we design a record method based on the headway, i.e., $H = \{h_1, h_2, \cdots, h_{|I|}\}$. Thereafter, we generate an initial solution by setting a constant headway h_c satisfying constraints (7), which is commonly used by the metro managers in practice to build train schedules with constant headways (Yin et al., 2021). Thereafter, we check the feasibility of this initial solution with respect to the timetabling constraints (4) - (5). If the solution is infeasible, we first check whether valid inequalities (34) are met or not. If yes, we randomly select a certain number of headways and update them to be $h_c + \varsigma$; otherwise, we update them as $h_c - \varsigma$, where ς is a tiny integer value. We repeat this procedure until a train schedule that satisfies constraints (4) - (5) and (34) is found.

Destroy operators. The ALNS algorithm uses both destroy and repair operators to generate candidate solutions. Therefore, how to design the destroy and repair operators is crucial to improve the efficiency of this algorithm. In this study, we propose four destroy operators. The first two leverage randomness to help escape local optima, while the latter two are designed based on the properties of the model to improve the effectiveness of the solution.

(i) ***Headway adjustment operator.***

This operator randomly selects train service i and increases or decreases the headway between this train and train $i+1$.

(ii) ***Headway swap operator.***

This operator randomly selects two headways between i and $i+1$, between j and $j+1$. Then, their corresponding values are exchanged.

(iii) ***Crowding-based operator.***

This operator identifies the train with the largest number of waiting passengers. Specifically, the values of $\sum_{v \in S, v \geq s+1} w_{isv}$ for all $i \in \mathcal{I}, s \in S$ obtained in the previous iteration are ranked, and the probability of selecting a train is assigned based on this ranking. The key idea of this destroy operator is to ensure that the higher the value of $\sum_{v \in S, v \geq s+1} w_{isv}$, the greater the probability that the corresponding train service i will be selected. Furthermore, train service i is selected according to a roulette rule and its headway with train $i+1$ is decreased, i.e., we have $h_{i+1} \leftarrow h_{i+1} - \delta$, where $\delta \in \mathbb{Z}_{>0}$ is a parameter.

(iv) *Low-demand adjustment operator.*

This operator identifies the train with the smallest number of waiting passengers. That is, based on the ranking results in the previous destroy operator, this operator aims to ensure that the smaller the value of $\sum_{v \in S, v \geq s} w_{isv}(w)$, the greater the probability that the corresponding train service i will be selected. The headway between the selected train service i and its follow-up train $i+1$ is updated as $h_{i+1} \leftarrow h_{i+1} + \delta$.

Example 3. An illustration of the crowding-based destroy operator is shown in Fig. 4. In the candidate timetable illustrated in Fig. 4(a), three train services are operated with uneven passenger distribution across trains.

By applying the crowding-based destroy operator, train service 2 is identified as having the highest number of waiting passengers. To alleviate congestion, this destroy operator reduces the headway between train services 1 and 2 by shifting train service 2 forward in time as shown in Fig. 4(b). This adjustment redistributes passengers more evenly across services by enabling some passengers to board an earlier train, thus reducing the waiting time of passengers.

Table 4
Running time on sections on Beijing metro Batong line.

Sections (Abbreviations)	Running time (min)
Tuqiao (TQ) station → Linheli (LHL) station	2
Linheli (LHL) station → Liyuan (LY) station	2
Liyuan (LY) station → Jiukeshu (JKS) station	2
Jiukeshu (JKS) station → Guoyuan (GY) station	2
Guoyuan (GY) station → Tongzhou North (TZN) station	2
Tongzhou North (TZN) station → Baliqiao (BLQ) station	2
Baliqiao (BLQ) station → Guanzhuang (GZ) station	3
Guanzhuang (GZ) station → Shuangqiao (SQ) station	3
Shuangqiao (SQ) station → Chuanmei (CM) station	3
Chuanmei (CM) station → Gaobei (GB) station	3
Gaobei (GB) station → Sihui East (SHE) station	2
Sihui East (SHE) station → Sihui (SH) station	3

Repair operators. At each iteration, if constraints (34) are violated, the repair operator randomly selects a headway and reduces its value to ensure that this family of constraints holds.

Adaptive searching strategy. In each iteration, a set of destroy and repair operators are selected to generate new solutions. Here, we design an adaptive search strategy to update the weights of each operator and select the most effective operators. To do so, we follow the strategies described in Yin et al. (2021).

The scores and weights of operators, which are automatically adjusted between 0 and 1, ensuring that the search direction always moves towards the improved solutions. Initially, the weight and score of each operator are set as 1 and 0, respectively. In each iteration, the weights and scores of each operator are dynamically updated. Specifically, we introduce ϵ_k^d and ζ_k^d as the score and weight of the k-th destroy operator. As the process evolves and depending on the operator performance, the weights are updated by:

$$\zeta_k^d = (1-\lambda)\zeta_k^d + \lambda \frac{\epsilon_k^d}{\sum_{k=1}^{4} \epsilon_k^d},$$

where $\lambda \in [0, 1]$ acts as a scaling factor to control how sensitive the weights are to the changes in the operators' performance.

Simulated annealing and termination criteria. At the end of each iteration, the initial solution for the subsequent iteration is selected based on the simulated annealing principle. This rule allows for controlled exploration of the solution space by occasionally accepting worse solutions to escape local optima, thus enhancing the algorithm's ability to converge to a global optimum. The probability of accepting a worse solution decreases as the algorithm progresses, governed by a cooling schedule that gradually reduces the temperature parameter.

The termination criteria of the algorithm are designed to ensure both computational efficiency and solution quality. Specifically, the algorithm terminates under the following conditions:

(i) The maximum number of iterations, $N^{\max}$, is reached, indicating the search has been sufficiently explored.

(ii) The best solution remains unchanged for M consecutive iterations, suggesting the search has likely converged to an optimal or near-optimal solution.

6. Numerical experiments

In this section, we present computational results from a real-world case study based on the Beijing metro Batong line, illustrating the potential applications and benefits of the proposed models and algorithm. The algorithm is implemented in Java and utilizes GUROBI version 9.5.1. All experiments are performed on a personal computer equipped with a 12th Gen Intel(R) Core(TM) i7-12700H CPU and 64 GB of RAM.

6.1. Numerical design

We now describe the dataset used to evaluate our methodology. The demand management and train scheduling problem requires input on the infrastructure layout, running times on sections, dwell times at stations, and time-varying passenger demand. To test our proposed approaches, we utilize real-world operational data from the Beijing metro Batong line. This dataset is used to generate several instances with varying numbers of timestamps and train services.

The Beijing metro Batong line is an important urban rail line connecting residential areas to workplaces with 13 stations, as illustrated in Fig. 5. In 2018, passenger flow control strategies were implemented at 9 stations during peak hours on weekdays to ensure operational safety and maintain the normal operation of trains. Table 4 provides the detailed running times between stations. Additionally, the dwell time at each station is set at 1 minute, with minimum and maximum headway of 2 minutes and 6 minutes, respectively. The train capacity is 1800 passengers. Passenger flow data from a weekday in 2016 is used as the input parameter and discretized with a 1-minute time granularity.

To evaluate the effectiveness of the proposed method, five instances with varying problem scales are constructed, with detailed parameters provided in Table 5. In these instances, the number of timestamps is gradually increased from 60 to 180, and the number

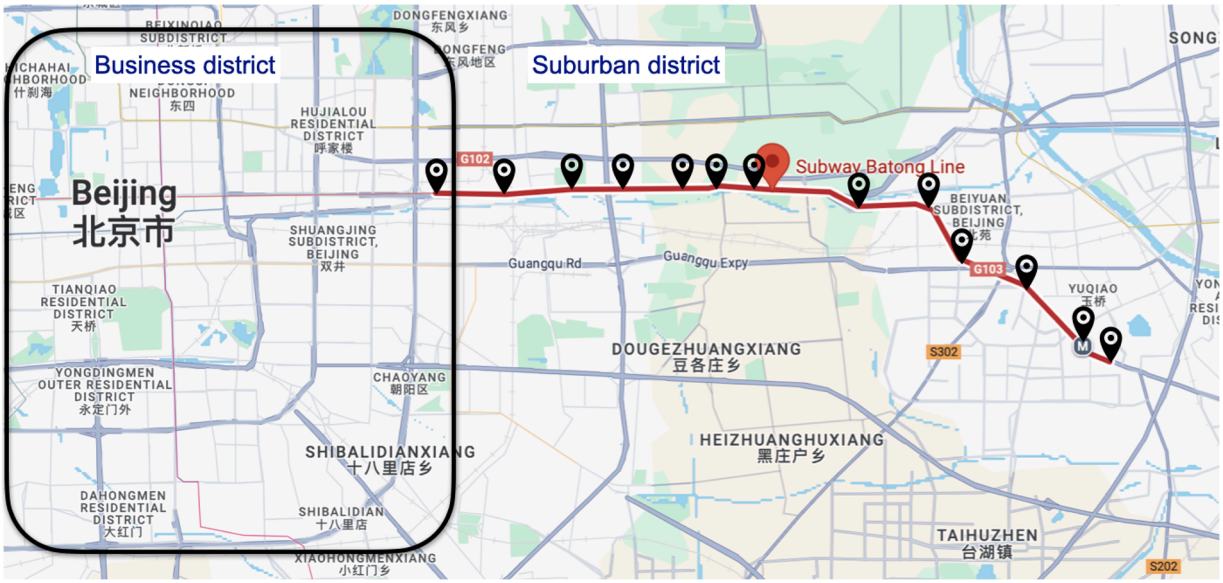

Fig. 5. Beijing metro Batong line. Source: Google map.

Table 5
Characteristics of the instances.

Instance index	Study time horizon	Timestamps (#)	Train services (#)
Q1	6:30 AM - 7:30 AM	60	6
Q2	6:30 AM - 8:00 AM	90	12
Q3	6:30 AM - 8:30 AM	120	23
Q4	6:30 AM - 9:00 AM	150	38
Q5	6:30 AM - 9:30 AM	180	48

of train services is scaled up from 6 to 48. Subsequently, seven sets of experiments are conducted to address the following questions and provide managerial insights for decision-makers:

(i) How do the computational efficiency and solution quality of the proposed algorithm compare to those of the state-of-the-art solver?

(ii) How do the computational efficiency and solution quality of the proposed algorithm compare to those of other metaheuristics, particularly Large Neighborhood Search?

(iii) To what extent does the joint optimization outperform the step-by-step optimization?

(iv) What is the impact of service fairness on operational efficiency?

(v) What are the benefits of the proposed line-wide and dynamic reservation slot allocation strategy compared to state-of-the-art benchmarks?

(vi) How robust is the proposed approach to deviations in passenger arrival times?

(vii) How do different settings of the weighting coefficients affect operational efficiency and safety?

6.2. Best settings of parameters for the proposed algorithm

We first investigate the impact of two key algorithmic parameters on the performance of the proposed algorithm (denoted as ALNS + GUROBI): the maximum number of iterations ($N^{\max}$) and the number of consecutive iterations without improvement before termination (M). Two sets of experiments are conducted using instance Q2 under various combinations of these parameters. The first set aims to evaluate the impact of $N^{\max}$ on solution quality by varying this parameter while keeping M fixed at 100. The second set of experiments seeks to determine the best value for M, given a fixed $N^{\max}$. For each parameter setting, the algorithm is executed five times.

Table 6 reports the performance of the ALNS + GUROBI algorithm under different settings of the maximum number of iterations. We can observe that increasing $N^{\max}$ from 500 to 1000 does not lead to consistent improvement in objective values. In all runs with $N^{\max} = 500$ and $N^{\max} = 1{,}000$, the algorithm terminates far before reaching the maximum iteration limit due to the other stopping criterion M, with the number of iterations ranging between 108 and 197. This observation indicates that setting a larger $N^{\max}$ has a limited practical effect, as the other termination condition M is triggered before the maximum iteration limitation is reached.

Table 6
Performance comparison of ALNS + GUROBI among different settings of the maximum number of iterations.

$N^{\max}$	Objective value	Iterations	Computation time (s)
	80,710.00	113.00	1,082.04
	77,260.00	197.00	1,885.75
500	83,783.00	133.00	1,272.44
	83,337.00	125.00	1,193.74
	80,399.00	146.00	1,397.82
	82,070.00	115.00	1,100.23
	82,015.00	108.00	1,003.26
1,000	80,979.00	135.00	1,319.28
	81,290.00	129.00	1,221.72
	79,894.00	131.00	1,243.59

Moreover, the best solution is achieved in one of the runs with $N^{\max} = 500$, suggesting that higher iteration limits do not necessarily yield better performance. Therefore, setting $N^{\max} = 500$ is a sufficient choice for ensuring solution quality without unnecessary computational overhead.

In the second set of experiments, we fix $N^{\max}$ at 500 and vary M from 10 to 100. Fig. 6 illustrates the objective values obtained under different settings of the termination threshold M. It can be observed that when M is set to relatively small values such as 10 or 30, the objective values are noticeably higher. This is primarily because the algorithm tends to terminate prematurely, before reaching high-quality solutions. As M increases to 50, both the quality and stability of the solutions improve significantly. Further increasing M to 100 yields the similar solution quality with slightly lower variance, but this comes at the cost of increased computation time. Therefore, $M = 50$ provides a good trade-off between solution quality and computational efficiency, and is adopted as the default setting in the subsequent experiments.

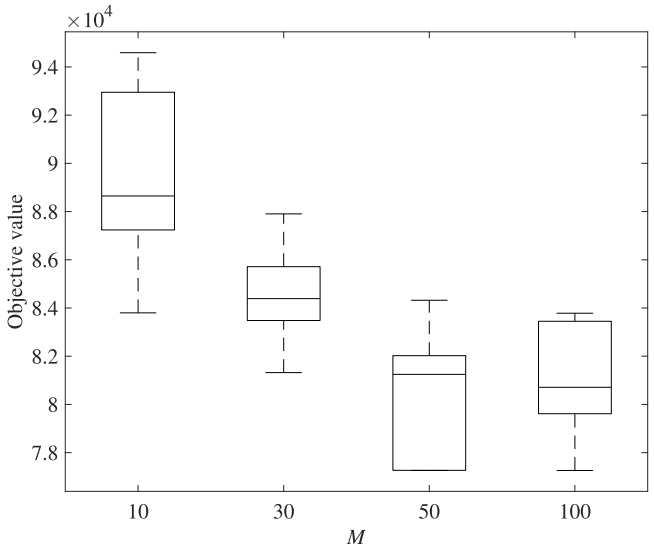

Fig. 6. Results of the algorithm among various settings concerning the termination criteria M.

6.3. Benefits of the proposed algorithm

To assess the computational efficiency and solution quality of the ALNS + GUROBI relative to GUROBI and the algorithm combining Large Neighborhood Search and GUROBI (denoted as LNS + GUROBI), a set of experiments is conducted using the five instances detailed in Table 5. GUROBI is used to solve Formulation (33) and serves as the benchmark. The key difference between ALNS and LNS lies in the adaptivity mechanism: LNS applies a fixed set of destroy and repair operators throughout the search process, whereas ALNS dynamically adjusts the selection probabilities of different operators based on their historical performance. The detailed procedure of LNS is provided in Algorithm 1 in Appendix A. In these experiments, following the findings obtained in Section 6.2, both ALNS and LNS use a maximum of $N^{\max} = 500$ iterations. The search process terminates if the best-found solution remains unchanged for 50 consecutive iterations, i.e., $M = 50$. The initial weights of all destroy operators are set to 1. For GUROBI, the maximum solution time is limited to 7,200 seconds. In addition, the minimum service percentage for passengers without reservations

Table 7
Performance comparison of GUROBI, ALNS + GUROBI, and LNS + GUROBI.

Instance	Solution method	Objective value	*Gap* (%)	*Dev* (%)	Computation time (s)
Q1	GUROBI	22,157.00	0	–	50.05
	ALNS + GUROBI	22,157.00	–	0	6.29
	LNS + GUROBI	22,157.00	–	0	10.12
Q2	GUROBI	75,420.00	0	–	2,770.41
	ALNS + GUROBI	77,260.00	–	2.44	650.57
	LNS + GUROBI	91,686.00	–	21.57	1,039.78
Q3	GUROBI	–	–	–	7,200.00
	ALNS + GUROBI	173,924.00	–	–	542.36
	LNS + GUROBI	220,758.00	–	–	463.56
Q4	GUROBI	–	–	–	7,200.00
	ALNS + GUROBI	238,863.00	–	–	1281.83
	LNS + GUROBI	257,612.00	–	–	1169.24
Q5	GUROBI	–	–	–	7,200.00
	ALNS + GUROBI	316,736.00	–	–	3,438.55
	LNS + GUROBI	351,490.00	–	–	3,796.43

($\kappa_{isv},\ \forall i \in \mathcal{I},\ s, v \in \mathcal{S},\ v > s$) is set to 20%. The weighting factors in the objective function are assigned values of 1 and 10, respectively, to balance the order-of-magnitude differences between the efficiency and congestion objectives.

Table 7 provides a detailed performance comparison of GUROBI, ALNS + GUROBI, and LNS + GUROBI for the investigated instances. We report the objective value and the computation time. We also report the optimality gap obtained by GUROBI, which is denoted as *Gap* (%). The fifth column reports the gap between ALNS + GUROBI (or LNS + GUROBI) and GUROBI, which is calculated by using the following formula:

$$Dev = \frac{\text{Objective value of ALNS + GUROBI} - \text{Objective value of GUROBI}}{\text{Objective value of GUROBI}} \times 100\ (\%).$$

As shown in the results in Table 7, the performance of GUROBI is far from optimal for instances with a long time horizon, such as Q3 with 120 timestamps, Q4 with 150 timestamps, and Q5 with 180 timestamps. Another observation is that GUROBI does not perform as well as the ALNS + GUROBI algorithm, even for the smallest-scale instance. The ALNS + GUROBI algorithm not only reduces computation time by approximately 87.43 % and 76.52 % for the smaller cases Q1 and Q2, but also achieves high-quality solutions, with a *Dev* of 2.44 % in Q2. The main reason for the limitations of GUROBI is that it has to solve a linear model with a large number of big-M constraints, which are required to linearize the nonlinear constraints. Additionally, as the size of the cases increases, the computation time of the algorithm also increases. However, the ALNS + GUROBI algorithm consistently finds approximate optimal solutions within 3,600 seconds, even for the larger cases (e.g., from Q2 to Q5). From these findings, we conclude that the ALNS + GUROBI algorithm is significantly more scalable and efficient for larger and more complex instances.

We further compare the ALNS + GUROBI and LNS + GUROBI algorithms to assess the effectiveness of the adaptive mechanism embedded in ALNS. For Q2, ALNS + GUROBI achieves an objective value of 77,260 with a 2.44 % deviation from the optimal solution, while LNS + GUROBI yields a considerably worse objective value of 91,686, corresponding to a 21.57 % deviation. A similar result is observed in Q3. The ALNS + GUROBI method obtains an objective value of 173,924, notably outperforming the 220,758 achieved by LNS + GUROBI. The main reason is the lack of adaptivity in LNS, which applies a fixed set of destroy and repair operators throughout the search process. Without the ability to adjust operator weights based on historical performance, LNS is more likely to get trapped in local optima and struggle to escape them. In contrast, ALNS dynamically adjusts the selection probabilities of different operators, enabling more effective exploration of the solution space. From these findings, we can conclude that ALNS consistently outperforms LNS in terms of solution quality, particularly for medium- and large-scale instances.

Furthermore, Fig. 7 illustrates the convergence trends of the objective values for the ALNS + GUROBI and LNS + GUROBI algorithms on instances Q2 and Q3. The horizontal axis represents the number of iterations, while the vertical axis shows the corresponding objective value. It can be observed that, for both instances, the ALNS + GUROBI algorithm demonstrates fast convergence. In instance Q2, the objective value rapidly decreases to 77,260 within the first 16 iterations, closely approaching the optimal objective value of 75,420. In instance Q3, a sharp decline is observed within the first 25 iterations, after which the solution stabilizes at a high-quality level. These results indicate that the ALNS + GUROBI algorithm proposed in this study exhibits both fast convergence and efficient problem-solving capabilities, making it well-suited for practical operational settings. In contrast, LNS + GUROBI exhibits slower convergence and a more limited ability to escape local optima. The objective values obtained by LNS + GUROBI are worse than those achieved by ALNS + GUROBI across both instances. These findings highlight the advantage of the adaptive operator selection mechanism in ALNS, which enhances the search process and guides it toward more promising regions of the solution space.

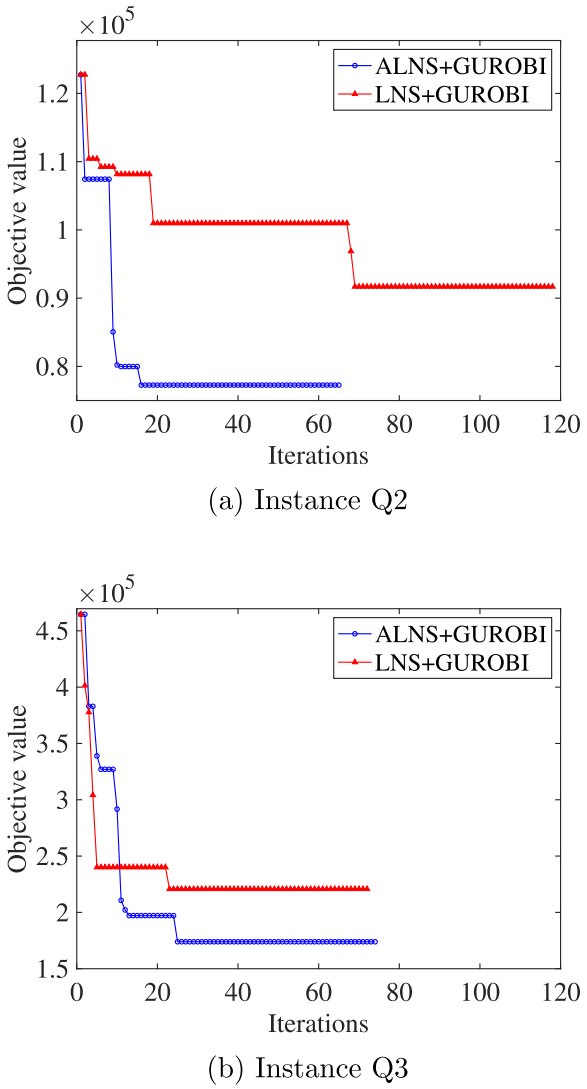

Fig. 7. Convergence tendency of the objective values for the proposed algorithm.

6.4. Benefits of the joint optimization approach of demand management and train scheduling

We then investigate the benefits of the joint optimization method for demand management and train scheduling. This set of experiments is based on instance Q3, still with the minimum service ratio set to 20 %. The proposed joint optimization approach is solved using the ALNS + GUROBI algorithm. The maximum number of iterations is set to 500, and computation stops if the objective value remains unchanged for 50 consecutive iterations. To provide a basis for comparison, we propose a step-by-step optimization framework that serves as the benchmark. This step-by-step strategy offers a computationally simpler decision framework and aligns with how metro operators currently plan and implement timetabling and demand management strategies. In practice, a commonly adopted decision-making process is that the timetable is determined first, and other plans, such as reservation slot allocations and passenger flow control strategies, are subsequently developed based on this fixed schedule. We adopt this sequential planning strategy as the benchmark to evaluate the effectiveness of the proposed joint optimization approach in improving system-level outcomes, including passenger waiting times and line congestion.

Specifically, in the first step, Formulation (37) is solved using GUROBI to determine the optimal train schedule. Formulation (37) incorporates train scheduling constraints and simulation constraints on passenger dynamics to ensure that all passengers are satisfied within the study time horizon. In the second step, the optimal train schedule obtained from Formulation (37) is used as input for subproblem (36), which is also solved using GUROBI. This step determines the optimal reservation slot allocation and passenger flow

Table 8
Comparison of results for step-by-step and joint optimization methods.

Optimization methods	Objective value	Waiting time F^e (min)	Line congestion F^c (persons)
Step-by-step optimization	185,405.00	118,595.00	6,681.00
Joint optimization	173,924.00	107,534.00	6,639.00
Dev (%)	−6.19	−9.33	−0.63

Table 9
Comparison of results under various minimum service ratios.

Minimum service ratio (%)	Waiting time F^e (min)	Line congestion F^c (persons)	Optimized service ratio below 10 % (20 %)(#)	Optimized reservation ratio (%)
0	103,887.00	6,605.00	22 (23)	25.65
10	103,898.00	6,635.00	0 (59)	77.24
20	107,534.00	6,639.00	0 (0)	84.83

control plans.

$$
\begin{cases}
\min\limits_{x} \quad \lambda^e F^e + \lambda^c F^c \\
\text{s.t.} \quad \sum\limits_{v\in\mathcal{S},v>s} b_{isv} = \min\{\hat{C}_{is}, \sum\limits_{v\in\mathcal{S},v>s} w_{isv}\}\ \forall i \in \mathcal{I}, s \in \mathcal{S}, \\
\quad \hat{C}_{is} = \begin{cases} C & s = 1 \\ C - o_{i(s-1)} + l_{is} & s \in \mathcal{S}\setminus\{1\} \end{cases} \forall i \in \mathcal{I}, \\
\quad w_{isv} = \begin{cases} \sum\limits_{t\in\mathcal{T}} x_{ist} D_{svt} & i = 1 \\ \sum\limits_{t\in\mathcal{T}} x_{ist} D_{svt} - \sum\limits_{j\in\mathcal{I},j<i} b_{jsv} & i \in \mathcal{I}\setminus\{1\} \end{cases} \forall s, v \in \mathcal{S}, v > s, \\
\quad \sum\limits_{i\in\mathcal{I}} b_{isv} = \sum\limits_{t\in\mathcal{T}} D_{svt}\ \forall s, v \in \mathcal{S}, v > s, \\
\quad o_{is} \le C\ \forall i \in \mathcal{I}, s \in \mathcal{S}, \\
\quad z_i \ge \sum\limits_{v\in\mathcal{S},v>s} w_{isv}\ \forall i \in \mathcal{I}, \\
\quad (4) - (11), (18), (23) - (24), (25), (30) - (29).
\end{cases} \tag{37}
$$

Table 8 presents a detailed comparison of the results, where *Dev* represents the relative difference between the results of the joint optimization method and those of the step-by-step optimization approach. The results show that the total passenger waiting time obtained from the joint optimization model is reduced by 9.33 % compared to the step-by-step optimization, while line congestion is decreased by 0.63 %. These findings highlight the benefits of the joint optimization approach, which leverages mutual feedback between the train scheduling and demand management subproblems to produce higher-quality solutions. In contrast, the step-by-step optimization approach lacks this feedback mechanism, leading to less effective operational plans.

6.5. The impact of service fairness on operational efficiency

Service fairness is typically achieved at the cost of reduced operational efficiency. In this section, we explore the impact of operational efficiency on the line congestion under different minimum service ratios, based on the instance Q3. Specifically, three cases are evaluated where the minimum service ratio ($\kappa_{isv}\ \forall i \in \mathcal{I}, s, v \in \mathcal{S}, v > s$) is set to 0 %, 10 %, and 20 %, respectively. The ALNS + GUROBI algorithm is employed to solve these instances with algorithmic parameters as described in Section 6.4.

Table 9 presents the optimization results for operational efficiency, line congestion, service ratio of passengers without reservations, and reservation ratio under various minimum service ratios. The service ratio is calculated as the number of unreserved passengers boarding a train divided by the total number of unreserved passengers waiting for this train. If no passengers are waiting, the value is excluded from the results. The fourth column reports the number of optimized trains with service ratios below 10 % or 20 % at each station, while the fifth column shows the reservation ratio (as a percentage), calculated by dividing the number of optimized reservation slots by the total number of passengers.

From the results in Table 9, we observe that increasing the minimum service ratio from 0 % to 10 % leads to a slight increase in passenger waiting time, from 103,887 to 103,898 (approximately 0.01 %), and an increase in line congestion from 6605 to 6635 (about 0.45 %). However, the number of service ratios below 10 % decreases significantly from 22 to zero. Further increasing the minimum service ratio to 20 % results in a 3.51 % increase in passenger waiting time and a 0.51 % increase in line congestion compared to a 0 % minimum service ratio. Additionally, the number of instances where a train transports passengers without reservations at a station with service ratios below 10 % or 20 % reduces to zero. These findings indicate that increasing the minimum service ratio effectively mitigates the extreme unfairness between reserved passengers (who are always served at 100 %) and unreserved passengers, ensuring a more balanced allocation of service. In addition, the results in Table 9 indicate that as the minimum service ratio increases, the ratio of reservations also rises. This result can be attributed to the minimum service ratio constraints, i.e., constraints (17). Specifically, since the minimum service ratio is calculated as the number of unreserved passengers boarding the train divided by the number of

waiting passengers, reducing the number of waiting passengers without reservations is necessary to satisfy this set of constraints (e.g., 10 % or 20 %).

Fig. 8 illustrates the reservation slot allocation plan for each station over time. It can be seen that the allocation of reservation slots under different minimum service ratios is highly dynamic. This highlights that time-varying passenger demand can be effectively managed by dynamically adjusting the number of reservation slots. Such adjustments enable better resource allocation and operational efficiency, even under varying operational constraints.

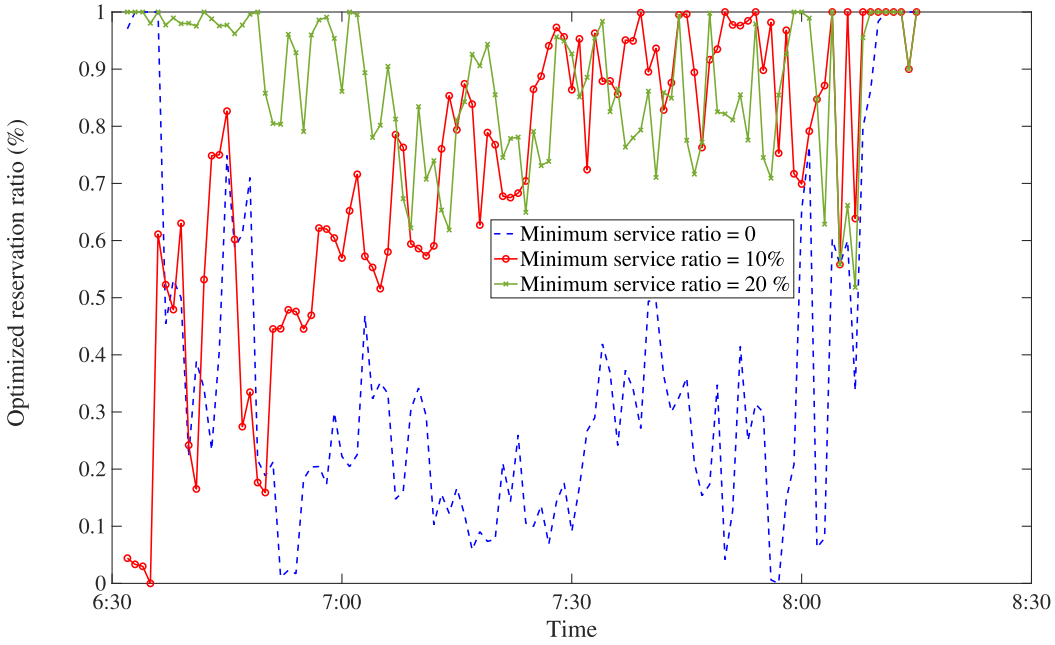

Fig. 8. The optimized reservation ratios under various minimum service ratio settings.

6.6. *Benefits of the proposed reservation slot allocation*

To further assess the benefits of our proposed line-wide and dynamic slot allocation strategy, we incorporate two benchmark strategies inspired by Tang et al. (2024) and Yang et al. (2025), respectively.

- **Single-station reservation allocation:** Based on Tang et al. (2024), this strategy applies the reservation mechanism only at the first station, assuming a fixed timetable with evenly spaced headways. Reserved passengers board directly, while non-reserved passengers follow a straightforward passenger flow control rule based on the remaining train capacity. This strategy is the real-life operational practice adopted in the pilot reservation program in the Beijing metro. The detailed formulation for this benchmark is presented in model (B.1) in Appendix B.
- **Fixed-slot reservation allocation:** Inspired by Yang et al. (2025), this strategy assumes a fixed number of reservation slots and jointly optimizes the train timetable and passenger flow control. The formulation is similar to our main model, except that the variable ϱ_{st} is treated as an input parameter, representing the number of reservation slots at station s and time t.

We construct this set of experiments based on instance Q1, where the number of fixed reservation slots in the *Fixed-slot reservation allocation* strategy is set to 50 % of the passenger demand at each station and time. Table 10 presents the results, reporting the total waiting time, line congestion, the relative deviation in waiting time achieved by our approach compared to the two benchmarks (denoted as Dev^e), and the relative deviation in line congestion (denoted as Dev^c). We observe that, compared to the *Single-station reservation allocation* strategy, our method improves operational efficiency by 10.77 % and reduces line congestion by 1.05 %. In contrast, the *Fixed-slot reservation allocation* strategy fails to yield a feasible solution in this instance. The primary reason is that when the pre-assigned number of reservation slots is too large, not all passengers with reservations can board the train due to limited capacity. This issue has also been acknowledged in Yang et al. (2025), where the authors propose a trip-shifting strategy to alleviate the resulting infeasibility. In our proposed model, this challenge is addressed by treating the reservation slot allocation as a decision variable rather than a fixed input. Compared to the *Fixed-slot reservation allocation* strategy, a key advantage of our approach is that it can provide feasible and efficient reservation slot allocation plans for metro operators regardless of how parameters are set in practice.

Table 10
Performance comparison of various reservation allocation strategies.

Strategy	Waiting time F^e (min)	Dev^e (%)	Line congestion F^c (persons)	Dev^c (%)
Single-station reservation allocation	12,457	–	1046	–
Fixed-slot reservation allocation	Infeasible	–	Infeasible	–
Our approach	11,115	−10.77	1035	−1.05

Table 11
Robustness of optimized and benchmark timetables under different ratios of arrival time perturbation.

Perturbation ratio ω (%)	Solution	Objective value	Dev (%)	Waiting time F^e (min)	Line congestion F^c (persons)
3	Benchmark timetable	77,481.00	–	45,581.00	3,190.00
	Optimized timetable	76,788.00	−0.89	45,008.00	3,178.00
5	Benchmark timetable	77,542.00	–	45,532.00	3,201.00
	Optimized timetable	76,808.00	−0.95	45,018.00	3,179.00
10	Benchmark timetable	77,396.00	–	45,446.00	3,195.00
	Optimized timetable	76,695.00	−0.91	44,915.00	3,178.00

6.7. Robustness analysis under arrival time perturbations

To evaluate the robustness of the proposed integrated optimization method, we consider a realistic setting in which passenger arrival times deviate slightly from the scheduled ones. Specifically, we introduce ± 1-minute perturbations to the original arrival profiles by randomly shifting ω% of the passengers at each station one minute earlier or later, while keeping the total demand unchanged. Here, ω denotes the perturbation ratio, representing the proportion of passengers subject to random arrival shifts. For each tested case, we fix the optimized timetables obtained by our approach, as well as the benchmark timetable derived from the step-by-step optimization method introduced in Section 6.4. We then re-evaluate the performance of these solutions under the perturbed demand profiles. Instance Q2 is used as the test case for this set of experiments.

Table 11 presents the results, reporting the objective values, waiting time, and line congestion under different optimized solutions, as well as the relative deviation of our approach compared to the benchmark step-by-step optimization method (denoted as Dev). As shown in Table 11, the optimized timetable consistently outperforms the benchmark timetable across all perturbation levels in terms of objective value, waiting time, and line congestion. This finding indicates that the proposed integrated optimization method is not only effective under ideal demand assumptions but also robust to minor and uncertain deviations in passenger arrival times. Besides, fixing the timetable while regenerating the reservation slot allocation and passenger flow control strategies under perturbed demand ensures operational feasibility.

6.8. Sensitivity analysis of key parameters

In multi-objective optimization, the weighting coefficients assigned to each objective influence the solution. To examine this effect, we conduct a sensitivity analysis on the weights λ^e and λ^c using instance Q2 by varying λ^e and λ^c within the range [1, 20], which balances service quality (F^e) and operational safety (F^c). Fig. 9 illustrates the trade-off between minimizing total passenger waiting time and reducing line congestion. When λ^e is large and λ^c is small, the optimization emphasizes efficiency, achieving the lowest waiting time but the highest congestion. In contrast, a higher λ^c prioritizes congestion mitigation at the cost of increased waiting time. From these findings, we can conclude that appropriate tuning of the weights is essential to strike a balance between service quality and operational safety in real-life operations.

In addition, Fig. 10 shows the number of passengers with reservations boarding the second train under two different settings of weighting coefficients. It can be observed that the distribution of boarding passengers with reservations varies between these two different settings. When operational safety is prioritized ($\lambda^e : \lambda^c = 1 : 10$), relatively fewer boardings occur at upstream stations such as stations GY and BLQ, while more reservation slots are allocated to downstream stations like GZ, SQ, and CM. In contrast, when service quality is prioritized ($\lambda^e : \lambda^c = 20 : 1$), more passengers with reservations are allowed to board at stations GY and BLQ, resulting in higher utilization at upstream stations but also potential crowding risks downstream (e.g., station GB). These findings highlight that different settings of weighting coefficients produce distinct boarding patterns, and thus the choice of weighting coefficients should align with the real-world operations.

6.9. Managerial insights

Based on the results, this study offers the following managerial insights for metro operators. First, when determining the reservation slot allocation plan, it is essential to adopt joint optimization methods that integrate demand management with train scheduling. This approach ensures an accurate alignment of capacity and demand while upholding service fairness between passengers with and without reservations. By implementing this strategy, metro operators can provide higher-quality, more efficient, and cost-effective travel services that better meet passenger needs.

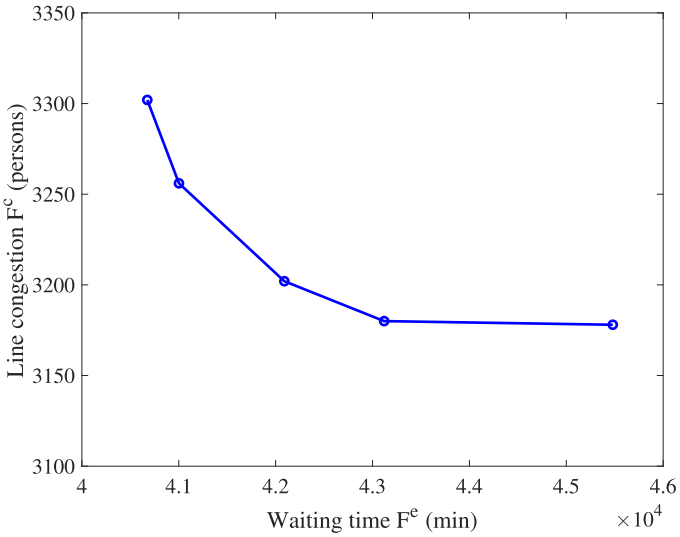

Fig. 9. Pareto frontier.

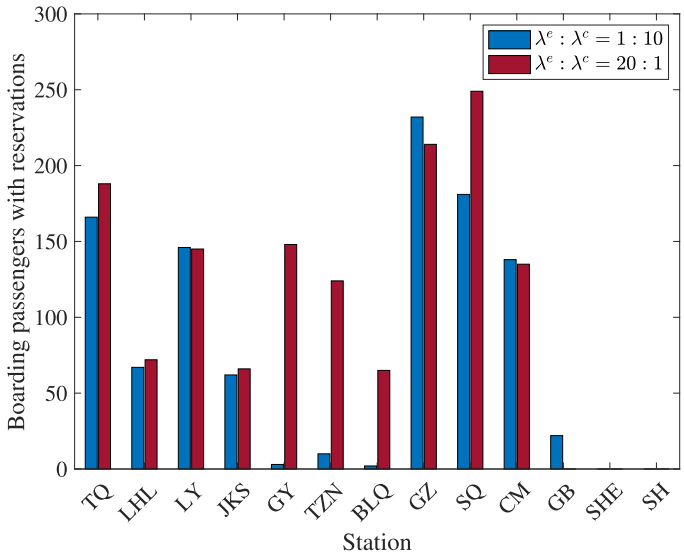

Fig. 10. The number of passengers with reservations boarding the second train under two settings of weighting coefficients.

Second, when implementing reservation-based travel as an emerging demand management strategy, a dynamic allocation approach for reservation slots should be adopted. Given the time-varying and highly interconnected nature of metro systems, which exhibit strong spatial and temporal coupling between train dynamics, passenger flows, and reservation statuses, time-varying demand management strategies are essential. These strategies not only enhance the passenger travel experience but also ensure the safe and stable operation of the metro system.

7. Conclusions

In this paper, we addressed the joint optimization problem for demand management and train scheduling with reservation strategies. An INLP model is developed by integrating train capacity constraints, reservation slot allocation, passenger flow control, and minimum service ratio constraints for passengers without reservations, with the objective of minimizing passengers' waiting time and line congestion. To enhance computational tractability, the model is reformulated into a linear programming model by introducing additional 0–1 variables and big-M constraints. The derivation of the values of big-M parameters ensures tighter upper bounds, resulting in a more efficient linear program. To solve the real-life problem effectively, we design an algorithm that combines ALNS with a commercial optimization solver. The algorithm employs heuristic rules to generate train schedules and uses the solver to optimize the smaller-scale demand management subproblem, iteratively searching high-quality solutions.

Numerical experiments based on the Beijing metro Batong Line show that the proposed algorithm can achieve high-quality solutions. For instance, it obtains a solution with an optimality gap of 2.44 % within 651 seconds, compared to GUROBI, which requires approximately 2771 seconds to find the optimal solution. For larger-scale practical problems, the proposed algorithm can find a feasible solution within a reasonable time (under one hour), whereas GUROBI fails to produce a feasible solution within two hours, verifying the practicality of the approach.

Additionally, compared to the step-by-step optimization method, the proposed joint optimization approach reduces passengers' waiting time by 9.33 % and line congestion by 0.63 %, highlighting the effectiveness and necessity of integrating demand management and train scheduling decisions. Sensitivity analysis on the minimum service ratio parameter for unreserved passengers reveals that increasing passengers' waiting time by 3.51 % and line congestion by 0.51 % eliminates extreme inequity between reserved (with service ratio of 100 %) and unreserved passengers (with service ratio less than 20 %). These results indicate the importance of implementing a dynamic reservation slot allocation strategy to achieve better service fairness and operational efficiency.

There are several promising directions for future research. One potential direction is to extend the joint optimization of demand management strategies and train schedules to the network level, capturing the interactions and dependencies across multiple metro lines. This extension may introduce significant computational complexity. Therefore, developing decomposition-based solution approaches would also be a valuable direction for future research. Additionally, incorporating uncertainty in passenger arrival times into the optimization framework represents a valuable direction for future research, which could further improve the service reliability under real-world operational conditions.

CRediT authorship contribution statement

Yahan Lu: Writing – original draft, Conceptualization, Methodology, Software, Visualization, Formal analysis; **Lixing Yang:** Writing – review & editing, Conceptualization, Investigation, Project administration, Funding acquisition; **Dongyang Xia:** Investigation, Software, Validation, Writing – review & editing; **Fanting Meng:** Project administration, Funding acquisition, Data curation; **Shadi Sharif Azadeh:** Investigation, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 72288101), and the R&D Program of Beijing Municipal Education Commission (No. KM202311417005).

Appendix A. Procedure of the LNS algorithm

In this section, we present the detailed procedure of the LNS algorithm.

Algorithm 1 Procedure of the LNS algorithm.

Input: Infrastructure layout; Passenger demand
Generate an initial feasible solution H
while termination conditions are not satisfied **do**
    Select a destroy operator and apply it to H to generate a partial solution H'
    Apply the repair operator: if constraints (34) are violated, randomly select a headway and reduce its value to ensure feasibility, obtaining a repaired solution $\tilde{H}$
    Call Gurobi to solve the DM subproblem (36) with fixed timetabling solution x from $\tilde{H}$ to compute detailed time-varying reservation slot allocation plans and passenger flow control plans
    Evaluate the objective value of the candidate solution
    if the new solution $\tilde{H}$ is accepted **then**
        $H \leftarrow \tilde{H}$
    end if
end while
Output: Train schedules; Time-varying reservation slot allocation plans; Passenger flow control plans

Appendix B. Formulation for the single-station reservation allocation optimization approach

Similar to Tang et al. (2024), we consider a fixed train timetable with even headways. Reservation is only available at the first station. Reserved passengers are guaranteed to board. Non-reserved passengers can board if capacity remains. Thus, y_{ist} now is an input parameter, which represents whether time t falls within the headway between train services $i-1$ and i. We define a new variable $C_{is}^{remaining}$ to indicate the remaining train capacity after passengers alights service i at station s. The variables $\hat{o}_{is}$ and $\hat{l}_{is}$ are defined to represent the number of in-vehicle passengers with reservations when train i departs from station s and the number of passengers with reservations leaves train i when it arrives at station s. The formulation for this single-station reservation allocation optimization approach can be expressed as follows:

$$
\begin{cases}
\min \quad \lambda^e F^e + \lambda^c F^c \\
\text{s.t.} \quad F^e = \Delta\left[\sum_{i\in\mathcal{I}}\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}}\left(y_{ist}\sum_{t'\in\mathcal{T},t'\le t}y_{ist'}\sum_{v\in\mathcal{S},v>s}D_{svt}\right)+\sum_{i\in\mathcal{I}}\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}}\left(y_{ist}\sum_{v\in\mathcal{S},v>s}r_{isv}\right)\right], \\
F^c = \sum_{i\in\mathcal{I}} z_i, \\
\hat{b}_{i1v} = \sum_{t\in\mathcal{T}} y_{i1t}\varrho_{1t} \qquad \forall i\in\mathcal{I}, v\in\mathcal{S}, v>1, \\
\sum_{i\in\mathcal{I}}\hat{b}_{i1v} = \sum_{t\in\mathcal{T}}\varrho_{1vt} \qquad \forall v\in\mathcal{S}, v>1, \\
w_{i1v} = \begin{cases} \sum_{t\in\mathcal{T}} x_{ist}(D_{1vt}-\varrho_{1vt}) & i=1 \\ \sum_{t\in\mathcal{T}} x_{ist}(D_{1vt}-\varrho_{1vt}) - \sum_{j\in\mathcal{I},j<i} b_{j1v} & i\in\mathcal{I}\setminus\{1\} \end{cases} \qquad \forall v\in\mathcal{S}, v>1, \\
w_{isv} = \begin{cases} \sum_{t\in\mathcal{T}} x_{ist}D_{svt} & i=1 \\ \sum_{t\in\mathcal{T}} x_{ist}D_{svt} - \sum_{j\in\mathcal{I},j<i} b_{jsv} & i\in\mathcal{I}\setminus\{1\} \end{cases} \qquad \forall s,v\in\mathcal{S}, s>1, v>s, \\
r_{isv} = w_{isv} - b_{isv} \qquad \forall i\in\mathcal{I}, s, v\in\mathcal{S}, v>s, \\
\sum_{i\in\mathcal{I}} b_{i1v} = \sum_{t\in\mathcal{T}}(D_{1vt}-\varrho_{1vt}) \qquad \forall v\in\mathcal{S}, v>1, \\
\sum_{i\in\mathcal{I}} b_{isv} = \sum_{t\in\mathcal{T}} D_{svt} \qquad \forall s,v\in\mathcal{S}, s>1, v>s, \\
\sum_{v\in\mathcal{S},v>s} b_{isv} = \min\{\sum_{v\in\mathcal{S},v>s} w_{isv}, C_{is}^{remaining}\} \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, \\
\hat{o}_{is} = \begin{cases} \sum_{v\in\mathcal{S},v>s}\hat{b}_{isv} & s=1 \\ \hat{o}_{i(s-1)} - \hat{l}_{is} & s\in\mathcal{S}\setminus\{1,|\mathcal{S}|\} \\ 0 & s=|\mathcal{S}| \end{cases} \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, \\
o_{is} = \begin{cases} \sum_{v\in\mathcal{S},v>s} b_{isv} & s=1 \\ o_{i(s-1)} - l_{is} + \sum_{v\in\mathcal{S},v>s} b_{isv} & s\in\mathcal{S}\setminus\{1,|\mathcal{S}|\} \\ 0 & s=|\mathcal{S}| \end{cases} \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, \\
C_{is}^{remaining} = \begin{cases} C & s=1 \\ C - O_{i(s-1)} - \hat{O}_{i(s-1)} + l_{is} & s\in\mathcal{S}\setminus\{1\} \end{cases} \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, \\
l_{is} = \begin{cases} 0 & s=1 \\ \sum_{m\in\mathcal{S},m\le s-1} b_{ims} & s\in\mathcal{S}\setminus\{1\} \end{cases} \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, \\
\hat{l}_{is} = \begin{cases} 0 & s=1 \\ \hat{b}_{i1s} & s\in\mathcal{S}\setminus\{1\} \end{cases} \qquad \forall i\in\mathcal{I}, s\in\mathcal{S}, \\
z_i \ge \sum_{s,v\in\mathcal{S},v>s}(w_{isv}+\hat{b}_{isv}) \qquad \forall i\in\mathcal{I}, \\
\varrho_{1vt}\in\mathbb{Z}_{\ge 0} \qquad \forall v\in\mathcal{S}, v>1, t\in\mathcal{T}, \\
b_{isv}\in\mathbb{Z}_{\ge 0} \qquad \forall i\in\mathcal{I}, s, v\in\mathcal{S}, v>s.
\end{cases} \tag{B.1}
$$

References

An, J., Mikhaylov, A., Jung, S.-U., 2021. A linear programming approach for robust network revenue management in the airline industry. J. Air Trans. Manag. 91, 101979.

Beijing Daily, 2021. Starting next week, five more subway stations will implement passenger flow control, bringing the total to 40 stations across 10 lines during morning rush hours Accessed: 2025-01-18. https://xinwen.bjd.com.cn/content/s61165587e4b0f21db0830769.HTML.

Beijing Municipal Commission of Transport, 2020. The station entry reservation trial will be launched at two beijing metro stations starting from march 6. https://www.beijing.gov.cn/gongkai/shuju/sjjd/202104/t20210419_2361873.html. [Accessed May 19, 2025].

Bertsimas, D., De Boer, S., 2005. Simulation-based booking limits for airline revenue management. Oper. Res. 53 (1), 90–106.

Binder, S., Maknoon, M.Y., Sharif Azadeh, S., Bierlaire, M., 2021. Passenger-centric timetable rescheduling: a user equilibrium approach. Transpor. Res. Part C: Emerg. Technol. 132, 103368.

Cacchiani, V., Qi, J., Yang, L., 2020. Robust optimization models for integrated train stop planning and timetabling with passenger demand uncertainty. Transpor. Res. Part B: Methodol. 136, 1–29.

Gkiotsalitis, K., Schmidt, M., van der Hurk, E., 2022. Subline frequency setting for autonomous minibusses under demand uncertainty. Transpor. Res. Part C: Emerg. Technol. 135, 103492.

Gong, C., Luan, X., Yang, L., Qi, J., Corman, F., 2024. Integrated optimization of train timetabling and rolling stock circulation problem with flexible short-turning and energy-saving strategies. Transpor. Res. Part C: Emerg. Technol. 166, 104756.

Guangzhou Municipal People's Government, 2023. Guangzhou metro adds 10 stations with regular passenger flow control. [Accessed January 18, 2025]. https://www.gz.gov.cn/zwfw/zxfw/jtfw/content/post_8828685.HTML.

Hu, Y., Li, S., Wang, Y., Zhang, H., Wei, Y., Yang, L., 2023. Robust metro train scheduling integrated with skip-stop pattern and passenger flow control strategy under uncertain passenger demands. Compt. Operat. Res. 151, 106116.
Jiang, H., Barnhart, C., 2009. Dynamic airline scheduling. Transp. Sci. 43 (3), 336–354.
Lamotte, R., De Palma, A., Geroliminis, N., 2017. On the use of reservation-based autonomous vehicles for demand management. Transp. Res. Part B: Methodol. 99, 205–227.
Lee, T.C., Hersh, M., 1993. A model for dynamic airline seat inventory control with multiple seat bookings. Transp. Sci. 27 (3), 252–265.
Li, X., Lu, Y., Yang, L., 2024. Collaborative optimization of passenger flow control and bus-bridging services in commuting metro lines. Appl. Math. Model. 130, 806–826.
Li, X., Yang, H., Ke, J., 2023. Booking cum rationing strategy for equitable travel demand management in road networks. Transp. Res. Part B: Methodol. 167, 261–274.
Liang, J., Ren, M., Huang, K., Gao, Z., 2024. Data-driven timetable design and passenger flow control optimization in metro lines. Transpor. Res. Part C: Emerg. Technol. 166, 104761.
Liu, R., Li, S., Yang, L., 2020. Collaborative optimization for metro train scheduling and train connections combined with passenger flow control strategy. Omega 90, 101990.
Liu, W., Yang, H., Yin, Y., 2015. Efficiency of a highway use reservation system for morning commute. Transpor. Res. Part C: Emerg. Technol. 56, 293–308.
Lu, Y., Yang, L., Yang, H., Zhou, H., Gao, Z., 2023. Robust collaborative passenger flow control on a congested metro line: a joint optimization with train timetabling. Transp. Res. Part B: Methodol. 168, 27–55.
Lu, Y., Yang, L., Yang, K., Gao, Z., Zhou, H., Meng, F., Qi, J., 2022. A distributionally robust optimization method for passenger flow control strategy and train scheduling on an urban rail transit line. Engineering 12, 202–220.
Ministry of Transport of the People's Republic of China, 2021. Smooth metro travel, comfortable and worry-free. [Accessed May 19, 2025]. https://www.mot.gov.cn/zhuanti/dangshixxjy/xuexidt/difang/202110/t20211009_3621269.HTML.
Molnar, G., De Almeida Correia, G.H., 2019. Long-term vehicle reservations in one-way free-floating carsharing systems: a variable quality of service model. Transpor. Res. Part C: Emerg. Technol. 98, 298–322.
Ouyang, Y., Yang, H., Daganzo, C.F., 2021. Performance of reservation-based carpooling services under detour and waiting time restrictions. Transp. Res. Part B: Methodol. 150, 370–385.
People's Daily Beijing, 2020. Beijing metro reservation entry will be gradually promoted: Passengers can choose voluntarily. [Accessed January 22, 2025]. http://bj.people.com.cn/n2/2020/0430/c82840-33987970.HTML.
Polinder, G.-J., Schmidt, M., Huisman, D., 2021. Timetabling for strategic passenger railway planning. Transp. Res. Part B: Methodol. 146, 111–135.
Robenek, T., Sharif Azadeh, S., Maknoon, Y., De Lapparent, M., Bierlaire, M., 2018. Train timetable design under elastic passenger demand. Transp. Res. Part B: Methodol. 111, 19–38.
Robinson, L.W., 1995. Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes. Oper. Res. 43 (2), 252–263.
Scheepmaker, G.M., Goverde, R. M.P., Kroon, L.G., 2017. Review of energy-efficient train control and timetabling. Eur. J. Oper. Res. 257 (2), 355–376.
Schettini, T., Gendreau, M., Jabali, O., Malucelli, F., 2023. An iterated local search metaheuristic for the capacitated demand-driven timetabling problem. Transp. Sci. 57 (5), 1379–1401.
Shentong Metro Group, 2024. Morning rush hour on August 12: These metro stations plan to implement passenger flow control. Accessed: 2025-01-18. https://finance.sina.com.cn/jjxw/2024-08-11/doc-incihsqa0093178.shtml.
Shi, J., Yang, J., Yang, L., Tao, L., Qiang, S., Di, Z., Guo, J., 2023. Safety-oriented train timetabling and stop planning with time-varying and elastic demand on overcrowded commuter metro lines. Transp. Res. Part E: Logist. Transp. Rev. 175, 103136.
Shi, J., Yang, L., Yang, J., Gao, Z., 2018. Service-oriented train timetabling with collaborative passenger flow control on an oversaturated metro line: an integer linear optimization approach. Transp. Res. Part B: Methodol. 110, 26–59.
Tang, J., Wu, J., Zhang, P., Zhang, Y., Cao, J., 2024. Modelling reservation strategies for managing peak-hour stranding on an oversaturated metro line. Transpor. Res. Part C: Emerg. Technol. 167, 104819.
Tundulyasaree, K., Martin, L., van Lieshout, R.N., Woensel, T.V., 2025. Optimal taxes and subsidies to incentivize modal shift for inner-city freight transport. arxiv:2501.09467.
Wang, H., Yang, L., Zhang, J., Luo, Q., Fan, Z., 2024. Real-time train timetabling with virtual coupling operations on a y-type metro line. Eur J Oper Res 319 (1), 168–190.
Xia, D., Ma, J., Sharif Azadeh, S., 2024a. Integrated timetabling and vehicle scheduling of an intermodal urban transit network: a distributionally robust optimization approach. Transpor. Res. Part C: Emerg. Technol. 162, 104610.
Xia, D., Ma, J., Sharif Azadeh, S., 2024b. Integrated timetabling, vehicle scheduling, and dynamic capacity allocation of modular autonomous vehicles under demand uncertainty. arxiv:2410.16409.
Xia, D., Ma, J., Sharif Azadeh, S., Zhang, W., 2023. Data-driven distributionally robust timetabling and dynamic-capacity allocation for automated bus systems with modular vehicles. Transpor. Res. Part C: Emerg. Technol. 155, 104314.
Yang, K., Lu, Y., Yang, L., Gao, Z., 2021. Distributionally robust last-train coordination planning problem with dwell time adjustment strategy. Appl. Math. Model. 91, 1154–1174.
Yang, L., Lu, Y., Yin, J., Sharif Azadeh, S., 2025. Integrated demand-side management and timetabling for an urban transit system: A Benders decomposition approach. arxiv:2502.12952.
Yin, J., D'Ariano, A., Wang, Y., Yang, L., Tang, T., 2021. Timetable coordination in a rail transit network with time-dependent passenger demand. Eur. J. Oper. Res. 295 (1), 183–202.
Yin, J., Pu, F., Yang, L., D'Ariano, A., Wang, Z., 2023. Integrated optimization of rolling stock allocation and train timetables for urban rail transit networks: a benders decomposition approach. Transp. Res. Part B: Methodol. 176, 102815.
Yuan, Y., Li, S., Liu, R., Yang, L., Gao, Z., 2023. Decomposition and approximate dynamic programming approach to optimization of train timetable and skip-stop plan for metro networks. Transpor. Res. Part C: Emerg. Technol. 157, 104393.
Yuan, Y., Li, S., Yang, L., Gao, Z., 2022. Real-time optimization of train regulation and passenger flow control for urban rail transit network under frequent disturbances. Transp. Res. Part E: Logist. Transp. Rev. 168, 102942.
Zhan, S., Xie, J., Wong, S.C., Zhu, Y., Corman, F., 2024. Handling uncertainty in train timetable rescheduling: a review of the literature and future research directions. Transp. Res. Part E: Logist. Transp. Rev. 183, 103429.
Zhang, D., Gao, Y., Yang, L., Cui, L., 2024a. Timetable synchronization of the last several trains at night in an urban rail transit network. Eur. J. Oper. Res. 313 (2), 494–512.
Zhang, Y., Li, S., Yuan, Y., Zhang, J., Yang, L., 2024b. Approximate dynamic programming approach to efficient metro train timetabling and passenger flow control strategy with stop-skipping. Eng. Appl. Artif. Intell. 127, 107393.
Zhu, Y., Goverde, R. M.P., 2019. Railway timetable rescheduling with flexible stopping and flexible short-turning during disruptions. Transp. Res. Part B: Methodol. 123, 149–181.
Zhu, Y., Goverde, R. M.P., 2021. Dynamic railway timetable rescheduling for multiple connected disruptions. Transpor. Res. Part C: Emerg. Technol. 125, 103080.