https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X, EISSN 1526-5463

Submitted to *Operations Research*

Urban rail rolling stock scheduling with short-turn operations and passenger flow management

(Authors' names are not included for peer review)

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and are not intended to be a true representation of the article's final published form. Use of this template to distribute papers in print or online or to submit papers to another non-INFORM publication is prohibited.

Abstract. Urban rail lines with uneven spatial and temporal passenger distributions often face severe congestion. In such cases, flexible short-turn operations and passenger flow management, including *passenger flow control* (i.e., restricting passenger entry at stations) and *passenger flow assignment* (i.e., allocating passengers to trips), can help balance transport supply and passenger demand, but increase the complexity of the rolling stock scheduling process. This paper addresses a rolling stock scheduling problem that jointly optimizes short-turn operations and passenger flow management. A novel path-based model is formulated with minimizing both the rolling stock operational cost and passenger-related cost. The period-based passenger flow management is incorporated to capture temporal variations of passenger distributions across physical segments and periods. To solve this model, we develop a customized Benders decomposition and column generation algorithm, where the pricing sub-problem is solved by a single-path labeling algorithm. To reduce the scale of the branch-and-bound search, we apply a column enumeration and reduction method with a multi-path labeling algorithm. The computational results demonstrate that the proposed joint optimization process achieves an average cost reduction of 15.19% compared to the sequential optimization process, while short-turn operations yield an average reduction of 2.94% compared to the traditional rolling stock scheduling scheme.

Key words: Rolling stock scheduling; Short-turn operations; Period-based passenger flow representation; Passenger flow control and assignment; Column enumeration and reduction

1. Introduction

Over-saturated passenger demand has increasingly constrained the service quality of urban rail systems, leading to longer waiting times, platform congestion, and uncomfortable travel experiences, which have become common in several megacities such as New York, Singapore, Hong Kong and Beijing. For instance, the Beijing urban rail system transported more than 3.454 billion passengers in 2024, surpassing the comfortable capacity of both trains and stations. Thus, it is difficult to manage peak-hour congestion with such high volumes, which can increase not only the passenger delay at stations but also the risk of safety incidents (e.g., stampede accidents on platforms and in train cars).

A primary cause of these issues is the uneven spatial and temporal distribution of arriving passenger flows along urban rail lines. Two effective and practical strategies for mitigating these imbalances are short-turn operations and passenger flow management (including *passenger flow control* and *passenger flow assignment*). Without passenger flow management, passengers will enter unrestrictedly during peak hours, leading to severe platform overcrowding as accumulated demand exceeds capacity. In comparison, passenger flow control can substantially reduce congestion and operational risks by regulating passengers from station halls to platforms through automatic fare collection (AFC) entrances (Liang et al. 2023, Lu et al. 2022). In addition, when trains become overcrowded at the earlier stations, passengers boarding at subsequent stations may face excessive delays or need to abandon the train. This could create frustration and drive them to abandon the train system, which ultimately leads to decreased long-term adoption of public transportation. Therefore, the coordinated passenger flow control across stations can help balance waiting passengers and the allocation of rolling stock capacity at the stations along the urban rail line, enabling service-oriented train operations. In short-turn operations, some rolling stock units travel on partial physical segments and turn around at some candidate stations, thereby increasing service frequency on certain high-demand physical segments. Lastly, to coordinate the benefits of short-turn operations and passenger flow management, *rolling stock scheduling* (defined by the train schedule and the rolling stock circulation) plays a pivotal role in balancing transport supply with passenger demand (Shi et al. 2018, Lu et al. 2023, Grimm et al. 2025).

Compared with emerging operational modes (e.g., flexible train compositions in Pan et al. (2023) or cross-line operations in Van Lieshout (2021)) with significant infrastructure investments (e.g., platform extensions, connection track construction), short-turn operations and passenger flow management can be implemented more quickly and at lower cost. These strategies have therefore been widely adopted in practice. For example, the 2021 annual report of the Shanghai (China) urban rail system noted 22,000 short-turn and 51,000 passenger flow control operations, accounting for the majority of temporary and regular dispatch adjustments.

The accurate modeling of platform congestion remains challenging because passenger queuing is a complex, time-dependent process (Shi and Li 2021, Zhang et al. 2024). In practice, passenger flow control schemes are typically updated at fixed periods (e.g., one hour) to avoid excessive operational workload, such as the limitation of the number of open gates at AFC entrances. This motivates the use of period-based passenger flow representations, in which both passenger arrival and control rate functions are defined over discrete periods rather than high-resolution timestamps. This method allows passenger-related objectives to be captured efficiently, enabling scalable solutions while remaining consistent with operational realities.

Focusing on providing system-optimal solutions, this paper investigates a rolling stock scheduling problem with short-turn operations and period-based passenger flow management (RSP-STPM). The aim is to jointly determine: paths for rolling stock units, continuous-time passenger control rate functions at stations, and

passenger flow assignments to trips and periods. Lastly, the objective is to minimize both rolling stock operational costs and passenger-related costs.

The remainder of this paper is organized as follows. Section 2 reviews the related literature and highlights the contributions of this study. Section 3 presents the problem statement, including definitions, concepts, and notation. Section 4 formulates a path-based model for joint rolling stock scheduling and passenger flow management problem. Section 5 designs a customized Benders decomposition with column generation (BDCG) algorithm for solving the model. Section 6 reports numerical experiments based on real-world data from the Beijing urban rail system. Finally, Section 7 concludes with key findings and directions for future research.

2. Literature review and contribution of this work

Passenger-oriented train operations are the core component of urban rail systems and have attracted considerable attention from both industry and academia in recent decades (see Caprara et al. (2007), Lusby et al. (2011), Sharma et al. (2023), Duarte et al. (2025)). In the following, we review the state of the art from three perspectives: passenger-oriented rolling stock scheduling, passenger flow modeling, and passenger flow management in urban rail systems.

2.1. Passenger-oriented rolling stock scheduling

The passenger-oriented rolling stock scheduling problem aims to jointly determine train schedules and rolling stock circulations based on passenger flow distributions (van der Hurk et al. 2018, Hoogervorst et al. 2020). In early studies, time-dependent passenger flows over the planning horizon were aggregated for each direction or physical segment into a single demand value, which was then satisfied by scheduling a sufficient number of trips and rolling stock units (Cordeau et al. 2000, 2001, Lingaya et al. 2002, Abbink et al. 2004). Other studies estimated the expected passenger demand for each trip from ticket sales data and seat capacity to determine the required trips (Fioole et al. 2006, Cacchiani et al. 2010b, 2013, Lin and Kwan 2016). Overall, such static methods cannot precisely capture passenger flows with dynamic spatial and temporal variations.

With the rapid development of urban rail systems, the topic of passenger-oriented rolling stock scheduling has evolved to incorporate time-dependent passenger flow distributions (Kroon et al. 2015). In this context, both trip frequencies and departure times depend on passenger flow distributions across different segments and periods, and the objective function aims to balance rolling stock operational cost with passenger waiting cost. To alleviate the passenger flow oversaturation, some emerging operational modes, such as flexible train compositions (Pan et al. 2023) and cross-line operations (Zhou et al. 2021, Van Lieshout 2021), are implemented to match uneven spatial and temporal passenger flow distributions, although these often require substantial infrastructure modifications.

Short-turn operations, in contrast, allow rolling stock units to turn around at some designated stations before reaching their terminal stations, resulting in more flexible transport capacity allocations across

different physical segments. That is, high-demand segments get more capacity, while low-demand segments get less. In the literature, short-turn operations are mainly applied in two typical scenarios, i.e., to address spatially uneven passenger flow distributions by scheduling additional short-turn trips on some physical segments with high demands (Schettini et al. 2022), and to respond to rail disruptions by rescheduling trips to unaffected physical segments (Ghaemi et al. 2017, Zhu and Goverde 2019, Ying et al. 2024).

2.2. Passenger flow modeling

In the literature, passenger flow is typically characterized from spatial and temporal perspectives. From the spatial perspective, time-dependent passenger flows are mainly modeled using three common methods: (1) the OD-based representation, which specifies the origin station, destination station, and the arrival time at the origin station (Niu and Zhou 2013, Lu et al. 2023) for each passenger or passenger group; (2) the segment-based representation, which aggregates passenger counts between each pair of adjacent stations (i.e., physical segments), without recording individual origins and destinations (Sun et al. 2014, Pan et al. 2023); and (3) the network-flow-based representation, which treats passengers as aggregated flows satisfying flow-balance constraints on space-time vertices and arcs, without tracking individual trajectories (Yin et al. 2022, Wang et al. 2024). Particularly, these last two approaches are well-suited to large-scale cases where individual passenger routes need not be modeled explicitly.

From the temporal perspective, the planning horizon is usually discretized into a set of timestamps with a predefined granularity (e.g., 1 min). Then, passenger-related parameters and variables are indexed by these timestamps in the mathematical model. To couple passenger loading with train departures at platforms, Niu and Zhou (2013) proposed a cumulative-flow variable formulation in a space-time network, which has been widely applied in passenger-oriented urban rail operation models (Niu et al. 2015, Lu et al. 2023). However, this formulation requires a large number of parameters, decision variables, and constraints for passenger arrival, waiting, boarding, and alighting movements at discrete timestamps, resulting in computational complexity for general-purpose solvers (e.g., CPLEX) in large networks with long planning horizons. To mitigate this issue, some continuous-time methods from the road traffic modelling might be adapted in urban rail operations, as they represent the flow evolution without discretizing the planning horizon into timestamps (e.g., the Newell model in Newell (2013)). Such methods can evaluate specific objectives (e.g., flow volumes, delays) by avoiding introducing a large number of discrete timestamps in the modeling process (Lu et al. 2022, Zhou et al. 2022).

2.3. Passenger flow management

Passenger flow management can be regarded as a capacity allocation process that determines passenger flow control rates and assigns passengers across trips by using guidance and capacity reservations (Nguyen et al. 2001, Lyu et al. 2019, Liang et al. 2023). It allocates shares of train capacity for boarding at stations along the urban rail line, without implying seat reservations as in advance ticketing.

In practice, passenger flow control can be applied locally at a single station or coordinately on the whole urban rail line, depending on different passenger flow distributions. Specifically, some local passenger flow control operations can reduce platform overcrowding and safety risks, such as pacing entries based on predicted dwelling times (Delgado et al. 2012), partitioning platform waiting zones via physical barriers (Davidich et al. 2013), and simulating pedestrians to design passenger-guidance facilities (Li et al. 2020). When the passenger flow distribution is uneven across stations, some line-wide passenger flow control operations coordinate entry rates and capacity allocations across multiple stations and trips to better match transport supply and passenger demand, such as reserving capacity shares for specific stations on upcoming trains (Yazdani et al. 2019) and implementing skip-stopping in peak corridors (Jiang et al. 2017).

In the literature, two major optimization objectives are commonly considered for passenger flow management problems. First, from the passenger perspective, it is important to reduce total waiting time (Lu et al. 2023) and improve fairness, such as minimizing the number of target-controlled passengers (Liang et al. 2023) and avoiding extremely long waiting times (Wu et al. 2021). Second, from the urban rail operator's perspective, the goals are to reduce congestion risk at stations and make full use of available rolling stock capacity (Shi et al. 2018, Liang et al. 2023, Lu et al. 2023). Finally, several studies have also begun to explore the benefits of jointly optimizing passenger flow control and train scheduling (Corman 2020, Luan and Corman 2022).

To highlight the contributions and differences between this paper and existing studies, Table 1 presents a detailed comparison of notable publications related to rolling stock scheduling and passenger flow management from the perspectives of problem considerations, models, and solution algorithms. Overall, our main contributions are the following:

• We introduce a joint rolling stock scheduling and passenger flow management model to determine trip departure/arrival times at stations, rolling stock circulations, period-based passenger flow control schemes at stations, and period-based passenger flow assignments on trips. Compared with common passenger-oriented rolling stock scheduling models (Schettini et al. 2022, Sharma et al. 2023, Wang et al. 2024), we extend the model to a more flexible mode by allowing short-turn operations at some candidate stations within a unified path-based column generation framework, and by controlling/assigning passenger flows under a novel period-based passenger flow representation.

• We develop an exact Benders decomposition with column generation algorithm to solve the RSP-STPM, in contrast to heuristic algorithms in existing studies on joint train scheduling and passenger flow control problems (Shi et al. 2018, Wu et al. 2021, Lu et al. 2023). Compared with existing column generation frameworks for general rolling stock scheduling problems (Cacchiani et al. 2008, 2010a, Duarte et al. 2025), we implement a column enumeration and reduction method to avoid introducing a large number of branching variables. In addition, two customized labeling algorithms, one with and one without recording explicit paths, are embedded in the column generation and enumeration processes, respectively.

Table 1 Characteristic comparisons of some closely related publications

Publication	Passenger flow					Train operation				Model	Algorithm
	Rep.	OD	Cont.	Assign.	P-Obj.	TS	RSC	ST	T-Obj.		
Niu and Zhou (2013)	DT	Yes	No	Yes	WP,RP, PWT	Yes	No	No	No	IP,NL	GA
Niu et al. (2015)	DT	Yes	No	Yes	PWT	Yes	No	No	TSC	IP,NL	N/A
Shi et al. (2018)	DT	Yes	Yes	Yes	PWT	Yes	No	No	No	IP,L	H,LS
Wu et al. (2021)	DT	Yes	Yes	Yes	PWT	SS	No	No	No	MIP,L	VNS
Yin et al. (2022)	DT,F	No	No	Yes	PTT	Re-	No	No	RSAC	NWF,PM,L	B&P
Liang et al. (2023)	S	Yes	Yes	No	EAP	Fixed	No	No	No	ORA,NL	SDP
Lu et al. (2023)	DT	Yes	Yes	Yes	WP	Yes	No	No	TSC	RO,NL	H,LS
Wang et al. (2024)	DT,F	No	No	Yes	PWT	Yes	Yes	Fixed	RSPC	NWF,PM,L	B&P
This paper	P,CT	Yes	Yes	Yes	PD,RP	Yes	Yes	Flexible	RSPC	PM,L	BDCG

Notes for Table 1:

Passenger flow representation (Rep.); Discrete time-dependent (DT); Continuous time-dependent (CT); Static (S); Network-flow-based (F); Period-based (P). OD-based passenger flow representation (OD); Passenger flow control (Cont.); Passenger flow assignment (Assign.); Passenger flow objective (P-Obj.); Waiting passengers (WP); Residual passengers (RP); Passenger waiting time (PWT); Passenger traveling time (PTT); Expected number of accepted passengers (EAP); Passenger delay (PD); Train schedule (TS); Only skip-stopping decisions (SS); Fixed train schedule (Fixed); Train rescheduling (Re-); Rolling stock circulation (RSC); Short-turn operation (ST); Short-turn operations at the fixed station (Fixed); Short-turn operations at candidate stations (Flexible); Train operation objective (T-Obj.); Train schedule cost (TSC); Rolling stock path cost (RSPC); Rolling stock allocation cost (RSAC); Non-linear model (NL); Linear model (L); Integer program (IP); Mixed integer program (MIP); Online resource allocation model (ORA); Robust optimization model (RO); Network-flow-based model (NWF); Path-based model (PM); Genetic algorithm (GA); Heuristic (H); Local search (LS); Variable neighbourhood search (VNS); Stochastic dynamic program (SDP); Branch-and-price (B&P); Benders decomposition with column generation (BDCG); Not available (N/A), i.e., the specific algorithm is not provided in the study.

• We demonstrate, through a series of computational experiments, that the proposed approach can obtain high-quality solutions for real-world instances by scheduling rolling stock units and managing passenger flows across multiple periods throughout the day. We analyze the sensitivity with respect to passenger flow rate function types, number of trips, and passenger flow evolution periods on small-scale and real-world instances, and compare against rolling stock scheduling without short-turn operations and the traditional sequential optimization scheme.

3. Problem statement

This paper considers a bidirectional urban rail line comprising a set of stations $\mathcal{I}$ (indexed by i), with candidate turn-around stations $\mathcal{I}^T \subseteq \mathcal{I}$, as illustrated in Fig. 1. A set of trips $\mathcal{R}$ (indexed by r) should be scheduled, specifying exact departure and arrival times at each station. The set $\mathcal{K}$ (indexed by k) represents the available rolling stock units, which are pre-stored at various depots to cover these trips. For convenience, homogeneous rolling stock units are divided into multiple groups (denoted by set $\mathcal{G}$ and index g), where all rolling stock units in the same group have the same origin depot and the same destination depot (but the origin and destination depots may differ). Specifically, let $\mathcal{K}_g \subseteq \mathcal{K}$ denote the rolling stock units in group g, and c_g^1 the fixed operational cost of assigning one rolling stock unit from group g.

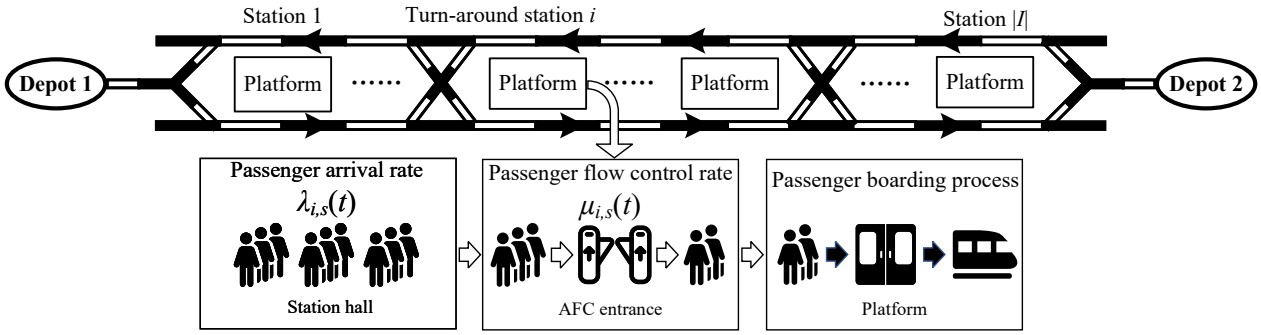


Figure 1 Urban rail line with multiple depots

In urban rail systems, both rolling stock units and trips are scheduled with explicit departure and arrival timestamps at stations. Thus, in the rolling stock scheduling process, the time horizon from time $t_0 = 0$ to time $|\mathcal{T}|$ is discretized into timestamps of a certain time granularity, denoted by a set $\mathcal{T} = \mathbb{Z}_{[t_0,|\mathcal{T}|]} = \{0, 1, \cdots, |\mathcal{T}|\}$. In contrast to the timestamp-based decisions in the rolling stock schedule, passenger arrival flows vary across time, often showing unimodal peaks in the morning/evening and near-constant rates during off-peak hours. In addition, passenger flow control schemes, such as the limitation of the number of open gates at AFC entrances, are typically updated at the start of each period (e.g., hourly). Thus, it is more practical to represent passenger flows over discrete periods than high-resolution timestamps, which can well balance computational efficiency and real-world requirements. Accordingly, the entire planning horizon is partitioned into a finite set of passenger flow evolution periods $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$. For each period s, the *passenger arrival rate function* at station i is $\lambda_{i,s}(t)$ over time windows $[\underline{t}_s, \bar{t}_s]$, where $\underline{t}_s$ and $\bar{t}_s$ are the start and end times of period s. Likewise, the *passenger flow control rate function* $\mu_{i,s}(t)$ regulates the passenger entry rate at the AFC entrance. Note that the passenger flow control rate function can be specified based on a pre-defined function family with practical requirements, e.g., the constant or linear rate function. Then, the decision of the passenger flow control rate function is equivalent to treating the parameters of the function $\mu_{i,s}(t)$ as decision variables in the mathematical model (see Section 4 for details). Compared with the existing passenger flow control approaches (Shi et al. 2018, Liang et al. 2023, Lu et al. 2023), our study generalizes timestamp- and trip-dependent schemes into a period-dependent passenger flow representation, which enhances the flexibility and scalability in passenger flow management scenarios. Lastly, Fig. 2 presents a three-period example with solid curve $\lambda_{i,s}(t)$ and dotted curve $\mu_{i,s}(t)$. In this case, we observe three over-saturated bottlenecks whenever $\lambda_{i,s}(t)$ temporarily exceeds $\mu_{i,s}(t)$.

This study focuses on a rolling stock scheduling problem for a bidirectional urban rail line, integrating short-turn operations and passenger flow management to address spatial and temporal imbalances in passenger flow distributions across multiple periods. The input data includes the urban rail line topology, available rolling stock units and groups, and the time-dependent passenger arrival rate functions. The objective is to determine rolling stock space-time trajectories (i.e., rolling stock schedule), passenger flow control schemes at stations, and passenger flow assignments for all trips and periods, while minimizing both rolling stock operational costs and passenger-related costs.

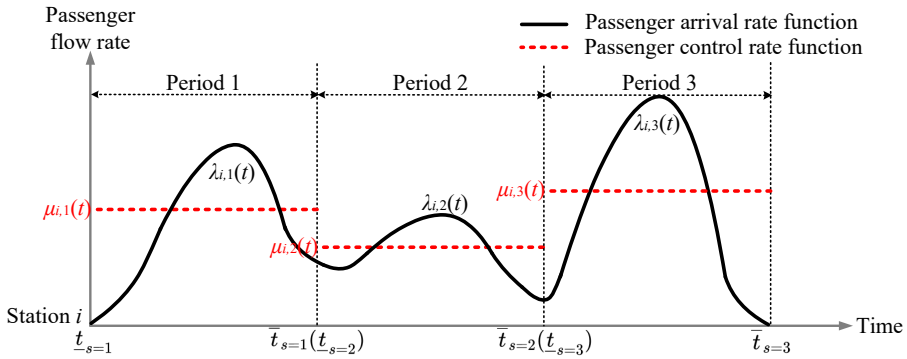


Figure 2 Illustration of period-based passenger flow representation

Example 1: Figs. 3(a–c) illustrates the benefits of different short-turn and passenger flow control operations under uneven spatial (i.e., stations 1 and 2) and temporal (i.e., periods 1 and 2) passenger distributions, where the shaded red denotes the waiting passenger flow volumes on station platforms. In Fig. 3(a), without short-turn and passenger flow control operations, the excess rolling stock capacity appears on parts of trips 1–2 in period 1 and trips 3–4 in period 2. Then, Fig. 3(b) shows that short-turn operations partially cancel trips 1–2 in period 1 and trips 3–4 in period 2 (i.e., dotted lines) with insufficient passenger demand. In this way, the rolling stock operational cost is reduced with shorter trips, but without increasing waiting passenger flow volumes (i.e., the same shaded red areas as Fig. 3(a)). Lastly, Fig. 3(c) shows the combination of short-turn and passenger flow control operations at station 1 in periods 1–2. In this way, compared to Fig. 3(b), by shaving peak waiting passenger flow in period 1 and filling off-peak waiting passenger flow in period 2, the waiting passenger flow volumes are more balanced at station 1 between peak period 1 and off-peak period 2 (i.e., the shaded red area at station 1 in Fig. 3(c)).

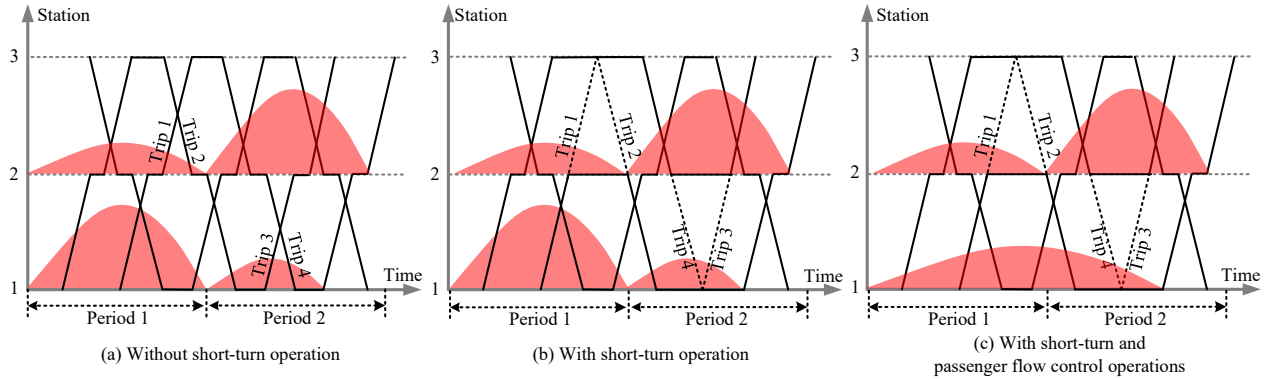


Figure 3 Rolling stock schedules with/without short-turn operations and passenger flow management

3.1. Events and activities

To represent the spatial and temporal movements of rolling stock units, an event-activity network $\mathcal{G}^N = (\mathcal{E}, \mathcal{A})$ is then constructed, which consists of operational events $e \in \mathcal{E}$ and activities $a = (e, e') \in \mathcal{A}$ that capture the movement of rolling stock units and trips between events (Dollevoet et al. 2012, Zhu and Goverde 2019). For each trip, each departure and arrival is represented by a departure event $e \in \mathcal{E}^D$ and arrival event

$e \in \mathcal{E}^A$ with an allowable start time window $\mathcal{T}_e = \mathbb{Z}_{[\underline{T}_e, \overline{T}_e]}$, where $\underline{T}_e$ and $\overline{T}_e$ are the earliest and latest feasible start times of event e. For any event $e \in \mathcal{E} = \mathcal{E}^D \cup \mathcal{E}^A$, let $r(e)$ and $i(e)$ denote its associated trip and station, respectively. The sets $\mathcal{E}_r^{DT}$ and $\mathcal{E}_r^{AT}$ contain the departure and arrival events at candidate turn-around stations for trip r, respectively. Lastly, considering that all rolling stock units in the same group have the same origin depot and the same destination depot, let $e_{\text{ori}(g)}$ and $e_{\text{des}(g)}$ denote the depot source and depot sink for rolling stock units in group g, respectively.

Five types of operational activities are defined: (1) running activity $(e, e') \in \mathcal{A}^R$, representing the movement from departure event e to arrival event e' with minimum running time $T_{e,e'}^{\min}$; (2) dwelling activity $(e, e') \in \mathcal{A}^D$, representing the movement from arrival event e to departure event e' with minimum dwell time $T_{e,e'}^{\min}$; (3) turn-around activity $(e, e') \in \mathcal{A}^T$, representing the reversing of direction from departure event $e \in \mathcal{E}_r^{DT}$ of trip r to arrival event $e' \in \mathcal{E}_{r'}^{AT}$ of trip r' at the same station but in opposite directions. Particularly, the fixed operational cost per minute for running, dwelling, or turn-around activities is denoted by $c_{g,e,e'}^2$ for activity $(e, e') \in \mathcal{A} = \mathcal{A}^R \cup \mathcal{A}^D \cup \mathcal{A}^T$ by rolling stock units in group g; (4) pull-out activity $(e, e') \in \mathcal{A}^{\text{Pout}}$, representing the movement from the depot source $e_{\text{ori}}(g)$ to an arrival event $e' \in \mathcal{E}^A$ at a terminal station; (5) pull-in activity $(e, e') \in \mathcal{A}^{\text{Pin}}$, representing the movement from a departure event $e \in \mathcal{E}^D$ at a terminal station to the depot sink $e_{\text{des}}(g)$.

Based on the definitions of rolling stock operational events and activities, a (rolling stock) path is a sequence of time-dependent events together with pull-in/out and turn-around activities, starting from the depot source $e_{\text{ori}}(g)$ and ending at the depot sink $e_{\text{des}}(g)$ for rolling stock units in group g. Let $\mathcal{P}_g$ (indexed by p) denote the set of feasible paths for group g, and define $\mathcal{P} = \bigcup_{g \in \mathcal{G}} \mathcal{P}_g$ as the set of feasible paths for all rolling stock groups. For each path p, let $g(p)$ indicate its associated rolling stock group. The cost of path $p \in \mathcal{P}$ is denoted by C_p^R and given by $C_p^R = c_{g(p)}^1 + \sum_{(e,e') \in \mathcal{A}} c_{g(p),e,e'}^2 \left(\sum_{t \in \mathcal{T}_{e'}} t \cdot \alpha_{p,e',t} - \sum_{t \in \mathcal{T}_e} t \cdot \alpha_{p,e,t} \right)$, where $\alpha_{p,e,t}$ is a binary parameter equal to 1 if event e is included in path p with start time t, and 0 otherwise.

Example 2: Fig. 4 illustrates the operational events and activities for the case in Fig. 1, which involves two rolling stock units and four trips. For instance, rolling stock unit k_1 pulls out from its depot d_2 and reaches the arrival event (r_1, i_3, A) via a pull-out activity. It then operates trip r_1 from station i_3 to station i_1. Afterward, rolling stock unit k_1 performs a turn-around activity at station i_1 from departure event (r_1, i_1, D) to arrival event (r_3, i_1, A), and operates trip r_3 back to station i_3 before pulling in to depot d_2. Meanwhile, the minimum headway time should be guaranteed between departure/arrival events at the same station in the same direction (i.e., the dotted curve in Fig. 4).

To represent passenger flows with different origins, destinations and trips, passenger activities (e, e') are defined by characterizing passenger flows from event e (associated with trip $r(e)$ and station $i(e)$) to event e' (associated with trip $r(e')$ and station $i(e')$). For modeling convenience in Section 4, passenger activities are classified into five sets: (1) set $\mathcal{A}_{i,j}^{OD}$, representing all passenger activities for OD pair (i, j); (2) set $\mathcal{A}_a^{\text{pre}}$, representing activities preceding passenger activity a with the same OD pair, as shown in Fig. 5(a); (3) set

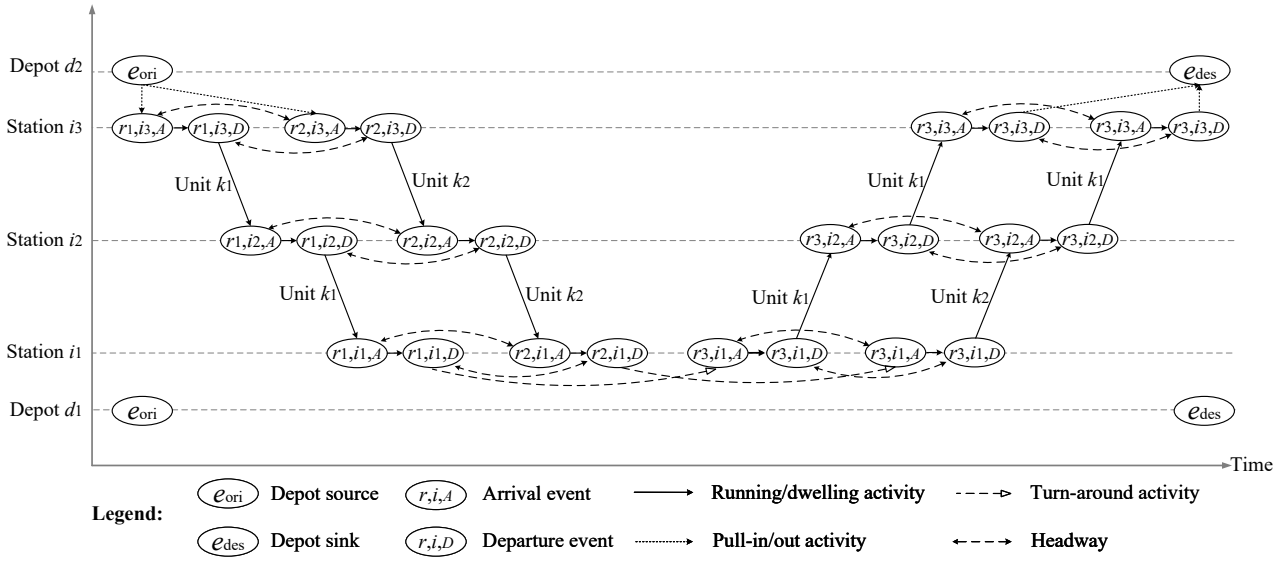
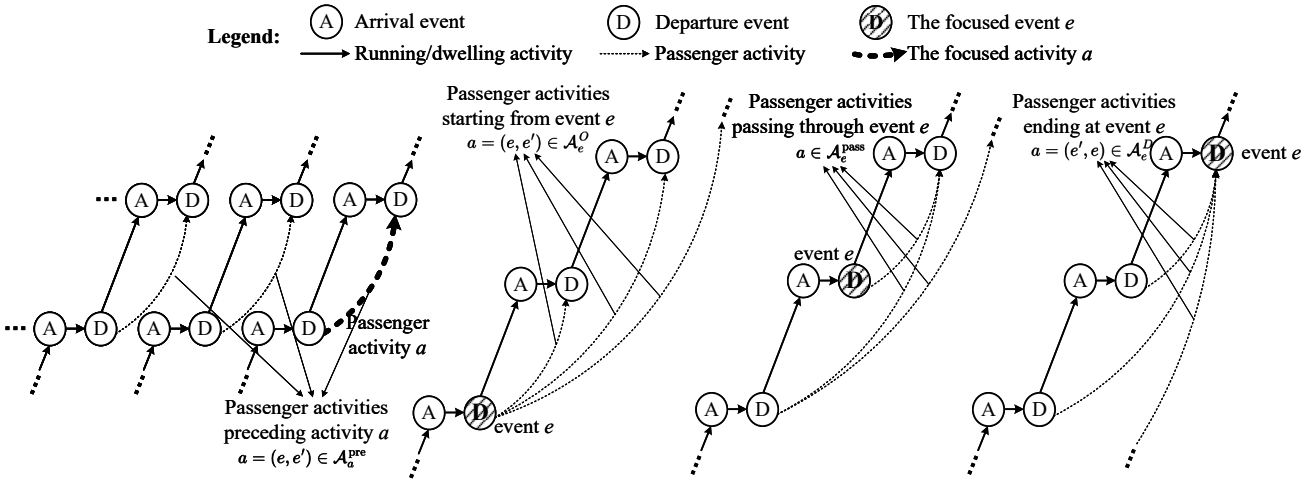


Figure 4 Illustration of rolling stock operational events and activities

$\mathcal{A}_e^O$, representing origin passenger activities starting from departure event e, as shown in Fig. 5(b); (4) set $\mathcal{A}_e^{\text{pass}}$, representing passenger activities passing through departure event e (including those starting at event e but excluding those ending at event e), as shown in Fig. 5(c); (5) set $\mathcal{A}_e^D$, representing destination passenger activities ending at departure event e, as shown in Fig. 5(d).



(a) Previous passenger activity set (b) Passenger origin activity set (c) Passenger pass activity set (d) Passenger destination activity set

Figure 5 Illustration of passenger activity sets

3.2. Assumptions

Without loss of generality, we adopt five assumptions regarding rolling stock operations and passenger flow management:

Assumption 1: As a medium-term planning problem, the rolling stock scheduling process aims to determine the number of required rolling stock units to cover the planned trips with the minimum cost. Therefore, no upper bound is imposed in advance on the number of available units in each group. The

objective is to cover all planned trips by assigning the minimum possible number of rolling stock units, as indicated in the objective function.

Assumption 2: The passenger travel time from the AFC entrance to the platform is neglected.

Assumption 3: To simplify the modeling process of passenger movements, all passengers are assumed to take direct trips to their destination stations rather than boarding a short-turn trip and transferring to another trip. In practice, passengers prefer direct trips because short-turn trips require additional waiting time after arriving at the terminal stations.

Assumption 4: Given the complexity of passenger movements within dwelling time windows, both boarding and alighting are assumed to occur at the departure events of trips. Consequently, only departure events are used to formulate OD-based passenger activities, reducing memory usage and improving computational efficiency.

Assumption 5: We assume that the passenger flow control rate function $\mu_{i,s}(t)$ is a constant function at station i in period s, i.e., $\mu_{i,s}(t) = \widehat{\mu}_{i,s}$, where $\widehat{\mu}_{i,s}$ is the passenger flow control constant rate at station i in period s. This aligns with standard practice, as the passenger flow control parameter is typically updated at fixed intervals (e.g., every hour) by adjusting the number of open gates at AFC entrances.

4. Mathematical formulation

This section presents a path-based mathematical formulation (M1) for the RSP-STPM. Section 4.1 defines the notation, followed by the objective function and system constraints in Section 4.2. Considering that the path-based model cannot be solved directly with general-purpose solvers (e.g., CPLEX) due to the impractical enumeration of the full set of paths, the equivalent primal model of the RSP-STPM is also formulated in Appendix B for reference.

4.1. Notation

The notation used in this study is summarized in Table 4 of Appendix A. The following four types of decision variables are used to formulate the model: path selection variable y_p taking value 1 if path p is selected (otherwise $y_p = 0$), passenger flow assignment variable $p_{a,s}$ indicating the boarding passenger flow volume for activity a in period s, and waiting passenger flow variable $W_{i,s}$ indicating the waiting passenger flow volume at station i at the start of period s. Based on Assumption 5, the decision of the passenger flow control scheme is equivalent to deciding the constant rate $\widehat{\mu}_{i,s}$ in the passenger flow control rate function $\mu_{i,s}(t) = \widehat{\mu}_{i,s}$ at station i in period s. Thus, the passenger flow control constant rate $\widehat{\mu}_{i,s}$ is treated as the decision variable in the mathematical formulation. Lastly, the following two types of auxiliary variables are used to formulate the passenger flow evolution: passenger delay variable $\widehat{w}_{i,s}$ indicating the total passenger delay at station i in period s, and residual passenger flow variable $p^{\text{re}}_{i,j}$ indicating the residual passenger flow volume for OD pair (i, j) at the end of the planning horizon.

4.2. Objective function and system constraints

The objective function is formulated over the aggregated path set $\mathcal{P} = \bigcup_{g\in\mathcal{G}} \mathcal{P}_g$ for all groups g. Specifically, the objective function in Eq. (1) minimizes the total cost of selected paths, passenger delay, passenger flow control schemes, and residual passenger penalties, where C_p^R is the cost of path p, $c_{i,s}^D$ and $c_{i,s}^C$ are the unit costs of passenger delay and flow control rate at station i in period s, and $c_{i,j}^{\text{re}}$ is the unit penalty cost of residual passenger flow volume for OD pair (i,j):

$$\min \quad \sum_{g\in\mathcal{G}} \sum_{p\in\mathcal{P}_g} C_p^R y_p + \sum_{i\in\mathcal{I}} \sum_{s\in\mathcal{S}_i} \left(c_{i,s}^D \widehat{w}_{i,s} + c_{i,s}^C \int_{\underline{t}_s}^{\bar{t}_s} \widehat{\mu}_{i,s} dt \right) + \sum_{(i,j)\in\mathcal{OD}} c_{i,j}^{\text{re}} p_{i,j}^{\text{re}}. \tag{1}$$

We observe that short-turn operations are associated with specific paths, and that the related constraints should be formulated in the pricing sub-problem within the column generation framework (see Section 5.3). Thus, the feasibility and consistency of the model are guaranteed by four groups of remaining path-based system constraints, including the rolling stock scheduling constraints, passenger flow management constraints, passenger delay calculation constraints, and the ranges of the decision and auxiliary variables, which are expressed and described in Groups I–IV as follows.

Group I: Rolling stock scheduling constraints

$$\sum_{p\in\mathcal{P}_g} y_p \le |\mathcal{K}_g| \quad \forall g \in \mathcal{G} \tag{2}$$

$$\sum_{p\in\mathcal{P}} \beta_{p,r} \cdot y_p = 1 \quad \forall r \in \mathcal{R} \tag{3}$$

$$\sum_{p\in\mathcal{P}} \sum_{t\in\mathcal{T}_e} \alpha_{p,e,t} \cdot y_p \le 1 \quad \forall e \in \mathcal{E} \tag{4}$$

$$\sum_{p\in\mathcal{P}} \alpha_{p,e,t} \cdot y_p + \sum_{p\in\mathcal{P}} \sum_{t'=t}^{t^H_{e,e',t}} \alpha_{p,e',t'} \cdot y_p \le 1 \quad \forall t \in \mathcal{T}^{\mathcal{V}}_{e,e'}, (e,e') \in \mathcal{V} \tag{5}$$

Constraints (2) limit the total number of selected paths (i.e., the total number of rolling stock units) for each group g to its capacity $|\mathcal{K}_g|$. Constraints (3) require that each trip r is served by one rolling stock unit, where the binary parameter $\beta_{p,r}$ equals to 1 if and only if trip r is included in path p. Constraints (4) ensure that each event e is assigned at most one start time (or is cancelled) within its allowable time window $\mathcal{T}_e$. Finally, constraints (5) enforce safety headway requirements between events in the same direction at the same station. Specifically, we pre-generate a set $\mathcal{V}$ of potentially conflicting event pairs, which belong to different events in the same direction at the same station, such that $\mathcal{V} = \{(e,e') \,|e,e' \in \mathcal{E}, i(e) = i(e'), r(e) \neq r(e')\}$. Then, we generate the set $\mathcal{T}^{\mathcal{V}}_{e,e'}$ of the possible conflicting timestamps for each potential conflicting event pair $(e,e') \in \mathcal{V}$. If event e starts at timestamp t, event e' is prohibited from starting within the interval $\left[t, t^H_{e,e',t}\right]$, where $t^H_{e,e',t} = \min\{t + h^{\min}_{e,e'} - 1, |\mathcal{T}|\}$ denotes the latest incompatible start time for event pair (e,e'), and $h^{\min}_{e,e'}$ is the minimum headway between events e and e'.

Group II: Passenger flow management constraints

$$\sum_{a' \in \mathcal{A}_a^{\text{pre}}} \sum_{s' \in \mathbb{Z}_{[0,s]}} p_{a',s'} \leq \sum_{s' \in \mathbb{Z}_{[0,s]}} \int_{\underline{t}_{s'}}^{\bar{t}_{s'}} \widehat{\mu}_{i,s'} dt \cdot \gamma_{i,j,s'} \quad \forall a \in \mathcal{A}_{i,j}^{OD}, (i,j) \in \mathcal{OD} \tag{6}$$

$$\sum_{a \in \mathcal{A}_{i,j}^{OD}} \sum_{s \in \mathcal{S}_i} p_{a,s} = \sum_{s \in \mathcal{S}_i} \int_{\underline{t}_s}^{\bar{t}_s} \widehat{\mu}_{i,s} dt \cdot \gamma_{i,j,s} \quad \forall (i,j) \in \mathcal{OD} \tag{7}$$

$$\sum_{s' \in \mathbb{Z}_{[0,s]}} \int_{\underline{t}_{s'}}^{\bar{t}_{s'}} \widehat{\mu}_{i,s'} dt \leq \sum_{s' \in \mathbb{Z}_{[0,s]}} \int_{\underline{t}_{s'}}^{\bar{t}_{s'}} \lambda_{i,s'}(t) dt \quad \forall s \in \mathcal{S}_i, i \in \mathcal{I} \tag{8}$$

Constraints (6) ensure that the cumulative boarding passenger flow volume prior to activity a and period s (i.e., $\sum_{a' \in \mathcal{A}_a^{\text{pre}}} \sum_{s' \in \mathbb{Z}_{[0,s]}} p_{a',s'}$) does not exceed the cumulative controlled passenger flow volume up to period s for OD pair (i,j), where $\mathcal{A}_a^{\text{pre}}$ denotes the set of passenger activities preceding a (see Section 3.1). The controlled passenger flow volume for OD pair (i,j) in period s' is given by $\int_{\underline{t}_{s'}}^{\bar{t}_{s'}} \widehat{\mu}_{i,s'} dt \cdot \gamma_{i,j,s'}$, where $\gamma_{i,j,s'}$ is the proportion of OD pair (i,j) in the total arrival passenger flow volume at station i during period s'. Constraints (7) require that the total boarding passenger flow volume across all periods (i.e., $\sum_{a \in \mathcal{A}_{i,j}^{OD}} \sum_{s \in \mathcal{S}_i} p_{a,s}$) matches the total controlled passenger flow volume for OD pair (i,j). Finally, constraints (8) ensure that the cumulative controlled passenger flow volume before period s (i.e., $\sum_{s' \in \mathbb{Z}_{[0,s]}} \int_{\underline{t}_{s'}}^{\bar{t}_{s'}} \widehat{\mu}_{i,s'} dt$) does not exceed the corresponding cumulative arrival passenger flow volume (i.e., $\sum_{s' \in \mathbb{Z}_{[0,s]}} \int_{\underline{t}_{s'}}^{\bar{t}_{s'}} \lambda_{i,s'}(t) dt$).

$$\begin{cases} W_{i,s} = 0 \quad s = 0 & (a) \\ W_{i,s-1} + \int_{\underline{t}_{s-1}}^{\bar{t}_{s-1}} \lambda_{i,s-1}(t) dt = \sum_{a \in \mathcal{A}_{i,s-1}^{D}} p_{a,s-1} + W_{i,s} \quad \forall s \in \mathcal{S}_i \backslash \{0\} & (b) \end{cases} \quad \forall i \in \mathcal{I} \tag{9}$$

Constraints (9) maintain passenger flow balance across periods by linking arriving, waiting, and boarding passenger flow volumes at station platforms. Constraints (9a) initialize the waiting passenger flow volume at zero for all stations at the start of the first period. Then, for station i, constraints (9b) state that the sum of the waiting passenger flow volume at the start of period $s-1$ (i.e., $W_{i,s-1}$) and the total arriving passenger flow volume during period $s-1$ (i.e., $\int_{\underline{t}_{s-1}}^{\bar{t}_{s-1}} \lambda_{i,s-1}(t) dt$) is equal to the sum of boarding passenger flow volumes in period $s-1$ (i.e., $\sum_{a \in \mathcal{A}_{i,s-1}^{D}} p_{a,s-1}$) and the waiting passenger flow volume $W_{i,s}$ at the start of the next period s.

$$\begin{cases} \sum_{a \in \mathcal{A}_e^O} p_{a,s} \leq \text{Cap}_{r(e)} \sum_{t \in \mathbb{Z}_{[\underline{t}_s, \bar{t}_s]}} \sum_{p \in \mathcal{P}} \alpha_{p,e,t} y_p & (a) \\ \sum_{a \in \mathcal{A}_e^{\text{pass}}} \sum_{s' \in \mathbb{Z}_{[0,s]}} p_{a,s'} \leq \text{Cap}_{r(e)} \sum_{t \in \mathbb{Z}_{[\underline{t}_s, \bar{t}_s]}} \sum_{p \in \mathcal{P}} \alpha_{p,e,t} y_p & (b) \\ \sum_{a \in \mathcal{A}_e^D} \sum_{s' \in \mathbb{Z}_{[0,s]}} p_{a,s'} \leq \text{Cap}_{r(e)} \sum_{t \in \mathbb{Z}_{[\underline{t}_s, \bar{t}_s]}} \sum_{p \in \mathcal{P}} \alpha_{p,e,t} y_p & (c) \end{cases} \quad \forall s \in \mathcal{S}, e \in \mathcal{E}^D \tag{10}$$

Constraints (10) link the rolling stock schedule to passenger flow assignments by imposing capacity limits at departure events. Specifically, if a departure event e is scheduled to start at any timestamp within the time

window $[\underline{t}_s, \bar{t}_s]$ of period s (i.e., $\sum_{t\in\mathbb{Z}_{[\underline{t}_s,\bar{t}_s]}}\sum_{p\in\mathcal{P}}\alpha_{p,e,t}y_p = 1$), then the total passenger flow volume from (a) activities $a \in \mathcal{A}_e^O$ starting from event e in period s, (b) activities $a \in \mathcal{A}_e^{\text{pass}}$ passing through event e and originating in earlier periods $s' \in \mathbb{Z}_{[\underline{t}_s,\bar{t}_s]} \cap \mathcal{S}_{e'}$, and (c) activities $a \in \mathcal{A}_e^D$ ending at event e and originating in earlier periods $s' \in \mathbb{Z}_{[\underline{t}_s,\bar{t}_s]} \cap \mathcal{S}_{e'}$ should not exceed the rolling stock capacity $\text{Cap}_{r(e)}$ for trip $r(e)$, as specified in constraints (10a)–(10c) based on the passenger activities in Fig. 5, respectively.

$$\sum_{a\in\mathcal{A}_{i,j}^{OD}}\sum_{s\in\mathcal{S}_i} p_{a,s} + p_{i,j}^{\text{re}} = P_{i,j}^{\text{in}} = \sum_{s\in\mathcal{S}_i}\int_{\underline{t}_s}^{\bar{t}_s} \lambda_{i,s}(t)\cdot\gamma_{i,j,s}dt \quad \forall (i,j)\in\mathcal{OD} \tag{11}$$

Constraints (11) ensure that, for each OD pair $(i,j) \in \mathcal{OD}$, the total boarding passenger flow volume over all periods (i.e., $\sum_{a\in\mathcal{A}_{i,j}^{OD}}\sum_{s\in\mathcal{S}_i} p_{a,s}$) plus the residual passenger flow volume $p_{i,j}^{\text{re}}$ should equal to the total passenger arrival flow volume $P_{i,j}^{\text{in}}$ for OD pair (i,j). The latter is computed as the sum of the integral of the passenger arrival rate function $\lambda_{i,s}(t)$ across its time window $[\underline{t}_s, \bar{t}_s]$ over all periods $s \in \mathcal{S}_i$.

Group III: Passenger delay calculation constraints

$$\widehat{w}_{i,s} = \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} q_{i,s}(t)dt = \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}}\int_{t_{i,s}^{QS}}^{t}[W_{i,s} + \lambda_{i,s}(\tau) - \widehat{\mu}_{i,s}]d\tau dt \quad \forall s\in\mathcal{S}_i, i\in\mathcal{I} \tag{12}$$

Based on Newell (2013) and Assumption 5, constraints (12) calculate the total passenger delay by integrating the queue length function $q_{i,s}(t)$ over the interval $[t_{i,s}^{QS}, t_{i,s}^{QE}]$, where $t_{i,s}^{QS}$ and $t_{i,s}^{QE}$ represent the start time and end time of the queuing process at station i in period s, respectively. Considering that $t_{i,s}^{QS}$ and $t_{i,s}^{QE}$ are generally unavailable as input data, the passenger delay can be derived in a closed form analytically under specific passenger flow rate functions (see Appendix C). For more complex passenger flow distributions, the passenger delay can be approximated by the period start time $\underline{t}_s$ and end time $\bar{t}_s$, yielding $\widehat{w}_{i,s} = \int_{\underline{t}_s}^{\bar{t}_s}\int_{\underline{t}_s}^{t}[W_{i,s} + \lambda_{i,s}(\tau) - \mu_{i,s}]d\tau dt$. Empirically, optimal passenger flow control rates $\mu_{i,s}$ tend to lie near the average passenger arrival rate $\overline{\lambda}_{i,s} = \frac{\int_{\underline{t}_s}^{\bar{t}_s}\lambda_{i,s}(t)dt}{\bar{t}_s - \underline{t}_s}$. To handle the nonlinearity of Eq. (12) in exact decomposition algorithms (e.g., column generation, Benders decomposition), we approximate it using tangent linear constraints at $\overline{\lambda}_{i,s}$, as given in Eq. (13).

$$\widehat{w}_{i,s}\left(\widehat{\mu}_{i,s}\right) \approx \widehat{w}_{i,s}\left(\bar{\lambda}_{i,s}\right) + \left.\frac{d\widehat{w}_{i,s}}{d\widehat{\mu}_{i,s}}\right|_{\widehat{\mu}_{i,s}=\bar{\lambda}_{i,s}}\left(\widehat{\mu}_{i,s} - \bar{\lambda}_{i,s}\right) \quad \forall s\in\mathcal{S}_i, i\in\mathcal{I} \tag{13}$$

Group IV: Ranges of the decision and auxiliary variables

All path selection decisions are binary variables, and all passenger flow management decisions are continuous variables, as formulated in the following constraints:

$$y_p \in \{0,1\} \quad \forall p\in\mathcal{P} \tag{14}$$

$$p_{a,s} \geq 0 \quad \forall s\in\mathcal{S}, a\in\mathcal{A}^{OD} \tag{15}$$

$$W_{i,s}, \widehat{w}_{i,s}, \widehat{\mu}_{i,s} \geq 0 \quad \forall s\in\mathcal{S}_i, i\in\mathcal{I} \tag{16}$$

$$p^{\text{re}}_{i,j} \geq 0 \quad \forall (i,j) \in \mathcal{OD}. \tag{17}$$

Using Eqs. (1)–(17), we formulate model M1 for the RSP-STPM. We observe that the complexity of model M1 is closely related to the scale of the involved events and activities, the number of periods, and the types of passenger flow control rate functions. To address this, we derive analytical expressions for passenger delays and total passenger arrival flow volumes under some specific passenger flow rate functions in Appendix C, and subsequently develop an exact and efficient BDCG algorithm to solve the path-based model M1, as detailed in the following section.

5. Solution algorithm

We propose an exact Benders decomposition algorithm for solving the RSP-STPM, with a combination of a column generation procedure to generate paths incorporating short-turn operations. The process begins with a pre-processing step that refines the rolling stock operational events and activities, yielding a more compact path-based model M1 (Section 5.1). Model M1 is then decomposed via Benders decomposition into a continuous sub-problem (BSP) M2 for passenger flow control and assignment, and an integer master problem (BMP) M3 for the rolling stock scheduling process. The BMP M3 is solved using a customized column generation algorithm by constructing event-activity paths that allow short-turns at candidate stations (Section 5.3). Subsequently, heuristic and branch-and-bound (B&B) algorithms are employed to obtain integer solutions and update lower/upper bounds (Sections 5.4 and 5.5). To reduce the number of branching variables in the B&B tree, we introduce a column enumeration and reduction method that facilitates finding exact feasible solutions and tightening lower/upper bounds (Section 5.6). In the column generation process, the single-path labeling is used to find the most promising column with the smallest reduced cost without recording paths on labels explicitly, which can reduce the memory use while preserving the optimality (Section 5.3.2). In the column enumeration process, to generate all promising columns that may appear in the optimal solution based on the reduced cost criteria, the multi-path labeling algorithm is used, which records partial paths on labels explicitly (Section 5.6.1). Finally, several acceleration techniques are incorporated to enhance computational performance (Section 5.7).

5.1. Pre-processing events and activities

We note that the computational complexity of model M1 is highly dependent on the start time windows $\mathcal{T}_e$ of events e, particularly when short-turn operations are involved. To mitigate this, we propose a pre-processing procedure that shrinks the time windows of rolling stock operational events, thereby producing a more compact formulation of model M1. Since all trips in the same direction share identical physical routes, a departure sequence can be pre-generated for each direction. Specifically, let r^{up}_q denote the q-th trip departing from the terminal station in the upstream direction. The earliest and latest start times of its first event $e_0(r^{\text{up}}_q)$ are computed using Eqs. (18)–(19), where $|\mathcal{R}^{\text{up}}|$ is the total number of trips in the upstream

direction, and $h^{\min}_{e_0(r^{\text{up}}_{q'-1}),e_0(r^{\text{up}}_{q'})}$ and $h^{\max}_{e_0(r^{\text{up}}_{q'-1}),e_0(r^{\text{up}}_{q'})}$ are the minimum and maximum headways between the first events of adjacent trips. The same method applies to trips in the downstream direction to determine $\underline{T}_{e_0(r^{\text{down}}_q)}$ and $\overline{T}_{e_0(r^{\text{down}}_q)}$.

$$\underline{T}_{e_0(r^{\text{up}}_q)} = \begin{cases} 0 \quad q=0 \\ \max\left\{ \sum_{q'=1}^{q} h^{\min}_{e_0(r^{\text{up}}_{q'-1}),e_0(r^{\text{up}}_{q'})}, |\mathcal{T}| - \sum_{q'=q+1}^{|\mathcal{R}^{\text{up}}|-1} h^{\max}_{e_0(r^{\text{up}}_{q'-1}),e_0(r^{\text{up}}_{q'})} \right\} \end{cases} \quad \forall q \in \{1,...,|\mathcal{R}^{\text{up}}|-1\} \tag{18}$$

$$\overline{T}_{e_0(r^{\text{up}}_q)} = \begin{cases} \min\left\{ \sum_{q'=1}^{q} h^{\max}_{e_0(r^{\text{up}}_{q'-1}),e_0(r^{\text{up}}_{q'})}, |\mathcal{T}| - \sum_{q'=q+1}^{|\mathcal{R}^{\text{up}}|-1} h^{\min}_{e_0(r^{\text{up}}_{q'-1}),e_0(r^{\text{up}}_{q'})} \right\} \\ |\mathcal{T}| \quad q=|\mathcal{R}^{\text{up}}|-1 \end{cases} \quad \forall q \in \{0,...,|\mathcal{R}^{\text{up}}|-2\} \tag{19}$$

Using the start time windows of the first event $e_0(r)$ for each trip r in both directions, the earliest and latest start times for any subsequent l-th event $e_l(r)$ of trip r can be calculated via Eqs. (20)–(21), where $T^{\min}_{e_l(r),e_{l+1}(r)}$ and $T^{\max}_{e_l(r),e_{l+1}(r)}$ are the minimum and maximum running/dwelling times between adjacent events $e_l(r)$ and $e_{l+1}(r)$.

$$\underline{T}_{e_l(r)} = \underline{T}_{e_0(r)} + \sum_{l'=0}^{l-1} T^{\min}_{e_l(r),e_{l+1}(r)} \quad \forall l \in \{1,\cdots,|\mathcal{E}_r|-1\}, r \in \mathcal{R} \tag{20}$$

$$\overline{T}_{e_l(r)} = \overline{T}_{e_0(r)} + \sum_{l'=0}^{l-1} T^{\max}_{e_l(r),e_{l+1}(r)} \quad \forall l \in \{1,\cdots,|\mathcal{E}_r|-1\}, r \in \mathcal{R} \tag{21}$$

Finally, based on the computed start time windows $\mathcal{T}_e$, we pre-process the turn-around activity set $\mathcal{A}^C$ by retaining only those event pairs in opposite directions at the same station whose time windows overlap sufficiently, i.e., $\mathcal{A}^C = \left\{ (e,e') \,\middle|\, e \in \mathcal{E}_r, e' \in \mathcal{E}_{r'}, e_0(r) \neq e_0(r'), i(e) = i(e'), \underline{T}_{e'} - \overline{T}_e \geq T^{\min}_{e,e'} \right\}$, where $T^{\min}_{e,e'}$ is the minimum turn-around time. Moreover, the presence of a feasible turn-around activity between event e and event e' allows further refinement of their start time windows as $\underline{T}_e \leftarrow \max\left\{\underline{T}_{e'} - T^{\max}_{e,e'}, \underline{T}_e\right\}$ and $\overline{T}_{e'} \leftarrow \min\left\{\overline{T}_{e'}, \overline{T}_e + T^{\max}_{e,e'}\right\}$, where $T^{\max}_{e,e'}$ is the maximum turn-around time.

5.2. Benders decomposition

From model M1, we observe that it contains numerous integer column selection variables and continuous passenger flow assignment variables, making it difficult to define a "column" and directly apply a traditional column generation algorithm. Notably, constraints (2)–(5) involve only the column selection variables y_p. Moreover, once y_p is fixed, the passenger flow management problem becomes a linear program (LP) with only continuous variables and linear constraints, which is far easier to solve. This motivates the use of Benders decomposition to first decompose model M1 into simpler models. Specifically, let $\overline{\mathbf{y}} = (\overline{y}_1, \cdots, \overline{y}_p, \cdots, \overline{y}_{|\mathcal{P}|})^T$ with $p \in \mathcal{P}$ be a path selection solution satisfying constraints (2)–(5) and (14). For any given path selection vector $\overline{\mathbf{y}}$, the LP relaxation of model M1 reduces to the BSP M2 in Eq. (22), which includes only the passenger-related objective $v(\overline{\mathbf{y}})$, decision variables, and constraints.

$$\text{M2:}\begin{cases}\min \quad v(\overline{\mathbf{y}})=\sum_{i\in\mathcal{I}}\sum_{s\in\mathcal{S}_i}\left(c^D_{i,s}\widehat{w}_{i,s}+c^C_{i,s}\int_{\underline{t}_s}^{\bar{t}_s}\hat{\mu}_{i,s}dt\right)+\sum_{(i,j)\in\mathcal{OD}}c^{\text{re}}_{i,j}p^{\text{re}}_{i,j} & (a)\\ \text{s.t.Constraints (6)–(9), (11)–(12), (13)–(17)} & (b)\\ \sum_{a\in\mathcal{A}^O_e}p_{a,s}\le \text{Cap}_{r(e)}\sum_{t\in\mathbb{Z}_{[\underline{t}_s,\bar{t}_s]}}\sum_{p\in\mathcal{P}}\alpha_{p,e,t}\overline{y}_p \quad \forall s\in\mathcal{S},e\in\mathcal{E}^D & (\delta_{e,s})(c)\\ \sum_{a\in\mathcal{A}^{\text{pass}}_e}\sum_{s'\in\mathbb{Z}_{[0,s]}}p_{a,s'}\le \text{Cap}_{r(e)}\sum_{t\in\mathbb{Z}_{[\underline{t}_s,\bar{t}_s]}}\sum_{p\in\mathcal{P}}\alpha_{p,e,t}\overline{y}_p \quad \forall s\in\mathcal{S},e\in\mathcal{E}^D & (\xi_{e,s})(d)\\ \sum_{a\in\mathcal{A}^D_e}\sum_{s'\in\mathbb{Z}_{[0,s]}}p_{a,s'}\le \text{Cap}_{r(e)}\sum_{t\in\mathbb{Z}_{[\underline{t}_s,\bar{t}_s]}}\sum_{p\in\mathcal{P}}\alpha_{p,e,t}\overline{y}_p \quad \forall s\in\mathcal{S},e\in\mathcal{E}^D & (\varphi_{e,s})(e)\end{cases} \tag{22}$$

The BSP M2 is always feasible due to the presence of the residual passenger flow variables $p^{\text{re}}_{i,j}$, which means unserved demand is simply penalized in the objective function. In this LP, the constant terms of constraints (6)–(8) and (9a) equal to zero, so their dual variables do not contribute to the Benders cuts. Therefore, we consider only the dual variables from the active constraints (9b), (22c)–(22e), and (11)–(12). Specifically, for iteration o of the Benders decomposition algorithm, we denote the corresponding dual variables as $\boldsymbol{\theta}_o=\{\theta_{o,i,s}\mid s\in\mathcal{S}_i\setminus\{0\},i\in\mathcal{I}\}$, $\boldsymbol{\delta}_o=\{\delta_{o,e,s}\ge 0\mid s\in\mathcal{S},e\in\mathcal{E}^D\}$, $\boldsymbol{\xi}_o=\{\xi_{o,e,s}\ge 0\mid s\in\mathcal{S},e\in\mathcal{E}^D\}$, $\boldsymbol{\varphi}_o=\{\varphi_{o,e,s}\ge 0\mid s\in\mathcal{S},e\in\mathcal{E}^D\}$, $\boldsymbol{\zeta}_o=\{\zeta_{o,i,j}\mid (i,j)\in\mathcal{OD}\}$, and $\boldsymbol{\kappa}_o=\{\kappa_{o,i,s}\mid s\in\mathcal{S}_i,i\in\mathcal{I}\}$. For convenience, we define parameters $\rho_{p,o}$ and ψ_o as in Eqs. (23)–(24).

$$\rho_{p,o}=\sum_{e\in\mathcal{E}^D}\sum_{s\in\mathcal{S}}\sum_{t\in\mathbb{Z}_{[\underline{t}_s,\bar{t}_s]}}\text{Cap}_{r(e)}\alpha_{p,e,t}\left(\delta_{o,e,s}+\xi_{o,e,s}+\varphi_{o,e,s}\right) \tag{23}$$

$$\psi_o=-\sum_{i\in\mathcal{I}}\sum_{s\in\mathcal{S}_i\setminus\{0\}}\left(\int_{\underline{t}_{s-1}}^{\bar{t}_{s-1}}\lambda_{i,s}(\tau)\,d\tau\theta_{o,i,s}\right)-\sum_{(i,j)\in\mathcal{OD}}P^{\text{in}}_{i,j}\zeta_{o,i,j}+\sum_{i\in\mathcal{I}}\sum_{s\in\mathcal{S}_i}\left(\widehat{w}_{i,s}(\bar{\lambda}_{i,s})-\frac{d\widehat{w}_{i,s}}{d\widehat{\mu}_{i,s}}\Big|_{\widehat{\mu}_{i,s}=\bar{\lambda}_{i,s}}\bar{\lambda}_{i,s}\right)\kappa_{o,i,s} \tag{24}$$

Finally, the LP relaxation of model M1 is reformulated as the relaxed BMP M3 in Eq. (25), with Benders optimality cuts in Eq. (25c), where $\mathcal{O}_\Delta$ denotes the set of enumerated extreme points of the dual solutions and η is an auxiliary variable representing total passenger-related costs.

$$\text{M3:}\begin{cases}\min \quad \sum_{p\in\mathcal{P}}C^R_p y_p+\eta & (a)\\ \text{s.t. Constraints (2)–(5)} & (b)\\ \eta\ge\psi_o-\sum_{p\in\mathcal{P}}\rho_{p,o}y_p \quad \forall o\in\mathcal{O}_\Delta & (c)\\ 0\le y_p\le 1 \quad \forall p\in\mathcal{P} & (d)\\ \eta\ge 0 & (e)\end{cases} \tag{25}$$

5.3. Column generation

In the BMP M3, the full enumeration of a large number of feasible paths is impractical, particularly when short-turn operations are allowed at candidate stations. Column generation is a well-established method for constructing feasible paths and obtaining lower bounds in urban rail operation problems (Freling et al.

2005, Cacchiani et al. 2010b, 2013, Pan et al. 2023, Wang et al. 2024). Accordingly, we solve BMP M3 within a column generation framework. We first formulate the pricing sub-problem for the relaxed BMP M3 (Section 5.3.1) and develop a customized single-path labeling algorithm (Section 5.3.2). Initially, a set $\overline{\mathcal{P}}$ of artificial columns with large costs is added to the relaxed BMP M3 to ensure feasibility of the restricted master problem. After solving the relaxed BMP M3, the pricing sub-problem, which is implemented via the single-path labeling algorithm, generates new paths (columns) using the dual solutions of the constraints in the relaxed master problem. The process terminates once no column with a negative reduced cost is identified.

5.3.1. Pricing sub-problem To formulate the pricing sub-problem of the relaxed BMP M3, we denote the dual variables $\boldsymbol{\pi}^1 = \left\{\pi^1_g\right\}_{g\in\mathcal{G}}$, $\boldsymbol{\pi}^2 = \{\pi^2_r\}_{r\in\mathcal{R}}$, $\boldsymbol{\pi}^3 = \{\pi^3_e\}_{e\in\mathcal{E}}$, $\boldsymbol{\pi}^4 = \left\{\pi^4_{e,e',t}\right\}_{(e,e')\in\mathcal{V},t\in\mathcal{T}}$, $\boldsymbol{\pi}^5 = \{\pi^5_o\}_{o\in\mathcal{O}_\Delta}$ associated with constraints (2)–(5) and (25c), respectively. The reduced cost of each path p is then computed using Eq. (26), where the binary parameter $\alpha_{p,e,t}$ equals to 1 if event e is included in path p with start time t, and binary parameter $\beta_{p,r}$ equals to 1 if trip r is included in path p. Thus, the pricing sub-problem for the relaxed BMP M3 becomes a constrained shortest path problem that finds the path $p \in \mathcal{P}$ with the minimum reduced cost (i.e., $\min_{p\in\mathcal{P}} \overline{C^R_p}$), thereby improving the objective value of the relaxed BMP M3. Unlike standard pricing sub problems in column generation procedures (Cacchiani et al. 2010b, 2013, Wang et al. 2024), we incorporate not only flow balance constraints but also additional constraints to account for short-turn operations.

$$\begin{aligned}\overline{C^R_p} = C^R_p - \pi^1_{g(p)} - \sum_{r\in\mathcal{R}} \beta_{p,r}\pi^2_r - \sum_{e\in\mathcal{E}}\sum_{t\in\mathcal{T}_e} \alpha_{p,e,t}\pi^3_e - \sum_{t\in\mathcal{T}}\sum_{(e,e')\in\mathcal{V}} \alpha_{p,e,t}\pi^4_{e',e,t} \\ - \sum_{t\in\mathcal{T}}\sum_{(e,e')\in\mathcal{V}}\sum_{t'=\widetilde{t^H_{e,e'}}}^{t} \alpha_{p,e',t'}\pi^4_{e,e',t'} - \sum_{o\in\mathcal{O}_\Delta} \rho_{p,o}\pi^5_o, \quad \widetilde{t^H_{e,e'}} = \max\{t - h^{\min}_{e,e'} + 1, 0\}\end{aligned} \tag{26}$$

5.3.2. Single-path labeling algorithm We observe that multiple paths may have the same minimum negative reduced cost in one column generation iteration. However, only one of these optimal columns needs to be added to the column pool. To achieve this while reducing memory usage and preserving optimality, we design a single-path labeling algorithm to find one of the most promising columns (i.e., a time-dependent event-activity path with short-turn operations) without explicitly recording paths on labels. Particularly, the labeling algorithm can be implemented in parallel for different rolling stock groups, ensuring both scalability and memory efficiency. First, we construct a time-dependent event-activity network $\mathcal{G}^T = \{\mathcal{V}^T, \mathcal{A}^T\}$ with the time-dependent event set $\mathcal{V}^T = \{(e,t) \mid t \in \mathcal{T}_e, e \in \mathcal{E}\}$ and the time-dependent activity set $\mathcal{A}^T = \{(e,e',t,t') \mid t \in \mathcal{T}_e, t' \in \mathcal{T}_{e'}, (e,e') \in \mathcal{A}\}$. Unlike in standard rolling stock scheduling problems, cyclic sub-paths can arise when short-turn operations are allowed at candidate stations. For instance, in Fig. 6(a), trips r_1 and r_2 offer two possible short-turn schemes at a station via activities (D_1, A_2) and (D_2, A_1). These create a sub-loop $A_1 \to D_1 \to A_2 \to D_2 \to A_1$, which violates the directed acyclic graph (DAG) requirement. To

address this, we eliminate the turn-around activities (D_1, A_2) and (D_2, A_1) between the last departure and first arrival events, and instead introduce new turn-around activities (A_1, D_2) and (A_2, D_1) between the last arrival and first departure events. Their activity times are set as $T_{A_1,D_2} = T_{A_1,D_1} + T_{D_1,A_2} + T_{A_2,D_2}$ and $T_{A_2,D_1} = T_{A_2,D_2} + T_{D_2,A_1} + T_{A_1,D_1}$, respectively, as illustrated in Fig. 6(b). This transformation prevents cyclic sub-paths from short-turn operations during the topological sort in the initialization stage.

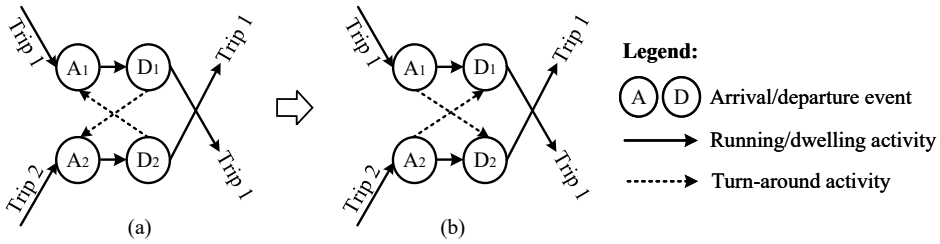


Figure 6 Illustration of turn-around activity transformation

In the forward dynamic label updating step, we first check two branching rules (see Section 5.6). That is, (1) if event e (or e') is prohibited from starting at timestamp t (or t'), i.e., $\overline{x}_{e,t} = 0$ or $\overline{x}_{e',t'} = 0$, or (2) if events e and e' cannot be linked by rolling stock units in group g, then the updates for labels with respect to vertices (e, t) and (e', t') are skipped. If neither branching rule applies, the label cost (i.e., reduced cost) is updated using the cost $\overline{c}_{e,e',t,t'}$ associated with the time-dependent activity arc (e, e', t, t'), as defined in Eqs. (27)–(28). Here, $r(e)$ and $r(e')$ denote the trips corresponding to events e and e', while $s(t')$ indicates the period of timestamp t'. Since only the single minimum reduced cost path is needed in each column generation iteration, we maintain only two matrices: the label cost matrix $RC_{e,t}$ and the label predecessor matrix $Pred_{e,t}$. This avoids tracking a potentially large number of paths, significantly reducing memory usage. After performing a topological sort of the time-dependent event-activity network, a single pass over all labels suffices to complete the updates. Lastly, the shortest path p^* with the minimum label cost $\overline{C^R_{p^*}}$, is then backtracked from the depot sink $(e_{\text{des}}(g), t_{|\mathcal{T}|})$ to the depot source $(e_{\text{ori}}(g), t_0)$ for rolling stock group g. If $\overline{C^R_{p^*}} < 0$, the path p^* is added to the column pool $\overline{\mathcal{P}}$ for subsequent iterations; otherwise, if $\overline{C^R_{p^*}} \geq 0$, the column generation process terminates. The full pseudocode of the single-path labeling algorithm is provided in Algorithm 1 of Appendix D.

$$\overline{c}_{e,e',t,t'} = \begin{cases} c^1_g - \pi^1_g - \pi^2_{r(e')} + \widetilde{c}_{e',t'}, & \text{if } (e,e',t,t') = (e_0(g), e', t_0, t'), t' \in \mathcal{T}_{e'}, e' = e_0(r'), r' \in \mathcal{R} \\ c^2_{g,e,e'}(t'-t) + \widetilde{c}_{e',t'}, & \text{if } t \in \mathcal{T}_e, t' \in \mathcal{T}_{e'}, (e,e') \in \mathcal{A}^R \cup \mathcal{A}^D \\ c^2_{g,e,e'}(t'-t) - \pi^2_{r(e')} + \widetilde{c}_{e',t'}, & \text{if } t \in \mathcal{T}_e, t' \in \mathcal{T}_{e'}, (e,e') \in \mathcal{A}^T \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

$$\widetilde{c}_{e',t'} = -\pi^3_e - \sum_{(e'',e') \in \mathcal{V}} \pi^4_{e'',e',t'} - \sum_{(e',e'') \in \mathcal{V}} \sum_{\tau = t^{\widetilde{H}}_{e,e'}}^{t'} \pi^4_{e',e'',\tau} - \sum_{o \in \mathcal{O}_\Delta} \rho_o \pi^5_o \tag{28}$$

5.4. Heuristic algorithm

In early iterations, the standard BDCG algorithm often exhibits large relative gaps due to the poor quality of feasible solutions and the corresponding upper bounds. Moreover, effective column enumeration and reduction (see Section 5.6) requires high-quality feasible solutions. To address these issues, we develop a heuristic algorithm that derives feasible integer solutions with practical finite upper bounds from the fractional solutions obtained by the BDCG algorithm. Specifically, we denote by $\overline{\mathbf{y}}$ the optimal path selection (fractional) solution vectors to the LP relaxation of model M1. We sort the path columns $p \in \overline{\mathcal{P}}$ in descending order of the fractional solution value y_p. We then select and fix the column based on this descending order. Particularly, the selected column should satisfy constraints (2)–(5) with other previously fixed columns. Simultaneously, we fix the occupied space-time resources and the associated headway resources of the selected column, which cannot be used by any subsequent columns. Next, if some trips remain uncovered by columns $p \in \overline{\mathcal{P}}$, we generate additional columns by applying the single-path labeling algorithm (see Section 5.3.2) with newly assigned available rolling stock units. Finally, after selecting and generating columns that cover all trips, we solve BSP M2 and update passenger-related decision variables based on the newly selected and generated rolling stock columns.

5.5. Branch-and-bound (B&B) algorithm

To enhance lower bounds and identify feasible solutions, we apply a B&B algorithm after performing Benders decomposition. We define two types of branching variables: (1) $\overline{x}_{e,t} = \sum_{p \in \overline{\mathcal{P}}} \overline{y}_p \alpha_{p,e,t}$, indicating whether event $e \in \mathcal{E}$ starts at time $t \in \mathcal{T}_e$, and (2) $\overline{z}_{g,e,e'} = \sum_{p \in \overline{\mathcal{P}}_g} \overline{y}_p \beta_{p,e} \beta_{p,e'}$, indicating whether rolling stock units in group $g \in \mathcal{G}$ performs the turn-around operation $(e, e') \in \mathcal{A}^C$. Here, $\overline{y}_p$ ($p \in \overline{\mathcal{P}}$) are the fractional optimal solutions of the relaxed model M1. At each branching step, we select the variable whose value is closest to 0.5, and create two child nodes: n_1 with the variable fixed to 0, and n_2 with it fixed to 1. During column generation and enumeration with the labeling algorithm, label updates involving branched event start times or event connections are skipped directly according to the branching rules. This preserves the structure of BMP M3 without adding extra branching constraints. We then update the lower bound as the minimum node lower bound in the current layer, and the upper bound whenever a new feasible solution is found. Finally, nodes are pruned if they are infeasible, integral, or have lower bounds exceeding the current upper bound.

5.6. Column enumeration and reduction (CER)

Preliminary experiments revealed that the B&B algorithm alone yields limited improvements in lower and upper bounds. The main cause is the large number of branching variables both for event start times $\overline{x}_{e,t}$ across all events and for event connections $\overline{z}_{g,e,e'}$ with (short) turn-around operations. This leads to many branch layers, high memory usage, and minimal reductions in relative gaps (Baldacci et al. 2008, Yang 2025). To address this, after solving the LP relaxation of BMP M3, we enumerate promising columns with small

reduced costs that are likely to appear in optimal solutions, and remove non-promising columns using a Lagrangian relaxation bound. This approach can produce tighter bounds and feasible (sometimes optimal) solutions. Specifically, let z_{LB} be the objective value of BMP M3 at a node after Benders decomposition and let z_{UB} be the incumbent objective value from the heuristic algorithm, serving as lower and upper bounds of integer model M1, respectively. By applying Proposition 1 (Desaulniers et al. 2006, Baldacci et al. 2008), if the gap $z_{UB} - z_{LB}$ is below a given threshold, all columns with reduced costs no greater than this gap can be enumerated. For clarity, we first describe the column enumeration using the multi-path labeling algorithm in Section 5.6.1.

PROPOSITION 1. *Given lower bound z_{LB} and upper bound z_{UB} for the optimal value of model M1, any column p with reduced cost exceeding $z_{UB} - z_{LB}$ can be eliminated from the column pool.*

5.6.1. Multi-path labeling algorithm for enumerating promising columns Compared to the labeling algorithm in the column generation process, which only generates one promising column with the minimum negative reduced cost, it is necessary to generate multiple (or all) promising columns with reduced costs less than $z_{UB} - z_{LB}$ in the column enumeration process. Thus, by explicitly recording paths on labels, we design a multi-path labeling algorithm to enumerate all possible paths that may appear in the optimal solution based on the reduced cost condition. Specifically, we first construct the time-dependent event-activity network $\mathcal{G}^T$. For each rolling stock group g, we initialize a label $l_{e_{\text{ori}}(g),t_0}$ at the depot source $(e_{\text{ori}}(g), t_0)$ with reduced cost $RC_{l_{e_{\text{ori}}(g),t_0}} = c_g^1 - \pi_g^1$, and push it into the deque list $LabelDeque$. In the label extension and update step, unlike the single-path labeling algorithm, we extend each label $l_{e,t}$ to all possible successor events $e' \in \mathcal{E}_e^+$ with their corresponding start times $t' \in \mathcal{T}_{e'}$, and then generate a new label $l_{e',t'}$. After checking the branching rules on event start times and connections, any feasible new label $l_{e',t'}$ is pushed back to $LabelDeque$, ensuring that no promising columns are missed. When all labels in $LabelDeque$ reach the depot sink $(e_{\text{des}}(g), t_{|\mathcal{T}|})$ and can no longer be extended, we filter those corresponding to feasible paths whose label costs (i.e., reduced costs) are less than the gap $z_{UB} - z_{LB}$. These paths are then used to construct the enumerated column pool $\overline{\mathcal{P}}^E$. The full pseudocode of the multi-path labeling algorithm is given in Algorithm 2 of Appendix D.

5.6.2. Lagrangian relaxation To further exploit the enumerated column pool and reduce the complexity of solving the integer model M1, we aim to identify and remove less promising columns using the Lagrangian relaxation bound (Balas and Carrera 1996, Holmberg and Yuan 2000, Yang 2025). Specifically, for an enumerated column p fixed with $y_p = 1$, if the Lagrangian relaxation bound exceeds the upper bound z_{UB}, column p can be removed from the enumerated column pool $\overline{\mathcal{P}}^E$ safely. For this purpose, we first define the Lagrangian relaxation function in Eq. (29) by dualizing the constraints of BMP M3 with respect to the enumerated column pool $\overline{\mathcal{P}}^E$.

$$
\begin{aligned}
&\mathcal{L}(\mathbf{y},\eta,\boldsymbol{\pi}^1,\cdots,\boldsymbol{\pi}^5)\\
&=\sum_{p\in\overline{\mathcal{P}}^E}\left(C_p^R-\pi^1_{g(p)}+\sum_{r\in\mathcal{R}}\beta_{p,r}\pi^2_r+\sum_{e\in\mathcal{E}}\sum_{t\in\mathcal{T}_e}\alpha_{p,e,t}\pi^3_e+\sum_{(e,e')\in\mathcal{V}}\sum_{t\in\mathcal{T}^v_{e,e'}}\left(\alpha_{p,e,t}\pi^4_{e',e,t}+\sum_{t'=t}^{\widetilde{H}_{e,e'}}\alpha_{p,e',t'}\pi^4_{e,e',t'}\right)-\sum_{o\in\mathcal{O}_\Delta}\rho_{p,o}\pi^5_o\right)y_p\\
&\quad-\sum_{g\in\mathcal{G}}|\mathcal{K}_g|\pi^1_g-\sum_{r\in\mathcal{R}}\pi^2_r-\sum_{e\in\mathcal{E}}\pi^3_e-\sum_{(e,e')\in\mathcal{V}}\sum_{t\in\mathcal{T}^v_{e,e'}}\pi^4_{e,e',t}+\sum_{o\in\mathcal{O}_\Delta}(\psi_o-\eta)\,\pi^5_o+\eta\\
&=\sum_{p\in\overline{\mathcal{P}}^E}\overline{C_p^R}y_p-\sum_{g\in\mathcal{G}}|\mathcal{K}_g|\pi^1_g-\sum_{r\in\mathcal{R}}\pi^2_r-\sum_{e\in\mathcal{E}}\pi^3_e-\sum_{(e,e')\in\mathcal{V}}\sum_{t\in\mathcal{T}^v_{e,e'}}\pi^4_{e,e',t}+\sum_{o\in\mathcal{O}_\Delta}(\psi_o-\eta)\,\pi^5_o+\eta
\end{aligned}
\tag{29}
$$

It is theoretically possible to validate the Lagrangian relaxation bound by solving $\min \mathcal{L}(\mathbf{y},\eta,\boldsymbol{\pi}^1,\ldots,\boldsymbol{\pi}^5)\big|_{y_p=1}$ for each column p, but it is computationally impractical to do it for all enumerated columns. Instead, we estimate the bound using a feasible set of dual values $\widehat{\boldsymbol{\pi}}^1,\ldots,\widehat{\boldsymbol{\pi}}^5$. These values are chosen to ensure non-negative reduced costs $\overline{C_p^R}$ for all columns $p\in\overline{\mathcal{P}}^E$, by solving the auxiliary model M4 in Eq. (30) which maximizes the total Lagrangian relaxation cost across all enumerated columns.

$$
\text{M4:}\begin{cases}
\max\limits_{\boldsymbol{\pi}_1,\cdots,\boldsymbol{\pi}_5}\sum\limits_{p\in\overline{\mathcal{P}}^E}\min\limits_{\{y_{p'}\}_{p'\in\overline{\mathcal{P}}^E\setminus\{p\}}}\mathcal{L}(\mathbf{y},\eta,\boldsymbol{\pi}_1,\cdots,\boldsymbol{\pi}_5)\big|_{y_p=1} & (a)\\
\quad\text{s.t. }\overline{C_p^R}\ge 0\quad\forall p\in\overline{\mathcal{P}}^E & (b)\\
\quad\pi^1_g,\pi^3_e,\pi^4_{e,e',t},\pi^5_o\in[-z_{UB},0]\quad\forall g\in\mathcal{G},e\in\mathcal{E},r\in\mathcal{R},(e,e')\in\mathcal{V},t\in\mathcal{T}^{\mathcal{V}}_{e,e'},o\in\mathcal{O}_\Delta & (c)\\
\quad\pi^2_r\in[-z_{UB},z_{UB}]\quad\forall r\in\mathcal{R} & (d)
\end{cases}
\tag{30}
$$

For simplicity, the objective function in Eq. (30a) can be reformulated by estimating its lower bound $Q(\overline{C_p^R},\boldsymbol{\pi}^1,\ldots,\boldsymbol{\pi}^5)$ in Eq. (31). Constraints (30b) enforce non-negative reduced costs, while (30c)–(30d) define the ranges of the dual variables. Because model M4 has more constraints (i.e., associated with enumerated columns) but fewer decision variables, it is more efficient to solve its primal model and extract the dual values of the constraints as the solution.

$$
\begin{aligned}
&\min\quad\mathcal{L}(\mathbf{y},\eta,\boldsymbol{\pi}^1,\cdots,\boldsymbol{\pi}^5)\big|_{y_p=1}\\
&=\sum_{p'\in\mathcal{P}^E\setminus\{p\}}\overline{C_p^R}y_{p'}+\overline{C_p^R}-\sum_{g\in\mathcal{G}}|\mathcal{K}_g|\pi^1_g-\sum_{r\in\mathcal{R}}\pi^2_r-\sum_{e\in\mathcal{E}}\pi^3_e-\sum_{(e,e')\in\mathcal{V}}\sum_{t\in\mathcal{T}^v_{e,e'}}\pi^4_{e,e',t}+\sum_{o\in\mathcal{O}_\Delta}(\psi_o-\eta)\,\pi^5_o+\eta\\
&\ge\overline{C_p^R}-\sum_{g\in\mathcal{G}}|\mathcal{K}_g|\pi^1_g-\sum_{r\in\mathcal{R}}\pi^2_r-\sum_{e\in\mathcal{E}}\pi^3_e-\sum_{(e,e')\in\mathcal{V}}\sum_{t\in\mathcal{T}^v_{e,e'}}\pi^4_{e,e',t}+\sum_{o\in\mathcal{O}_\Delta}\psi_o\pi^5_o=Q\left(\overline{C_p^R},\boldsymbol{\pi}^1,\cdots,\boldsymbol{\pi}^5\right)
\end{aligned}
\tag{31}
$$

With the obtained dual solution $\widehat{\boldsymbol{\pi}}^1,\ldots,\widehat{\boldsymbol{\pi}}^5$, we can safely remove any column p for which $Q(\overline{C_p^R},\widehat{\boldsymbol{\pi}}^1,\ldots,\widehat{\boldsymbol{\pi}}^5)>z_{UB}$ (i.e., $\min \mathcal{L}(\mathbf{y},\eta,\widehat{\boldsymbol{\pi}}^1,\ldots,\widehat{\boldsymbol{\pi}}^5)\Big|_{y_p=1}>z_{UB}$), and then solve the integer BMP M3 and BSP M4 again after updating the enumerated column pool. The resulting objective values of these models can further refine the lower and upper bounds.

In our numerical experiments, we observe that for small-scale and medium-scale instances, it is possible to enumerate all effective columns with reduced costs below the gap $z_{UB}-z_{LB}$ without exceeding memory limits, and to achieve the exact optimum for integer BMP M3. In some large-scale cases, it remains challenging to reach the exact optimality. To address this, we solve the integer BMP M3 with a maximum computing time limit (e.g., 1 minute), then update z_{LB} and z_{UB} from the resulting lower and upper bounds, which still yields improvements over the traditional B&B approach for finding integer solutions.

5.7. Acceleration techniques

We observe that the lower bounds improve only slightly in the early iterations of the Benders decomposition process, while each iteration remains time-consuming due to the large number of redundant constraints and columns tackled in the column generation. To address these issues, we enhance the standard BDCG framework with three acceleration techniques: (1) Dynamic constraint generation: In the Benders decomposition process, we handle the large number of inactive constraints (2) and (5) dynamically. After each iteration, we identify the violated constraints (2) and (5) in the obtained solution and add them to the active constraint set of the relaxed BMP M3. All such constraints are still retained in the column enumeration and reduction stage to ensure primal feasibility and a complete dual solution space. (2) Column and cut inheritance: To avoid regenerating the same columns and Benders cuts at new B&B nodes, the child nodes inherit all columns with positive selection values ($y_p > 0, p \in \overline{\mathcal{P}}$) and all generated Benders cuts from their parent node. (3) Parallel computing: Labeling algorithms for different rolling stock groups, as well as the solution of B&B nodes within the same branching layer, are executed independently across CPU threads in parallel.

6. Numerical experiments

In this section, we present a series of numerical experiments to evaluate the performance of the proposed approach, including both small-scale and real-world case studies based on the Changping Line of the Beijing urban rail system. All algorithms were implemented in C++ and use CPLEX 22.10. They were executed on a computer equipped with an Intel(R) Core(TM) i7-13700 CPU (2.50 GHz) and 32 GB of RAM. For convenience, the small-scale instances are solved using both the CPLEX solver and the proposed algorithm to verify the correctness and efficiency of our approach (see Appendix E). The large-scale instances, however, are solved exclusively with the proposed algorithm, as it is computationally challenging to directly tackle the primal model M5 (see Appendix B) with general-purpose solvers.

6.1. Experimental settings

We investigate a real-world case to evaluate computational performance and derive managerial insights based on the Changping line of the Beijing urban rail system (a bidirectional line with 12 stations). Fig. 7 illustrates the station layouts and physical segments, along with their minimum running times. The planning horizon spans 19 hours, from 5:00 to 24:00. Four rolling stock groups are considered in the real-world case study, and rolling stock units are allowed to depart from or return to depots at any time within the planning horizon. Specifically, rolling stock units in group 1 start from and end at the XEQ depot, rolling stock units in group 2 start from the XEQ depot and end at the CPXSK depot, rolling stock units in group 3 start from the CPXSK depot and end at the XEQ depot, rolling stock units in group 4 start from the CPXSK depot and end at the CPXSK depot. Each rolling stock unit has a maximum capacity of 1,440 passengers. Short-turn operations are permitted at stations ZXZ, GHC, SH, SUP, NS, BSW, CPDG, and CP during morning peak hours (5:00–10:00) and evening peak hours (16:00–24:00). To maintain service quality and computational

efficiency, half of trips are allowed to perform candidate short-turn operations, reflecting actual operations in urban rail systems. The minimum and maximum turn-around times for all rolling stock units are set to 3 minutes and 30 minutes, respectively. Based on real-world operations of the Beijing urban rail system, the fixed assignment cost is set to ¥1,000 per rolling stock unit, accounting for amortized depreciation and routine maintenance expenses. The costs associated with running/dwelling operations and turn-around operations are set to ¥10 per minute and ¥5 per minute, representing the energy consumption during train movements and the occupation cost of terminal station tracks, respectively. In addition, the passenger capacity of each rolling stock unit is assumed to be 1000 persons, corresponding to a six-car B-type train composition.

For passenger flow distribution, we use time-dependent passenger arrival data from the AFC system and transform it into a passenger arrival rate function $\lambda_{i,s}(t)$ at station i and period s. The passenger flow proportion $\gamma_{i,j,s}$ is computed from the share of the passenger flow volume from origin station i to destination station j during period s. Fig. 8 shows the resulting time-varying passenger arrival rates, with most passenger flow volumes concentrated in the morning and evening peak hours. Finally, according to the prevailing income levels and the estimated value of passengers' time, the unit costs of passenger delay and control operations are specified as ¥2 and ¥0.1 per person per minute, respectively. In addition, a penalty cost of ¥10 per person is imposed for residual passengers.

Figure 7 **Beijing Changping urban rail line**

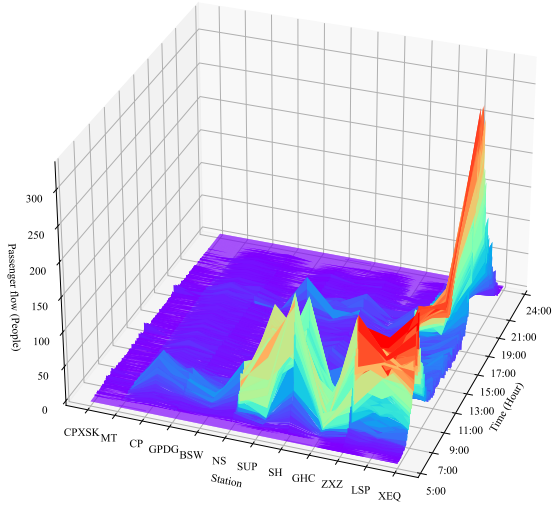


Figure 8 **Passenger arrival flow distribution**

6.2. Experimental results

To evaluate the computational performance of the proposed approach, we design 21 instances with varying parameter settings, indexed as $(F, |\mathcal{R}|, |\mathcal{E}^T|)$, where "$F$" denotes the regular frequency of full-length trips, "$|\mathcal{R}|$" denotes the number of trips, and "$|\mathcal{E}^T|$" denotes the number of events allowing turn-around operations. Initially, we set an average desired headway of 30 minutes during off-peak hours [10:00, 16:00] and an average desired frequency of 15 minutes during peak hours [5:00, 10:00] and [16:00, 23:00], yielding a

total of 120 trips and 608 available turn-around events. Note that the average desired frequencies are used solely for generating the trip set and do not represent a fixed train schedule. Then, instances 2–21 are generated by incrementally increasing the average desired frequencies by 1 minute for off-peak hours and 0.5 minutes for peak hours. We conduct 42 experiments by comparing the standard BDCG algorithm (i.e., column "Standard") with the BDCG algorithm incorporating all acceleration techniques (i.e., column "All techniques"). The enhanced BDCG algorithm achieves notable improvements over the standard approach: the average lower bound, upper bound, and gap (i.e., 570,111.1, 577,811.1, and 2.0%) represent improvements of 6.55%, 0.46%, and 85.40% compared with the standard BDCG results (i.e., 535,067.2, 580,471.1, and 13.7%), respectively.

Table 2 Computational results with different acceleration technique combinations

ID	$(F, \|\mathcal{R}\|, \|\mathcal{E}^T\|)$	Standard					All techniques				
		#Col	LB	UB	Gap(%)	CT(s)	#Col	LB	UB	Gap(%)	CT(s)
1	(30,120,608)	77253	1040934.2	1043344.4	0.2	727.9	179043	1041178.4	1045073.8	0.4	5610.1
2	(29,122,612)	76897	1016000.7	1017702.6	0.2	5288.3	173234	1016575.6	1017830.9	0.1	637.7
3	(28,128,640)	63215	932539.4	937029.7	0.5	7076.0	187169	932731.6	938863.9	0.7	4895.2
4	(27,130,644)	90701	901687.9	905271.0	0.4	4035.0	162719	902552.2	906507.7	0.4	6906.4
5	(26,138,692)	60779	782466.9	786292.6	0.5	5739.9	147535	783370.8	785922.8	0.3	651.0
6	(25,140,696)	56931	754141.1	760713.2	0.8	1944.9	154110	754925.5	759727.8	0.7	2379.7
7	(24,148,744)	44955	680807.8	683255.1	0.4	5444.7	129673	681034.3	683640.0	0.4	1948.6
8	(23,154,772)	58754	662950.1	665951.6	0.4	3865.3	114427	663027.6	667081.7	0.6	6644.3
9	(22,160,800)	45770	630235.5	633427.6	0.5	3755.1	108157	631458.4	633655.7	0.3	6219.4
10	(21,170,852)	36384	578270.0	582250.6	0.7	3704.9	119124	579033.9	581078.7	0.4	4404.6
11	(20,178,900)	30203	537590.1	546028.6	1.5	3828.7	78245	539264.8	541392.0	0.4	4438.3
12	(19,186,932)	34757	498426.5	503844.5	1.1	3719.1	67777	500123.3	500521.4	0.1	4401.1
13	(18,200,1008)	23773	420923.6	428018.8	1.7	3728.6	54763	421541.6	426374.8	1.1	3711.3
14	(17,210,1060)	13514	384790.8	411295.9	6.3	3944.5	33493	385339.4	399338.9	3.6	3606.7
15	(16,222,1116)	15556	330665.3	355555.3	7.0	3728.6	45607	335898.6	351614.1	4.5	3644.1
16	(15,240,1216)	7911	297130.3	340916.1	12.8	3962.4	20667	317124.1	336522.1	5.8	3613.6
17	(14,254,1276)	6523	105366.6	328413.1	67.9	4362.0	19975	300974.8	315302.2	4.5	3891.0
18	(13,274,1380)	7873	299266.1	325479.1	8.1	5419.6	15142	305104.4	317095.3	3.8	4079.3
19	(12,298,1508)	4452	116768.0	310461.0	62.4	3727.0	12633	292993.3	307276.2	4.6	4320.7
20	(11,322,1620)	3687	122784.5	308954.8	60.3	6363.6	18292	288434.7	303899.7	5.1	3783.0
21	(10,356,1800)	4804	142665.4	315687.3	54.8	5981.8	10773	299645.4	315313.8	5.0	4049.8
Ave.		36413.9	535067.2	580471.1	13.7	4302.3	88217.0	570111.1	577811.1	2.0	3992.2

Notes for Table 2: ID: Instance ID; #Col: Number of columns; LB: Lower bound; UB: Upper bound; Gap = $\frac{\text{UB - LB}}{\text{UB}} \times 100\%$; CT: Computing times until incumbent bounds are obtained with no further improvement; Ave.: Average performance for instances.

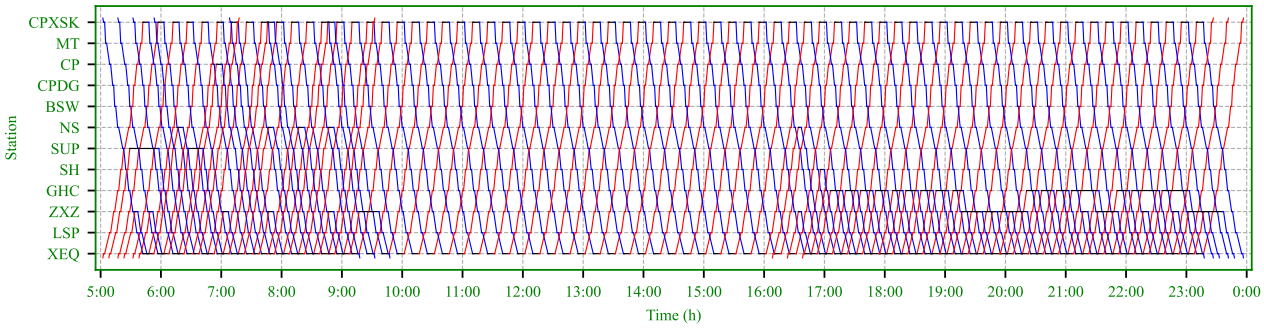


Figure 9 Visualization for rolling stock schedule in Beijing Changping urban rail line

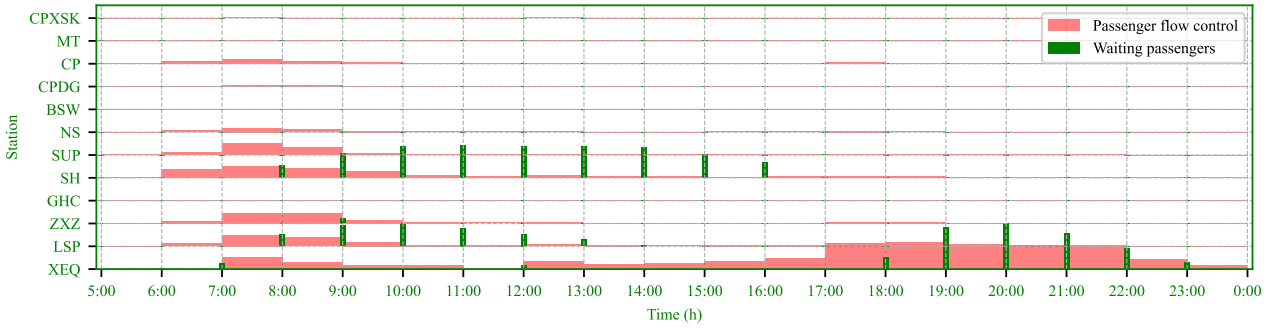


Figure 10 **Visualization for passenger flow evolution in Beijing Changping urban rail line**

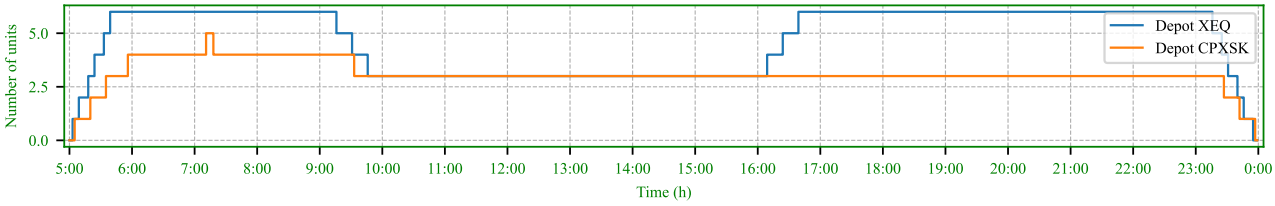


Figure 11 **Visualization for deficit function of rolling stock units**

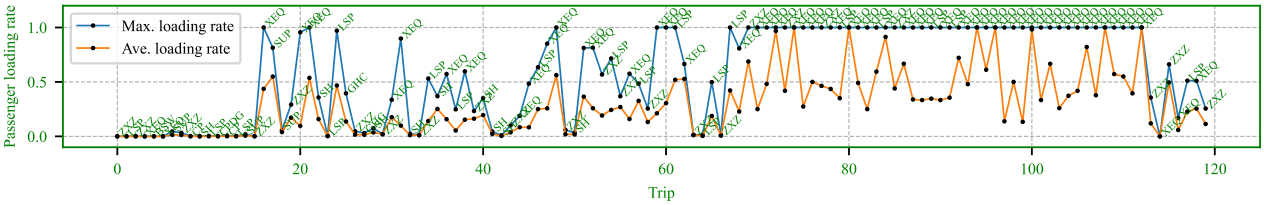


Figure 12 **Visualization for passenger loading rates of trips in upstream direction**

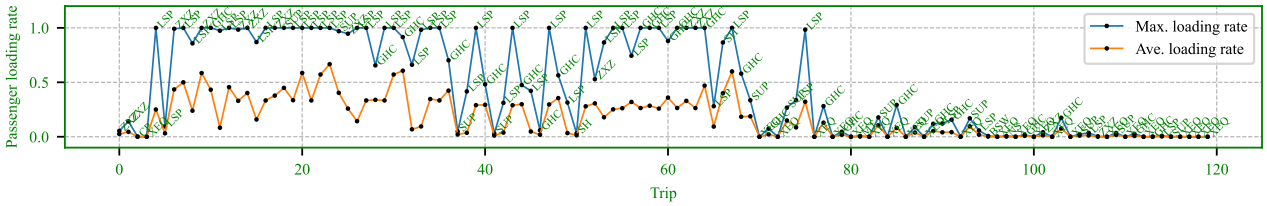


Figure 13 **Visualization for passenger loading rates of trips in downstream direction**

To illustrate the results, we take instance 16 from Table 2 as an example, displaying the rolling stock schedule, passenger flow evolution, deficit function of rolling stock units, and passenger loading rates in Figs. 9–13. As shown in Fig. 9, short-turn operations are scheduled during both morning and evening peak hours, enabling a better match to the uneven passenger flow distributions across physical segments while reducing operational costs. From the passenger flow evolution in Fig. 10, we observe clear passenger flow accumulation and dissipation periods: [8:00, 13:00] at the LSP station, [8:00, 16:00] at the SH station, and [18:00, 23:00] at the XEQ station. In addition, higher passenger flow control rates are applied at multiple stations in the suburb-to-downtown direction (i.e., CPXSK→XEQ) during the morning peak hour [6:00, 10:00], and at the XEQ station in the downtown-to-suburb direction (i.e., XEQ→CPXSK) during the evening peak hour [17:00, 22:00]. Finally, passenger loading rate patterns in Figs. 12–13 indicate that most trips

operate at the maximum 100% loading rate during peak hours, demonstrating the efficiency and service quality achieved by the proposed approach.

6.3. Performance analyses of the number of trips

We find that the number of scheduled trips plays a crucial role in both rolling stock scheduling and passenger flow management performance. For clarity, Fig. 14 and the "Joint RSP-STPM" column in Table 3 present six performance indicators, including total objective value, rolling stock operational cost, total passenger-related cost, passenger delay cost, passenger control cost, and residual passenger penalty cost. Overall, total cost, total passenger-related cost, passenger delay cost, and residual passenger penalty cost decline sharply when transport supply falls short of passenger demand (i.e., $\leq$240 trips), but stabilize once supply meets or exceeds demand (i.e., $\geq$240 trips) (see Figs. 14(a,c,d,f)). This indicates that the marginal benefits of increasing the number of trips diminish once the balance between transport capacity and passenger demand is achieved. Notably, instance 16 with 240 trips represents this critical balance point, delivering the most effective trade-off between rolling stock scheduling and passenger flow management. Lastly, both the rolling stock operational cost and passenger flow control cost rise as more trips are scheduled due to the increase in transport supplies. While, the passenger flow control cost levels off once transport supply and passenger demand are balanced (i.e., $\geq$240 trips).

6.4. Performance analyses of sequential vs. joint optimization schemes

To highlight the advantages of the joint optimization scheme, we define the sequential rolling stock scheduling and passenger flow management model (i.e., sequential RSP-STPM) for comparison with the joint RSP-STPM. Specifically, let $\mathbf{Y}$ denote the set of path selection vectors satisfying constraints (2)–(5) and (14), i.e., $\mathbf{Y} = \left\{ \mathbf{y} = \left(y_1, \cdots, y_p, \cdots, y_{|\mathcal{P}|}\right)^T | \text{constraints (2)–(5), (14)}, p \in \mathcal{P} \right\}$. Then, the sequential RSP-STPM is solved in two stages. First, the rolling stock scheduling model (i.e., $\min_{\mathbf{y}\in\mathbf{Y}} \left\{ \sum_{p\in\mathcal{P}} C_p^R \cdot y_p \right\}$) is solved via the standard B&P algorithm. Next, the passenger flow management model is solved based on the resulting path selection $\overline{\mathbf{y}} \in \mathbf{Y}$, subject to constraints (6)–(12) and (15)–(17).

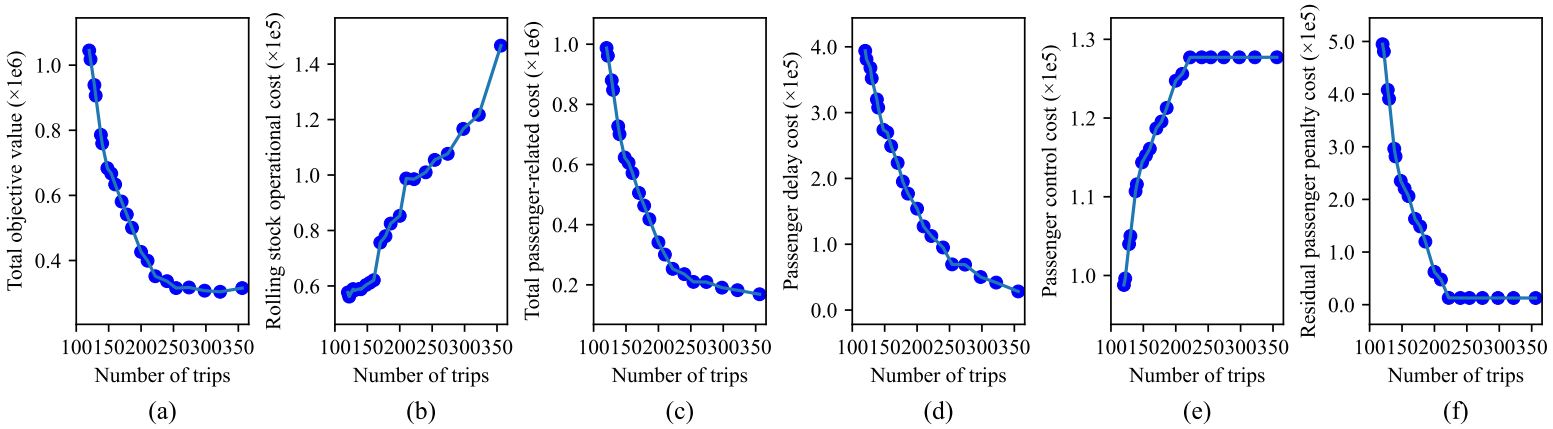


Figure 14 Performance indicators with respect to different numbers of trips

Table 3 Summary of optimization results in performance analyses

ID	$(F, \|\mathcal{R}\|)$	Sequential RSP-STPM with short-turn operations ($\times 10^3$)						Joint RSP-STPM with short-turn operations ($\times 10^3$)						Joint RSP-STPM without short-turn operations ($\times 10^3$)					
		TC	RSC	PC	PDC	PCC	RPC	TC	RSC	PC	PDC	PCC	RPC	TC	RSC	PC	PDC	PCC	RPC
1	(30,120)	1160.76	52.35	1108.41	439.51	93.98	574.91	1045.07	57.66	987.42	393.90	98.79	494.73	1054.68	69.49	985.20	392.41	98.84	493.94
2	(29,122)	1128.93	52.42	1076.51	421.49	94.87	560.15	1017.83	56.15	961.68	381.05	99.62	481.02	1029.83	69.37	960.46	381.27	99.71	479.48
3	(28,128)	1056.14	53.76	1002.38	405.67	98.59	498.12	938.86	58.86	880.00	367.96	103.99	408.04	946.76	71.08	875.68	363.93	104.01	407.74
4	(27,130)	1025.93	53.72	972.21	389.71	99.50	483.00	906.51	58.38	848.13	352.13	105.02	390.98	916.13	70.84	845.29	349.25	105.02	391.02
5	(26,138)	936.75	55.24	881.51	364.16	103.66	413.69	785.92	58.96	726.97	320.00	110.70	296.26	798.04	73.26	724.78	317.71	110.69	296.37
6	(25,140)	905.34	55.08	850.26	346.34	104.51	399.40	759.73	58.81	700.92	307.71	111.58	281.62	769.64	72.88	696.76	303.40	111.57	281.79
7	(24,148)	834.70	56.40	778.30	322.85	107.61	347.84	683.64	60.34	623.30	273.82	114.37	235.11	695.30	75.05	620.25	270.79	114.37	235.10
8	(23,154)	790.39	57.28	733.11	303.20	109.24	320.67	667.08	61.19	605.90	269.95	115.23	220.72	678.01	76.20	601.81	265.89	115.24	220.69
9	(22,160)	756.09	58.06	698.03	281.86	110.11	306.05	633.66	62.08	571.58	249.16	116.10	206.31	646.86	77.24	569.62	247.21	116.10	206.31
10	(21,170)	710.28	72.30	637.98	252.66	112.08	273.23	581.08	75.65	505.43	223.63	118.69	163.12	596.07	92.22	503.85	222.04	118.69	163.11
11	(20,178)	669.91	73.06	596.85	225.41	112.97	258.47	541.39	77.95	463.45	195.18	119.55	148.71	562.67	100.36	462.31	194.07	119.56	148.69
12	(19,186)	641.98	79.09	562.89	213.70	114.39	234.80	500.52	82.50	418.02	176.90	121.29	119.84	522.43	104.64	417.79	176.64	121.29	119.86
13	(18,200)	573.89	81.76	492.13	187.39	117.23	187.51	426.37	85.28	341.09	154.06	124.74	62.30	446.11	108.56	337.55	150.59	124.74	62.21
14	(17,210)	546.78	82.52	464.26	167.83	117.76	178.67	399.34	98.76	300.58	127.12	125.61	47.85	410.34	110.28	300.06	126.66	125.61	47.80
15	(16,222)	487.01	84.27	402.74	142.40	120.06	140.27	351.61	98.50	253.11	112.51	127.70	12.90	363.82	112.57	251.25	110.84	127.72	12.70
16	(15,240)	427.21	87.04	340.17	124.99	122.94	92.23	336.52	100.98	235.55	95.10	127.71	12.74	347.64	116.97	230.67	90.28	127.72	12.68
17	(14,254)	385.38	101.43	283.95	99.13	124.88	59.94	315.30	105.46	209.84	69.38	127.71	12.74	341.04	131.70	209.34	68.96	127.72	12.66
18	(13,274)	326.46	104.28	222.18	81.74	127.71	12.73	317.10	107.65	209.45	68.99	127.71	12.74	343.86	140.93	202.93	62.55	127.72	12.66
19	(12,298)	311.88	118.14	193.75	53.28	127.71	12.76	307.28	116.64	190.64	50.18	127.71	12.74	331.18	150.94	180.24	39.86	127.72	12.66
20	(11,322)	310.55	124.03	186.52	46.06	127.71	12.74	303.90	121.72	182.18	41.72	127.71	12.74	330.38	155.30	175.08	34.70	127.72	12.66
21	(10,356)	321.18	150.18	171.01	30.63	127.72	12.66	315.31	146.66	168.65	28.27	127.72	12.66	371.01	205.06	165.95	25.57	127.72	12.66
Ave.		681.31	78.69	602.62	233.33	113.11	256.18	577.81	83.34	494.47	202.80	118.06	173.61	595.32	104.04	491.28	199.74	118.07	173.47

Notes for Table 3: ID: Instance ID; TC: Total cost; RSC: Rolling stock operational cost; PC: Total passenger-related cost; PDC: Passenger delay cost; PCC: Passenger control cost; RPC: Residual passenger penalty cost; Ave.: Average performance for instances.

Table 3 compares computational results for rolling stock operational costs and passenger-related costs between sequential and joint RSP-STPM solutions. On average, the joint optimization scheme (see column "Joint RSP-STPM with short-turn operations") achieves an objective value of 577,811.1, representing a 15.19% reduction compared with the sequential approach (see column "Sequential RSP-STPM with short-turn operations") at 681,310.5 (the reduction is as high as 26.97% in some instances). The reductions mainly come from substantial decreases in passenger delay costs (i.e., 13.08% on average) and residual passenger penalty costs (i.e., 32.23% on average), accompanied by slight increases in rolling stock operational costs (i.e., 5.92% on average) and passenger flow control costs (i.e., 4.20% on average). This performance is mainly because the sequential RSP-STPM tends to schedule all trips at their minimum possible lengths to reduce rolling stock operational costs, which inevitably increases passenger delay and residual passenger penalties. In contrast, the joint scheme coordinates both rolling stock scheduling and passenger flow management, achieving a better balance between rolling stock operational efficiency and passenger service quality.

6.5. Performance analyses of short-turn operations

Compared with full-length trips (i.e., turn-around operations are only permitted at terminal stations), short-turn operations can enhance rolling stock scheduling flexibility and better align with passenger flow distributions across peak and off-peak segments. To evaluate the effectiveness, we test 21 additional instances both with and without short-turn operations. Table 3 presents the computational results, including total rolling stock operational cost, total passenger-related cost, passenger delay cost, passenger control cost, and residual passenger penalty cost. On average, the objective value with short-turn operations (i.e., 577,811.1 in column "Joint RSP-STPM with short-turn operations") performs an average 2.94% reduction compared with the traditional scheme without short-turn operations (i.e., 595,323.0 in column "Joint RSP-STPM without short-turn operations"), with reductions reaching up to 15.01% in some instances. This improvement is primarily driven by a significant 19.90% decrease in rolling stock operational costs, offset slightly by a modest 0.65% increase in total passenger-related costs. This is mainly because some trips can turn around at intermediate stations, servicing only the partial physical segments with high passenger flow volumes. This approach reduces rolling stock operational costs compared to turning around only at terminal stations. Additionally, short-turn operations improve rolling stock circulation efficiency by allowing partial-length trips and reducing the total number of required rolling stock units.

7. Conclusions and future research

In this study, we addressed the challenge of jointly controlling and assigning passenger flows in urban rail systems, which has often been treated separately from rolling stock decisions. To this end, we formulated the RSP-STPM as a path-based model that integrates short-turn operations with a period-based passenger flow representation. To handle large-scale real-world instances, we developed an exact BDCG algorithm enhanced with multiple acceleration techniques. The proposed approach was evaluated on both small-scale

and real-world cases. The computational results show that the approach can effectively schedule rolling stock units with flexible short-turn operations to closely match uneven spatial and temporal passenger flow distributions. This enables operators to design precise and effective passenger flow control schemes. Moreover, the methodology can be generalized to other passenger-oriented urban rail operation problems, avoiding the need to explicitly formulate passenger flow evaluation and passenger delay calculation with numerous discrete time-dependent parameters and variables.

From a managerial perspective, the rolling stock scheduling efficiency and the passenger service quality can be improved in two ways: (1) The joint optimization scheme reduces total rolling stock operational and passenger-related costs. (2) Short-turn operations, rather than restricting turn-around operations to terminal stations, can lower rolling stock operational costs without sacrificing passenger service quality. Further research will focus on the following aspects: (1) The proposed period-based passenger flow approach can be extended to address uneven passenger flow distributions across multiple scenarios. (2) A promising research direction is to control and assign OD-based passenger flows across multiple physical lines in an urban rail network using the period-based passenger flow framework.

References

Abbink, E., van den Berg, B., Kroon, L., Salomon, M., 2004. Allocation of railway rolling stock for passenger trains. Transportation Science 38, 33–41.

Balas, E., Carrera, M.C., 1996. A dynamic subgradient-based branch-and-bound procedure for set covering. Operations Research 44, 875–890.

Baldacci, R., Christofides, N., Mingozzi, A., 2008. An exact algorithm for the vehicle routing problem based on the set partitioning formulation with additional cuts. Mathematical Programming 115, 351–385.

Cacchiani, V., Caprara, A., Toth, P., 2008. A column generation approach to train timetabling on a corridor. 4OR 6, 125–142.

Cacchiani, V., Caprara, A., Toth, P., 2010a. Non-cyclic train timetabling and comparability graphs. Operations Research Letters 38, 179–184.

Cacchiani, V., Caprara, A., Toth, P., 2010b. Solving a real-world train-unit assignment problem. Mathematical Programming 124, 207–231.

Cacchiani, V., Caprara, A., Toth, P., 2013. A Lagrangian heuristic for a train-unit assignment problem. Discrete Applied Mathematics 161, 1707–1718.

Caprara, A., Kroon, L., Monaci, M., Peeters, M., Toth, P., 2007. Chapter 3 passenger railway optimization, in: Barnhart, C., Laporte, G. (Eds.), Transportation. Elsevier. volume 14 of *Handbooks in Operations Research and Management Science*, pp. 129–187.

Cordeau, J.F., Soumis, F., Desrosiers, J., 2000. A Benders decomposition approach for the locomotive and car assignment problem. Transportation Science 34, 133–149.

Cordeau, J.F., Soumis, F., Desrosiers, J., 2001. Simultaneous assignment of locomotives and cars to passenger trains. Operations Research 49, 531–548.

Corman, F., 2020. Interactions and equilibrium between rescheduling train traffic and routing passengers in microscopic delay management: A game theoretical study. Transportation Science 54, 785–822.

Davidich, M., Geiss, F., Mayer, H.G., Pfaffinger, A., Royer, C., 2013. Waiting zones for realistic modelling of pedestrian dynamics: A case study using two major german railway stations as examples. Transportation Research Part C: Emerging Technologies 37, 210–222.

Delgado, F., Munoz, J.C., Giesen, R., 2012. How much can holding and/or limiting boarding improve transit performance? Transportation Research Part B: Methodological 46, 1202–1217.

Desaulniers, G., Desrosiers, J., Solomon, M.M., 2006. Column generation. volume 5. Springer Science & Business Media.

Dollevoet, T., Huisman, D., Schmidt, M., Schöbel, A., 2012. Delay management with rerouting of passengers. Transportation Science 46, 74–89.

Duarte, P.J.C., Dauzère-Pérès, S., Huisman, D., Mannino, C., Sartor, G., Weik, N., Widmann, P., 2025. 50 years of or in railway timetabling and rolling stock planning. EURO Journal on Transportation and Logistics 14, 100155.

Fioole, P.J., Kroon, L., Maróti, G., Schrijver, A., 2006. A rolling stock circulation model for combining and splitting of passenger trains. European Journal of Operational Research 174, 1281–1297.

Freling, R., Lentink, R.M., Kroon, L.G., Huisman, D., 2005. Shunting of passenger train units in a railway station. Transportation Science 39, 261–272.

Ghaemi, N., Cats, O., Goverde, R.M., 2017. A microscopic model for optimal train short-turnings during complete blockages. Transportation Research Part B: Methodological 105, 423–437.

Grimm, B., Hoogervorst, R., Borndörfer, R., 2025. A comparison of two models for rolling stock scheduling. Transportation Science 59, 1101–1129.

Holmberg, K., Yuan, D., 2000. A Lagrangian heuristic based branch-and-bound approach for the capacitated network design problem. Operations Research 48, 461–481.

Hoogervorst, R., Dollevoet, T., Maróti, G., Huisman, D., 2020. Reducing passenger delays by rolling stock rescheduling. Transportation Science 54, 762–784.

van der Hurk, E., Kroon, L., Maróti, G., 2018. Passenger advice and rolling stock rescheduling under uncertainty for disruption management. Transportation Science 52, 1391–1411.

Jiang, F., Cacchiani, V., Toth, P., 2017. Train timetabling by skip-stop planning in highly congested lines. Transportation Research Part B: Methodological 104, 149–174.

Kroon, L., Maróti, G., Nielsen, L., 2015. Rescheduling of railway rolling stock with dynamic passenger flows. Transportation Science 49, 165–184.
Li, Z., Lo, S., Ma, J., Luo, X., 2020. A study on passengers' alighting and boarding process at metro platform by computer simulation. Transportation Research Part A: Policy and Practice 132, 840–854.
Liang, J., Lyu, G., Teo, C.P., Gao, Z., 2023. Online passenger flow control in metro lines. Operations Research 71, 768–775.
Lin, Z., Kwan, R.S., 2016. A branch-and-price approach for solving the train unit scheduling problem. Transportation Research Part B: Methodological 94, 97–120.
Lingaya, N., Cordeau, J.F., Desaulniers, G., Desrosiers, J., Soumis, F., 2002. Operational car assignment at VIA Rail Canada. Transportation Research Part B: Methodological 36, 755–778.
Lu, J., Nie, Q., Mahmoudi, M., Ou, J., Li, C., Zhou, X.S., 2022. Rich arc routing problem in city logistics: Models and solution algorithms using a fluid queue-based time-dependent travel time representation. Transportation Research Part B: Methodological 166, 143–182.
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. Transportation Research Part B: Methodological 168, 27–55.
Luan, X., Corman, F., 2022. Passenger-oriented traffic control for rail networks: An optimization model considering crowding effects on passenger choices and train operations. Transportation Research Part B: Methodological 158, 239–272.
Lusby, R.M., Larsen, J., Ehrgott, M., Ryan, D., 2011. Railway track allocation: Models and methods. OR Spectrum 33, 843–883.
Lyu, G., Cheung, W.C., Chou, M.C., Teo, C.P., Zheng, Z., Zhong, Y., 2019. Capacity allocation in flexible production networks: Theory and applications. Management Science 65, 5091–5109.
Newell, C., 2013. Applications of queueing theory. volume 4. Springer Science & Business Media.
Nguyen, S., Pallottino, S., Malucelli, F., 2001. A modeling framework for passenger assignment on a transport network with timetables. Transportation Science 35, 238–249.
Niu, H., Zhou, X., 2013. Optimizing urban rail timetable under time-dependent demand and oversaturated conditions. Transportation Research Part C: Emerging Technologies 36, 212–230.
Niu, H., Zhou, X., Gao, R., 2015. Train scheduling for minimizing passenger waiting time with time-dependent demand and skip-stop patterns: Nonlinear integer programming models with linear constraints. Transportation Research Part B: Methodological 76, 117–135.
Pan, H., Yang, L., Liang, Z., 2023. Demand-oriented integration optimization of train timetabling and rolling stock circulation planning with flexible train compositions: A column-generation-based approach. European Journal of Operational Research 305, 184–206.
Schettini, T., Jabali, O., Malucelli, F., 2022. Demand-driven timetabling for a metro corridor using a short-turning acceleration strategy. Transportation Science 56, 919–937.
Sharma, B., Pellegrini, P., Rodriguez, J., Chaudhary, N., 2023. A review of passenger-oriented railway rescheduling approaches. European Transport Research Review 15, 14.
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. Transportation Research Part B: Methodological 110, 26–59.
Shi, X., Li, X., 2021. Operations design of modular vehicles on an oversaturated corridor with first-in, first-out passenger queueing. Transportation Science 55, 1187–1205.
Sun, L., Jin, J.G., Lee, D.H., Axhausen, K.W., Erath, A., 2014. Demand-driven timetable design for metro services. Transportation Research Part C: Emerging Technologies 46, 284–299.
Van Lieshout, R.N., 2021. Integrated periodic timetabling and vehicle circulation scheduling. Transportation Science 55, 768–790.
Wang, E., Yang, L., Yin, J., Zhang, J., Gao, Z., 2024. Passenger-oriented rolling stock scheduling in the metro system with multiple depots: Network flow based approaches. Transportation Research Part B: Methodological 180, 102885.
Wu, Y., Yang, H., Zhao, S., Shang, P., 2021. Mitigating unfairness in urban rail transit operation: A mixed-integer linear programming approach. Transportation Research Part B: Methodological 149, 418–442.
Yang, Y., 2025. Deluxing: Deep Lagrangian underestimate fixing for column-generation-based exact methods. Operations Research 73, 1184–1207.
Yazdani, D., Omidvar, M.N., Deplano, I., Lersteau, C., Makki, A., Wang, J., Nguyen, T.T., 2019. Real-time seat allocation for minimizing boarding/alighting time and improving quality of service and safety for passengers. Transportation Research Part C: Emerging Technologies 103, 158–173.
Yin, J., Yang, L., D'Ariano, A., Tang, T., Gao, Z., 2022. Integrated backup rolling stock allocation and timetable rescheduling with uncertain time-variant passenger demand under disruptive events. INFORMS Journal on Computing 34, 3234–3258.
Ying, C., Chow, A.H., Yan, Y., Kuo, Y.H., Wang, S., 2024. Adaptive rescheduling of rail transit services with short-turnings under disruptions via a multi-agent deep reinforcement learning approach. Transportation Research Part B: Methodological 188, 103067.
Zhang, P., Wu, J., Wang, K., Qu, Y., Long, J., 2024. Dynamic flow control model and algorithm for metro network under FIFO condition. Transportation Research Part B: Methodological 190, 103089.
Zhou, X.S., Cheng, Q., Wu, X., Li, P., Belezamo, B., Lu, J., Abbasi, M., 2022. A meso-to-macro cross-resolution performance approach for connecting polynomial arrival queue model to volume-delay function with inflow demand-to-capacity ratio. Multimodal Transportation 1, 100017.
Zhou, Y., Yang, H., Wang, Y., Yan, X., 2021. Integrated line configuration and frequency determination with passenger path assignment in urban rail transit networks. Transportation Research Part B: Methodological 145, 134–151.
Zhu, Y., Goverde, R.M., 2019. Railway timetable rescheduling with flexible stopping and flexible short-turning during disruptions. Transportation Research Part B: Methodological 123, 149–181.

Appendix
A. Definition of indices, sets and parameters

Table 4: Definition of indices, sets and parameters

Notation	Definition
Indices	
i, j	Index of stations
r	Index of trips
k	Index of rolling stock units
e/e'	Index of events
a	Index of activities
t	Index of timestamps
s	Index of periods
p	Index of (rolling stock) paths
g	Index of rolling stock groups
$r(e)/i(e)$	Associated trip/station of event e
$g(k)/g(p)$	Group of rolling stock unit k / path p
$e_0(r)/e_{\text{end}}(r)$	First/last event of trip r
$e_{\text{ori}}(g)/e_{\text{des}}(g)$	Depot source/sink of rolling stock units in group g
Sets	
$\mathcal{I}$	Set of stations
$\mathcal{I}^T$	Set of candidate turn-around stations
$\mathcal{R}$	Set of trips
$\mathcal{K}$	Set of rolling stock units
$\mathcal{K}_g$	Set of rolling stock units in group g
$\mathcal{T}$	Set of timestamps
$\mathcal{S}$	Set of periods
$\mathcal{S}_i$	Set of periods at station i
$\mathcal{OD}$	Set of passenger OD pairs
$\mathcal{G}$	Set of rolling stock groups
$\mathcal{P}$	Set of (rolling stock) paths
$\mathcal{P}_g$	Set of paths of rolling stock units in group g
$\mathcal{T}_e$	Set of start times of event e
$\mathcal{G}^N$	Event-activity network
$\mathcal{E}$	Set of events
$\mathcal{E}^D/\mathcal{E}^A$	Set of departure/arrival events
$\mathcal{E}_r^D$	Set of departure events of trip r
$\mathcal{E}_r^{DT}/\mathcal{E}_r^{AT}$	Set of departure/arrival events at candidate turn-around stations of trip r
$\mathcal{E}_e^-/\mathcal{E}_e^+$	Set of incoming/outgoing events of event e
$\mathcal{E}_{<e}/\mathcal{E}_{\geq e}$	Set of events before/after e along the involved trip
$\mathcal{A}$	Set of activities
$\mathcal{A}^R/\mathcal{A}^D/\mathcal{A}^T/\mathcal{A}^{\text{Pin}}/\mathcal{A}^{\text{Pout}}$	Set of running/dwelling/turn-around/pull-in/pull-out activities
$\mathcal{A}^{OD}$	Set of passenger activities
$\mathcal{A}_{i,j}^{OD}$	Set of passenger activities for OD pair (i, j)
$\mathcal{A}_a^{\text{pre}}$	Set of passenger activities preceding activity a
$\mathcal{A}_e^O/\mathcal{A}_e^D$	Set of origin/destination passenger activities starting from / ending at departure event e
$\mathcal{A}_e^{\text{pass}}$	Set of passenger activities passing through event e
$\mathcal{V}$	Set of potential conflicting event pairs
$\mathcal{T}_{e,e'}^{\mathcal{V}}$	Set of the possible conflicting timestamps for potential conflicting event pair $(e, e') \in \mathcal{V}$
Parameters	
c_g^1	Fixed operational cost of assigning one rolling stock unit from group g
$c_{g,e,e'}^2$	Fixed operational cost per minute for activity (e, e') by rolling stock units in group g
$c_{i,s}^D/c_{i,s}^C$	Unit cost of passenger delay / flow control rate at station i and period s
$c_{i,j}^{\text{re}}$	Unit penalty cost of residual passenger flow volume for OD pair (i, j)
C_p^R	Cost of (rolling stock) path p
$\underline{T}_e/\overline{T}_e$	Earliest/latest start time of event e
$h_{e,e'}^{\min}$	Minimum headway time between events e and e'
$\underline{t}_s/\overline{t}_s$	Start/end time of period s
$T_{e,e'}^{\min}$	Minimum running/dwelling/turn-around time of activity (e, e')

Table 4: Definition of indices, sets and parameters

Notation	Definition
Cap_r	Passenger capacity of trip r
$\gamma_{i,j,s}$	The proportion of OD pair (i,j) in the total boarding passenger flow volume at station i during period s
$t_{i,s}^{QS}/t_{i,s}^{QE}$	Start/end time of the passenger queuing process at station i in period s
$T_{i,s}^{Q\max}$	Time of the maximum passenger queue length at station i in period s
$\widehat{\lambda}_{i,s}$	Shape parameter of passenger arrival rate function at station i in period s
$\lambda_{i,s}^{\min}/\lambda_{i,s}^{\max}$	Minimum/maximum passenger arrival rate at station i in period s
$\alpha_{p,e,t}$	Binary parameter, =1 if event e is included in path p with start time t
$\beta_{p,r}$	Binary parameter, =1 if trip r is included in path p
Variables	
y_p	Binary variables, =1 if (rolling stock) path p is selected; =0 otherwise
$\widehat{w}_{i,s}$	Continuous variables, representing total passenger delay at station i in period s
$\widehat{\mu}_{i,s}$	Continuous variables, representing passenger flow control constant rate at station i in period s
$W_{i,s}$	Continuous variables, representing waiting passenger flow volume at station i at the start of period s
$p_{a,s}$	Continuous variables, representing boarding passenger flow volume for activity a in period s
$p_{i,j}^{\text{re}}$	Continuous variables, representing residual passenger flow volume for OD pair (i,j) at the end of the planning horizon
$x_{e,t}$	Binary variables, =1 if event e starts at timestamp t; =0 otherwise
$z_{k,e_{\text{ori}}(g(k)),e}$	Binary variables, =1 if event e is connected from source $e_{\text{ori}}(g(k))$ for rolling stock unit k; =0 otherwise
$z_{k,e,e'}$	Binary variables, =1 if events e and e' are connected with the turn-around operation for rolling stock unit k; =0 otherwise
$z_{k,e,e_{\text{des}}(g(k))}$	Binary variables, =1 if event e is connected to depot sink $e_{\text{des}}(g(k))$ for rolling stock unit k; =0 otherwise
n_g	Integer variable, representing the total number of assigned rolling stock units in group g
$v_{k,e,e'}$	Integer variable, representing the actual time required for running, dwelling, or turn-around activity (e,e') by rolling stock unit k

B. Equivalent primal model of the RSP-STPM

In this section, we present a compact mathematical optimization model of the RSP-STPM for validation purposes, which is equivalent to the path-based model M1 in Section 4 and can be directly solved by general-purpose solvers (e.g., CPLEX). The notation, objective function, and system constraints of the model M1 are introduced in Sections B.1 and B.2, respectively.

B.1. Notation

The primal model M5 employs four types of decision variables: event start time selection variable $x_{e,t}$ taking value 1 if event e starts at timestamp t (otherwise $x_{e,t}=0$); rolling stock starting variable $z_{k,e_{\text{ori}}(g(k)),e}$ taking value 1 if event e is connected from source $e_{\text{ori}}(g(k))$ for rolling stock unit k (otherwise $z_{k,e_{\text{ori}}(g(k)),e}=0$); rolling stock turn-around variable $z_{k,e,e'}$ taking value 1 if events e and e' are connected with the turn-around operation for rolling stock unit k (otherwise $z_{k,e,e'}=0$); rolling stock ending variable $z_{k,e,e_{\text{des}}(g(k))}$ taking value 1 if event e is connected to depot sink $e_{\text{des}}(g(k))$ for rolling stock unit k (otherwise $z_{k,e,e_{\text{des}}(g(k))}=0$). Additionally, two auxiliary variables are introduced for objective function calculation: positive integer variable n_g representing the total number of assigned rolling stock units in group g; and positive integer variable $v_{k,e,e'}$ representing the actual time required for running, dwelling, or turn-around activity (e,e') by rolling stock unit k.

B.2. Objective function and system constraints

The objective function in Eq. (32) minimizes the sum of rolling stock operational costs and passenger-related costs, and is equivalent to Eq. (1) in the path-based model M1. Specifically, the first term represents the total cost of assigning rolling stock units. The second term accounts for the total cost of actual running, dwelling, and turn-around activity times. The third term captures the total passenger delay and flow control cost, while the final term imposes residual passenger penalties.

$$\min \quad \sum_{g\in\mathcal{G}} c_g^1 n_g + \sum_{k\in\mathcal{K}} \sum_{(e,e')\in\mathcal{A}} c^2_{g(k),e,e'} v_{k,e,e'} + \sum_{i\in\mathcal{I}} \sum_{s\in\mathcal{S}_i} \left(c^D_{i,s} \widehat{w}_{i,s} + c^C_{i,s} \int_{\underline{t}_s}^{\bar{t}_s} \widehat{\mu}_{i,s} dt \right) + \sum_{(i,j)\in\mathcal{OD}} c^{\mathrm{re}}_{i,j} p^{\mathrm{re}}_{i,j} \tag{32}$$

The feasibility and consistency of the model are guaranteed through four groups of system constraints, formulated as below, where the train scheduling constraints, rolling stock circulation constraints, passenger flow management constraints, and the ranges of the decision variables are held in Groups I–IV, respectively.

Group I: Train scheduling constraints

$$\sum_{t\in\mathcal{T}_e} x_{e,t} \le 1 \quad \forall e \in \mathcal{E} \tag{33}$$

$$\sum_{t'\in\mathcal{T}_e\cap\{t\}} x_{e,t'} + \sum_{t'\in\mathcal{T}_{e'}\cap\mathbb{Z}_{\left[t,t^H_{e,e',t}\right]}} x_{e',t'} \le 1 \quad \forall t \in \mathcal{T}^{\mathcal{V}}_{e,e'}, (e,e') \in \mathcal{V} \tag{34}$$

$$\sum_{e\in\mathcal{E}_r} \sum_{t\in\mathcal{T}_e} x_{e,t} \ge N^{\min}_r \quad \forall r \in \mathcal{R} \tag{35}$$

$$\sum_{t\in\mathcal{T}_{e'}} t\cdot x_{e',t} - \sum_{t\in\mathcal{T}_e} t\cdot x_{e,t} \ge T^{\min}_{e,e'} + M_{e,e'} \left(\sum_{t\in\mathcal{T}_e} x_{e,t} + \sum_{t\in\mathcal{T}_{e'}} x_{e',t} - 2 \right) \quad \forall (e,e') \in \mathcal{A}^R \cup \mathcal{A}^D \tag{36}$$

Constraints (33) ensure that each event e is assigned at most one start timestamp within its time window $\mathcal{T}_e$ or is canceled, which is equivalent to constraints (4) in the path-based model. Constraints (34) enforce safety headway, which is equivalent to constraints (5). Here, $t^H_{e,e',t} = \min\left\{t + h^{\min}_{e,e'} - 1, |\mathcal{T}|\right\}$ denotes the latest incompatible start time for event pair $(e,e') \in \mathcal{V}$, given that event e starts at time t, and $h^{\min}_{e,e'}$ represents the minimum headway between events e and e'. Constraints (35) require each trip r to operate with at least $N^{\min}_r$ events, preventing full trip cancellations. Finally, constraints (36) enforce that each running or dwelling activity (e,e') meets the minimum travel time $T^{\min}_{e,e'}$ when both events e and e' are not cancelled (i.e., $\sum_{t\in\mathcal{T}_e} x_{e,t} + \sum_{t\in\mathcal{T}_{e'}} x_{e',t} = 2$).

Group II: Rolling stock circulation constraints

$$\sum_{r\in\mathcal{R}} z_{k,e_{\mathrm{ori}}(g),e_0(r)} \le 1 \quad g = g(k), \forall k \in \mathcal{K} \tag{37}$$

$$z_{k,e_{\mathrm{ori}}(g),e_0(r)} + \sum_{e\in\mathcal{E}^{AT}_r} \sum_{e'\in\mathcal{E}^-_e} z_{k,e',e} = z_{k,e_{\mathrm{end}}(r),e_{\mathrm{des}}(g)} + \sum_{e\in\mathcal{E}^{DT}_r} \sum_{e'\in\mathcal{E}^+_e} z_{k,e,e'} \quad g = g(k), \forall r \in \mathcal{R}, k \in \mathcal{K} \tag{38}$$

$$\sum_{t\in\mathcal{T}_{e'}} t\cdot x_{e',t} \ge \sum_{t\in\mathcal{T}_e} t\cdot x_{e,t} + T^{\min}_{e,e'} + M_{e,e'}(z_{k,e,e'} - 1) \quad \forall (e,e') \in \mathcal{A}^C, k \in \mathcal{K} \tag{39}$$

Constraints (37) assign each rolling stock unit k to at most one initial trip start event $e_0(r)$ from its depot origin $e_{\text{ori}}(g)$ with respect to rolling stock group $g = g(k)$. Constraints (38) ensure flow conservation on each trip r, balancing all incoming flows (i.e., $z_{k,e_{\text{ori}}(g),e_0(r)} + \sum_{e\in\mathcal{E}_r^{AT}} \sum_{e'\in\mathcal{E}_e^-} z_{k,e',e}$) with outgoing flows (i.e., $z_{k,e_{\text{end}}(r),e_{\text{des}}(g)} + \sum_{e\in\mathcal{E}_r^{DT}} \sum_{e'\in\mathcal{E}_e^+} z_{k,e,e'}$). As shown in Fig. 15(a), for one specific flow conservation constraint on trip r, the incoming flows (i.e., red turn-around activity) are imposed at all arrival events $e \in \mathcal{E}_r^{AT}$, and the outgoing flows (i.e., blue turn-around activity) are imposed at all departure events $e \in \mathcal{E}_r^{DT}$. Constraints (39) enforces the minimum turn-around time $T_{e,e'}^{\min}$ when a turn-around activity (e, e') is assigned, where $M_{e,e'} = \overline{T}_e + T_{e,e'}^{\max}$ is the smallest valid big-M constant, $\overline{T}_e$ is the latest start time of event e, and $T_{e,e'}^{\max}$ is the maximum turn-around time between events e and e'.

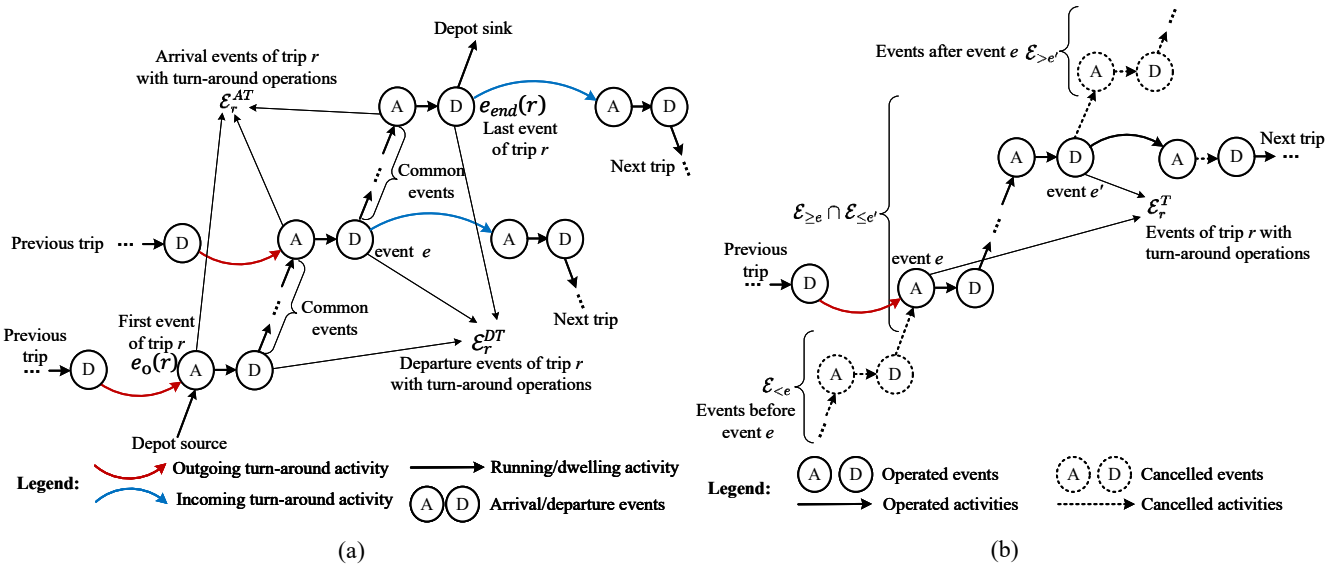


Figure 15 Illustration of rolling stock flow balance

$$
\begin{cases}
\sum\limits_{t''\in\mathcal{T}_{e''}} x_{e'',t''} \leq M_{e,e'} \left(1 - z_{k,e_{\text{ori}}(g),e} - \sum\limits_{e'''\in\mathcal{E}_e^-} z_{k,e''',e}\right) & \forall e'' \in \mathcal{E}_{<e}(a) \\
\sum\limits_{t''\in\mathcal{T}_{e''}} x_{e'',t''} \geq 1 - M_{e,e'} \left(2 - z_{k,e_{\text{ori}}(g),e} - z_{k,e',e_{\text{des}}(g)} - \sum\limits_{e'''\in\mathcal{E}_e^-} z_{k,e''',e} - \sum\limits_{e''\in\mathcal{E}_{e'}^+} z_{k,e',e'''}\right) & \forall e'' \in \mathcal{E}_{\geq e} \cap \mathcal{E}_{\leq e'}(b) \\
\sum\limits_{t''\in\mathcal{T}_{e''}} x_{e'',t''} \leq M_{e,e'} \left(1 - z_{k,e',e_{\text{des}}(g)} - \sum\limits_{e'''\in\mathcal{E}_{e'}^+} z_{k,e',e'''}\right) & \forall e'' \in \mathcal{E}_{>e'}(c) \\
& g = g(k), \forall e, e' \in \mathcal{E}_r^T, r \in \mathcal{R}, k \in \mathcal{K}
\end{cases}
\tag{40}
$$

To avoid infeasible partial trips, the short-turn coupling constraints (40) classify all events in trip r relative to candidate turn-around events e and e' into set $\mathcal{E}_{<e}$ with events before event e, set $\mathcal{E}_{\geq e}$ with events after e, set $\mathcal{E}_{<e'}$ with events before e', set $\mathcal{E}_{\geq e'}$ with events after e', as shown in Fig. 15(b). If the trip is operated from events e to e' (i.e., $z_{e_{\text{ori}}(g),e} + \sum_{e'''\in\mathcal{E}_e^-} z_{e''',e} = 1$ and $z_{k,e',e_{\text{des}}(g)} + \sum_{e''\in\mathcal{E}_e^+} z_{e',e''} = 1$), events before e

(i.e., $e'' \in \mathcal{E}_{<e}$) and after e' (i.e., $e'' \in \mathcal{E}_{>e'}$) are canceled (i.e., constraints (40a) and (40c)), while all events between events e and e' (i.e., $e'' \in \mathcal{E}_{\geq e} \cap \mathcal{E}_{\leq e'}$) are operated (i.e., constraints (40b)).

$$\sum_{t \in \mathcal{T}_e} x_{e,t} \leq \sum_{k \in \mathcal{K}} \left(z_{k,e_{\text{ori}}(g),e_0(r)} + \sum_{e' \in \mathcal{E}_r^T \cap \mathcal{E}_{\leq e}} \sum_{e'' \in \mathcal{E}_e^-} z_{k,e'',e'} \right) \quad g = g(k), \forall e \in \mathcal{E}_r, r \in \mathcal{R} \tag{41}$$

Lastly, constraints (41) link event activation to rolling stock assignment, ensuring that each active event (i.e., $\sum_{t \in \mathcal{T}_e} x_{e,t} = 1$) is covered either by a rolling stock unit from a depot (i.e., $\sum_{k \in \mathcal{K}} z_{k,e_{\text{ori}}(g),e_0(r)} = 1$) or by a turn-around activity from a preceding event (i.e., $\sum_{k \in \mathcal{K}} \sum_{e' \in \mathcal{E}_r^T \cap \mathcal{E}_{\leq e}} \sum_{e'' \in \mathcal{E}_e^-} z_{k,e'',e'} = 1$) .

Group III: Passenger flow management constraints

$$\begin{cases} \sum_{a \in \mathcal{A}_e^O} p_{a,s} \leq \text{Cap}_{r(e)} \sum_{t \in \mathbb{Z}_{[\underline{t}_s, \overline{t}_s]}} x_{e,t} & (a) \\ \sum_{a \in \mathcal{A}_e^{\text{pass}}} \sum_{s' \in \mathbb{Z}_{[0,s]} \cap \mathcal{S}_{e'}} p_{e',e'',s'} \leq \text{Cap}_{r(e)} \sum_{t \in \mathbb{Z}_{[\underline{t}_s, \overline{t}_s]}} x_{e,t} & (b) \\ \sum_{a \in \mathcal{A}_e^D} \sum_{s' \in \mathbb{Z}_{[0,s]} \cap \mathcal{S}_{e'}} p_{a,s'} \leq \text{Cap}_{r(e)} \sum_{t \in \mathbb{Z}_{[\underline{t}_s, \overline{t}_s]}} x_{e,t} & (c) \end{cases} \quad \forall s \in \mathcal{S}, e \in \mathcal{E}^D \tag{42}$$

Constraints (42) link rolling stock schedules to passenger flow assignments by imposing capacity limits at departure events. These constraints are applied to three passenger activity types: boarding from event e (i.e., $a \in \mathcal{A}_e^O$), passing through event e (i.e., $a \in \mathcal{A}_e^{\text{pass}}$), and alighting at event e (i.e., $a \in \mathcal{A}_e^D$), which are equivalent to constraints (10). Combined with constraints (6)–(8), (11)–(12), and (15)–(17), we can ensure consistent period-based passenger flow management.

Group IV: Auxiliary constraints and ranges of the decision variables

$$n_g = \sum_{k \in \mathcal{K}_g} \sum_{r \in \mathcal{R}} z_{k,e_{\text{ori}}(g),e_0(r)} \quad \forall g \in \mathcal{G} \tag{43}$$

$$v_{k,e,e'} \geq \sum_{t \in \mathcal{T}_{e'}} t x_{e',t} - \sum_{t \in \mathcal{T}_e} t x_{e,t} - M_{e,e'} \left(2 - \sum_{t \in \mathcal{T}_{e'}} x_{e',t} - \sum_{t \in \mathcal{T}_e} x_{e,t} \right) \quad \forall (e,e') \in \mathcal{A}^R \cup \mathcal{A}^D, k \in \mathcal{K} \tag{44}$$

$$v_{k,e,e'} \geq \sum_{t \in \mathcal{T}_{e'}} t x_{e',t} - \sum_{t \in \mathcal{T}_e} t x_{e,t} - M_{e,e'} \left(1 - \sum_{k \in \mathcal{K}} z_{k,e,e'} \right) \quad \forall (e,e') \in \mathcal{A}^C, k \in \mathcal{K} \tag{45}$$

$$x_{e,t} \in \{0,1\} \quad \forall t \in \mathcal{T}_e, e \in \mathcal{E} \tag{46}$$

$$z_{k,e,e'} \in \{0,1\} \quad \forall (e,e') \in \mathcal{A}^C, k \in \mathcal{K} \tag{47}$$

$$v_{k,e,e'} \in \mathbb{Z}^+ \quad \forall (e,e') \in \mathcal{A}, k \in \mathcal{K} \tag{48}$$

$$n_g \in \mathbb{Z}_{[0,|\mathcal{K}_g|]} \quad \forall g \in \mathcal{G} \tag{49}$$

Constraints (43) compute the total number n_g of rolling stock units used in each group g. Constraints (44) and (45) calculate actual running/dwelling times and turn-around times, respectively, for activities (e, e') assigned to rolling stock units k. Finally, constraints (46)–(49) define the ranges of all decision variables.

$$\text{M5:} \begin{cases} \min \sum_{g\in\mathcal{G}} c_g^1 n_g + \sum_{k\in\mathcal{K}} \sum_{(e,e')\in\mathcal{A}} c^2_{g(k),e,e'} v_{k,e,e'} + \sum_{i\in\mathcal{I}} \sum_{s\in\mathcal{S}_i} \left(c^D_{i,s} \widehat{w}_{i,s} + c^C_{i,s} \int_{\underline{t}_s}^{\bar{t}_s} \widehat{\mu}_{i,s} dt \right) + \sum_{(i,j)\in\mathcal{OD}} c^{\text{re}}_{i,j} p^{\text{re}}_{i,j} & (a) \\ \text{s.t. Constraints (6)–(8), (11)–(12), (15)–(17), (33)–(49)} & (b) \end{cases} \tag{50}$$

With Eqs. (1)–(17) and (33)–(49), we summarize the primal mathematical model M5 in Eq. (50) for the RSP-STPM. A general-purpose solver (e.g., CPLEX) can solve the primal model M5 for small-scale cases (see Section E for the numerical experiments). However, it remains challenging to solve the primal model M5 for large-scale instances due to the high number of coupling constraints and complex integer variables, making the equivalent path-based model M1 a necessary alternative.

C. Passenger delay and total passenger flow volume calculations

Based on Eqs. (1)–(17), we develop a general period-based passenger flow management model that accommodates arbitrary passenger arrival rate function $\lambda_{i,s}(t)$ and control rate function $\mu_{i,s}(t)$ for station i and period s. In practice, these functions often display some characteristics that enable analytical derivation of the passenger delay calculations for specific functional forms. For example, passenger arrival rates commonly follow "unimodal" or "bimodal" distributions over the daily horizon, reflecting morning and evening peak hours. A unimodal distribution can often be approximated by a quadratic function, while passenger flow control is typically implemented at a constant entry rate, e.g., by installing barriers at station entrances within a fixed period (e.g., one hour). In addition, The path-based model M1 is not directly solvable by general-purpose solvers because it is challenging to determine the queue start time $t^{QS}_{i,s}$ and end time $t^{QE}_{i,s}$ in Eq. (12), which is not straightforward available from the input data. However, when explicit passenger arrival and control rate functions are given, Eq. (12) can be reformulated analytically into a simple closed-form expression requiring fewer parameters and decision variables.

Given this observation, we focus here on two propositions based on a linear or quadratic passenger arrival rate function and a constant passenger control rate function, where the total passenger delay $\widehat{w}_{i,s}$ for station i and period s can be computed explicitly. Lastly, to accommodate LP-based decomposition algorithms that require linear formulations, we further approximate the non-linear delay calculation with tangential linear constraints (13) and assess the deviation between the exact and approximated results.

PROPOSITION 2. *The passenger arrival rate $\lambda_{i,s}(t)$ is assumed to be the linear function in Eq. (51a) and shown in Fig. 16(a) for station i and period s, where $\widehat{\lambda}_{i,s} \geq 0$ is the shape parameter and $t^{Q\max}_{i,s}$ is the time of maximum queue length. Let $\lambda^{\max}_{i,s}$ be the maximum arrival rate at the queue start time $t^{QS}_{i,s}$ in*

Eq. (51b). The control rate is constant function $\mu_{i,s}(t) = \widehat{\mu}_{i,s}$ with $0 \le \widehat{\mu}_{i,s} \le \mu_{i,s}^{\max}$, where $\mu_{i,s}^{\max}$ denotes the maximum passenger control rate determined by the infrastructure capacity of station i at period s.

$$\begin{cases} \lambda_{i,s}(t) = -\widehat{\lambda}_{i,s}\left(t - t_{i,s}^{Q\max}\right) + \mu_{i,s} & (a) \\ \lambda_{i,s}(t_{i,s}^{QS}) = \lambda_{i,s}^{\max} & (b) \end{cases} \quad \underline{t}_s \le t_{i,s}^{QS} \le t \le t_{i,s}^{ES} \le \bar{t}_s, \forall s \in \mathcal{S}_i, i \in \mathcal{I} \tag{51}$$

Without loss of generality, we assume that there is no waiting passenger flow volume at station i at the start of period s (i.e., $W_{i,s} = 0$). By substituting Eq. (51) *into Eq.* (12)*, we obtain the analytical expression for the total delay $\widehat{w}_{i,s}$ at station i during period s, as given in Eq.* (52)*. Similarly, the total arrival passenger flow volume within the time window $[\underline{t}_s, \bar{t}_s]$ of period s can be explicitly derived, as given in Eq.* (53)*.*

$$\widehat{w}_{i,s} = \frac{2}{3} \cdot \frac{\left(\lambda_{i,s}^{\max} - \widehat{\mu}_{i,s}\right)^3}{\widehat{\lambda}_{i,s}^2} \quad \forall s \in \mathcal{S}_i, i \in \mathcal{I} \tag{52}$$

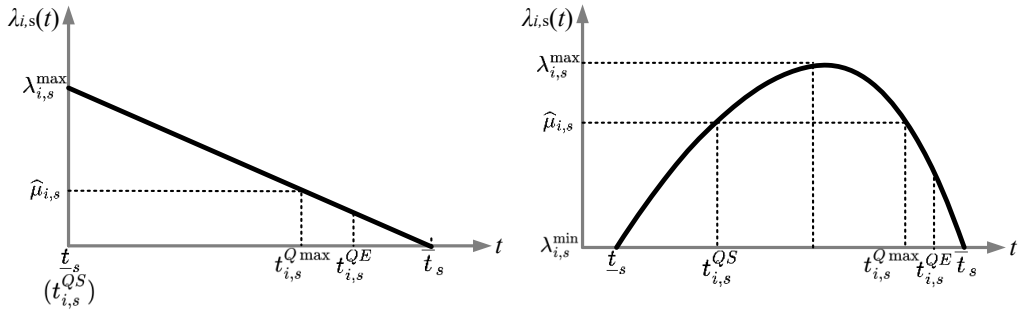
$$P_{i,j}^{in} = \sum_{s \in \mathcal{S}_i} \frac{1}{2} \lambda_{i,s}^{\max}\left(\bar{t}_s - \underline{t}_s\right) \cdot \gamma_{i,j,s} \quad \forall i, j \in \mathcal{I}, i \ne j \tag{53}$$

Proof: For simplicity, we assume that there is no waiting passenger flow volume at station i at the start of period s (i.e., $W_{i,s} = 0$). By substituting Eq. (51) into the passenger delay calculation in Eq. (12), we obtain the total passenger delay at station i in period s, as given in Eq. (54). We note that the passenger queue dissipates at time $t_{i,s}^{QE}$, i.e., $q_{i,s}(t_{i,s}^{QE}) = 0$. This leads to the relationship $t_{i,s}^{Q\max} = \frac{1}{2}\left(t_{i,s}^{QS} + t_{i,s}^{QE}\right)$, where $t_{i,s}^{QS}$, $t_{i,s}^{Q\max}$ and $t_{i,s}^{QE}$ denote the queue start, peak, and end times, respectively. By substituting this relationship into Eq. (54b), together with the maximum arrival rate condition in Eq. (51b), we simplify the expression for total passenger delay to the closed form in Eq. (54e). This final form of $\widehat{w}_{i,s}$ depends only on the maximum passenger arrival rate $\lambda_{i,s}^{\max}$, the shape parameter $\widehat{\lambda}_{i,s}$, and the passenger control rate $\widehat{\mu}_{i,s}$. It can thus be embedded directly into the objective function (12) of model without introducing additional passenger-related decision variables or constraints.

$$\begin{aligned}
\widehat{w}_{i,s} &= \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} q_{i,s}(t)dt = \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} \int_{t_{i,s}^{QS}}^{t} [\lambda_{i,s}(\tau) - \widehat{\mu}_{i,s}]d\tau dt = \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} \int_{t_{i,s}^{QS}}^{t} -\widehat{\lambda}_{i,s}\left(\tau - t_{i,s}^{Q\max}\right)d\tau dt & (a) \\
&= \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} -\frac{\widehat{\lambda}_{i,s}}{2}\left(t - t_{i,s}^{QS}\right)\left(t + t_{i,s}^{QS} - 2t_{i,s}^{Q\max}\right)dt \leftarrow \text{Substitute } q_{i,s}(t_{i,s}^{QE}) = 0, t_{i,s}^{Q\max} = \frac{1}{2}\left(t_{i,s}^{QS} + t_{i,s}^{QE}\right) & (b) \\
&= \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} -\frac{\widehat{\lambda}_{i,s}}{2}\left(t - t_{i,s}^{QS}\right)\left(t - t_{i,s}^{QE}\right)dt & (c) \\
&= \frac{\widehat{\lambda}_{i,s}\left(t_{i,s}^{QS} - t_{i,s}^{QE}\right)^3}{12} \leftarrow \text{Substitute } \lambda_{i,s}(t_{i,s}^{QS}) = \lambda_{i,s}^{\max} & (d) \\
&= \frac{2}{3} \cdot \frac{\left(\lambda_{i,s}^{\max} - \widehat{\mu}_{i,s}\right)^3}{\widehat{\lambda}_{i,s}^2} & (e)
\end{aligned} \tag{54}$$

Lastly, the total passenger flow volume for OD pair (i, j) in constraints (11) can be explicitly determined by integrating the arrival rate function in Eq. (51) over the time window $[\underline{t}_s, \bar{t}_s]$ of period s, yielding Eq. (55).

$$P_{i,j}^{\text{in}} = \sum_{s \in \mathcal{S}_i} \int_{\underline{t}_s}^{\bar{t}_s} \left[-\widehat{\lambda}_{i,s}\left(t - t_{,i,s}^{Q\max}\right) + \widehat{\mu}_{i,s}\right] dt \cdot \gamma_{i,j,s} = \frac{1}{2}\lambda_{i,s}^{\max}\left(\bar{t}_s - \underline{t}_s\right) \cdot \gamma_{i,j,s} \quad \forall i \in \mathcal{I} \tag{55}$$



(a) Linear function (b) Quadratic function

Figure 16 Illustration of the passenger arrival rate functions

Next, we extend the passenger arrival rate function from a linear to a quadratic form in Proposition 3, which better reflects the "unimodal" pattern of passenger distributions during peak hours.

PROPOSITION 3. *The passenger arrival rate $\lambda_{i,s}(t)$ is assumed to be the quadratic function in Eq. (56a) and Fig. 16(b) for station i and period s, where $\widehat{\lambda}_{i,s} \geq 0$ is the shape parameter and $t_{i,s}^{Q\max}$ is the time of maximum passenger queue length. We impose two bounding conditions on the maximum arrival rate $\lambda_{i,s}^{\max}$ at peak time $\frac{t_{i,s}^{QS}+t_{i,s}^{Q\max}}{2}$ in Eq. (56b), as well as the minimum arrival rate $\lambda_{i,s}^{\min}$ at the start time $\underline{t}_s$ and the end time $\bar{t}_s$ of the period in Eq. (56c). As in Proposition 2, the passenger control rate function $\mu_{i,s}(t) = \mu_{i,s}$ is assumed constant.*

$$\begin{cases} \lambda_{i,s}(t) = -\widehat{\lambda}_{i,s}\left(t - t_{i,s}^{QS}\right)\left(t - t_{i,s}^{Q\max}\right) + \widehat{\mu}_{i,s} & (a) \\ \lambda_{i,s}\left(\frac{t_{i,s}^{QS}+t_{i,s}^{Q\max}}{2}\right) = \lambda_{i,s}^{\max} & (b) \\ \lambda_{i,s}\left(\underline{t}_s\right) = \lambda_{i,s}\left(\bar{t}_s\right) = \lambda_{i,s}^{\min} & (c) \end{cases} \quad \underline{t}_s \leq t_{i,s}^{QS} \leq t \leq t_{i,s}^{QE} \leq \bar{t}_s, \forall s \in \mathcal{S}_i, i \in \mathcal{I} \tag{56}$$

By substituting Eq. (56) *into Eq.* (12)*, we obtain the analytical expression for total passenger delay $\widehat{w}_{i,s}$ at station i during period s, as show in Eq.* (57)*. Similarly, the total passenger arrival flow volume within the time window $[\underline{t}_s, \bar{t}_s]$ of period s can be explicitly derived, as given Eq.* (58)*.*

$$\widehat{w}_{i,s} = \frac{27}{4}\frac{\left(\lambda_{i,s}^{\max} - \widehat{\mu}_{i,s}\right)^2}{\widehat{\lambda}_{i,s}} \quad \forall s \in \mathcal{S}_i, i \in \mathcal{I} \tag{57}$$

$$P_{i,j}^{in} = \sum_{s \in \mathcal{S}_i}\left(\frac{2}{3}\lambda_{i,s}^{\max} + \lambda_{i,s}^{\min}\right)\left(\bar{t}_s - \underline{t}_s\right) \cdot \gamma_{i,j,s} \quad \forall i, j \in \mathcal{I}, i \neq j \tag{58}$$

Proof: To calculate the passenger delay $\widehat{w}_{i,s}$, we substitute Eq. (56) into the passenger delay calculation in Eq. (12), yielding the total passenger delay at station i and period s, as given in Eq. (59). By applying the passenger dispatching condition $q_{i,s}(t_{i,s}^{QE}) = 0$ at the queue end time $t_{i,s}^{QE}$, we derive the relationship $t_{i,s}^{Q\max} = \frac{1}{3}t_{i,s}^{QS} + \frac{2}{3}t_{i,s}^{QE}$ linking the queue start time $t_{i,s}^{QS}$, peak time $t_{i,s}^{Q\max}$, and end time $t_{i,s}^{QE}$. Then, using the

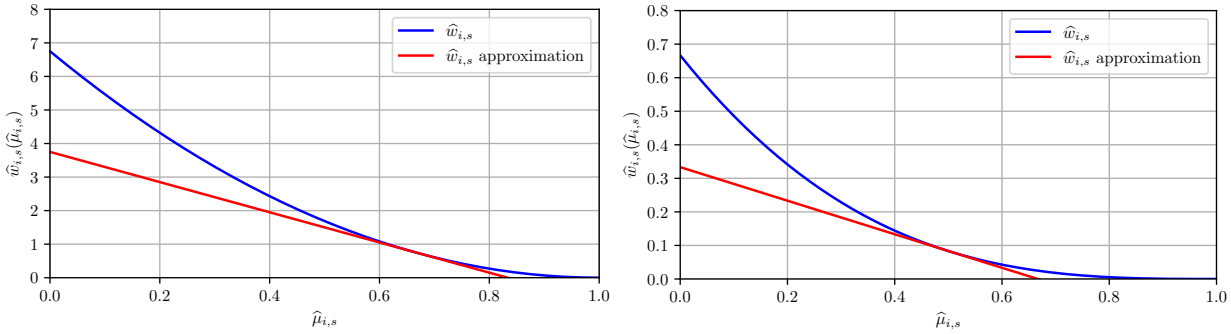
maximum arrival rate condition in Eq. (56b), the passenger delay $\widehat{w}_{i,s}$ is reduced to the closed form in Eq. (59e).

$$\begin{aligned}
\widehat{w}_{i,s} &= \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} q_{i,s}(t)dt = \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} \int_{t_{i,s}^{QS}}^{t} [\lambda_{i,s}(\tau) - \widehat{\mu}_{i,s}]d\tau dt = \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} \int_{t_{i,s}^{QS}}^{t} -\widehat{\lambda}_{i,s}\left(\tau - t_{i,s}^{QS}\right)\left(\tau - t_{i,s}^{Q\max}\right)d\tau dt & (a)\\
&= \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} -\widehat{\lambda}_{i,s}\left(t - t_{i,s}^{QS}\right)^2\left(\frac{t_{i,s}^{Q\max} - t_{i,s}^{QS}}{2} - \frac{t - t_{i,s}^{QS}}{3}\right)dt \leftarrow \text{Substitute } t_{i,s}^{Q\max} = \frac{1}{3}t_{i,s}^{QS} + \frac{2}{3}t_{i,s}^{QE} & (b)\\
&= \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} -\widehat{\lambda}_{i,s}\left(t - t_{i,s}^{QS}\right)^2\left(t - t_{i,s}^{QE}\right)dt = \int_{t_{i,s}^{QS}}^{t_{i,s}^{QE}} -\widehat{\lambda}_{i,s}\left(t - t_{i,s}^{QS}\right)^2\left(t - t_{i,s}^{QE}\right)dt & (c)\\
&= \frac{\widehat{\lambda}_{i,s}\left(t_{i,s}^{QE} - t_{i,s}^{QS}\right)^4}{12} \leftarrow \text{Substitute } \lambda_{i,s}\left(\frac{t_{i,s}^{QS} + t_{i,s}^{Q\max}}{2}\right) = \lambda_{i,s}^{\max} \text{ and } t_{i,s}^{Q\max} = \frac{1}{3}t_{i,s}^{QS} + \frac{2}{3}t_{i,s}^{QE} & (d)\\
&= \frac{27}{4}\frac{\left(\lambda_{i,s}^{\max} - \widehat{\mu}_{i,s}\right)^2}{\widehat{\lambda}_{i,s}} & (e)
\end{aligned} \tag{59}$$

Lastly, the total passenger flow volume for OD pair (i,j) in constraints (11) can be explicitly determined by integrating the arrival rate function in Eq. (56) over the time window $[\underline{t}_s, \overline{t}_s]$ of period s, yielding Eq. (60).

$$\begin{aligned}
P_{i,j}^{\text{in}} &= \sum_{s\in\mathcal{S}_i}\int_{\underline{t}_s}^{\overline{t}_s} [-\widehat{\lambda}_{i,s}(t - \underline{t}_s)(t - \overline{t}_s) + \widehat{\mu}_{i,s}]dt \cdot \gamma_{i,j,s} & (a)\\
&= \left(\frac{\widehat{\lambda}_{i,s}}{6}\left(\overline{t}_s - \underline{t}_s\right)^2 + \widehat{\mu}_{i,s}\right)\left(\overline{t}_s - \underline{t}_s\right)\cdot\gamma_{i,j,s} \leftarrow \text{Substitute Eqs. (56b)–(56c)} \quad (b)\ \forall i,j\in\mathcal{I}, i\neq j & (60)\\
&= \sum_{s\in\mathcal{S}_i}\left(\frac{2}{3}\lambda_{i,s}^{\max} + \lambda_{i,s}^{\min}\right)\left(\overline{t}_s - \underline{t}_s\right)\cdot\gamma_{i,j,s} & (c)
\end{aligned}$$

Example: Fig. 17 illustrates the exact and approximate passenger delay calculations for two explicit passenger arrival rate functions, including a linear form $\lambda_{i,s}(t) = -t + 1$, $t \in [0,1]$ and a quadratic form $\lambda_{i,s}(t) = -(t-1)^2 + 1$, $t \in [0,2]$. First, by substituting these functions into Eqs. (52) and (57), we obtain the exact passenger delays $\widehat{w}_{i,s}(\widehat{\mu}_{i,s}) = \frac{2(1-\widehat{\mu}_{i,s})^3}{3}$, $\widehat{\mu}_{i,s} \in [0,1]$ for the linear case, and $\widehat{w}_{i,s}(\widehat{\mu}_{i,s}) = \frac{27(1-\widehat{\mu}_{i,s})^2}{4}$, $\widehat{\mu}_{i,s} \in [0,1]$ for the quadratic case, shown as the blue curves in Fig. 17. In both cases, passenger delays decrease monotonically as the passenger flow control rate increases, reaching zero when $\widehat{\mu}_{i,s}$ equals to the maximum passenger arrival rate $\lambda_{i,s}^{\max}$. Using Eq. (13), we then approximate the passenger delay at the average arrival rate $\overline{\lambda}_{i,s}$. The resulting red curves provide a close approximation to the nonlinear constraints (52) and (57) in the neighborhood of the average arrival rate $\overline{\lambda}_{i,s}$.



(a) Linear function (b) Quadratic function

Figure 17 Passenger delay approximations with flow control rates

D. Pseudocodes of labeling algorithms

Algorithm 1 ***Single-path labeling algorithm for solving the pricing sub-problem***

Input: *Time-dependent event-activity network $\mathcal{G}^T = (\mathcal{E}^T, \mathcal{A}^T)$ and dual solutions $\boldsymbol{\pi}^1$–$\boldsymbol{\pi}^5$ corresponding to the constraints of the relaxed BMP M3.*

Output: *The optimal event-activity path p_g^* for each rolling stock unit group g.*

1: # parallel for *(each rolling stock group g)*

2: ***Step 1 (Initialization):***

3: • *Generate labels $l_{e,t}$ for all events e and associated start times $t \in \mathcal{T}_e$.*

4: • *Initialize label costs: $RC_{e_{ori(g)},t_0} = c_g^1 - \pi_g^1$ for the depot source $(e_{ori(g)}, t_0)$, $RC_{e,t} = \infty$ for all other vertices.*

5: • *Sequence labels from the depot source using topological sorting.*

6: ***Step 2 (Forward dynamic label updating):***

7: ***for*** *each label $l_{e,t}$ and $l'_{e',t'}$, where $l \neq l'$* ***do***

8: ***if*** *$e' \in \mathcal{E}_e^+$ and $T_{e,e'}^{\min} \leq t' - t \leq T_{e,e'}^{\max}$* ***then***

9: • *Check branching rules*

10: ***if*** *$\overline{x}_{e,t} = 0$ or $\overline{x}_{e',t'} = 0$ or $\overline{z}_{g,e,e'} = 0$* ***then***

11: ***Continue***

12: ***end if***

13: • *Compute $RC'_{e',t'} = RC_{e,t} + \overline{c}_{e,e',t,t'}$ using Eqs. (27)–(28).*

14: ***if*** *$RC'_{e',t'} < RC_{e',t'}$* ***then***

15: • *Update $RC_{e',t'} \leftarrow RC'_{e',t'}$ and $Pred_{e',t'} \leftarrow (e,t)$.*

16: ***end if***

17: ***end if***

18: ***end for***

19: ***Step 3 (Backtrace):***

20: • *For each group g, trace back from the depot sink $(e_{des}(g), |\mathcal{T}|)$ to the depot source $(e_{ori}(g), t_0)$ using $Pred_{e,t}$ and $RC_{e,t}$.*

21: • *Record the resulting optimal column and add it to the column pool $\overline{\mathcal{P}}$ for subsequent iterations.*

Algorithm 2 ***Multi-path labeling algorithm for column enumeration***

Input: *Time-dependent event-activity network $\mathcal{G}^T = (\mathcal{E}^T, \mathcal{A}^T)$ and dual solutions $\boldsymbol{\pi}^1$–$\boldsymbol{\pi}^5$ from the final iteration of the Benders decomposition algorithm for the constraints of the relaxed BMP M3 .*

Output: *Enumerated column pool $\overline{\mathcal{P}}^E$ containing event-activity paths with reduced costs smaller than the optimality gap $z_{UB} - z_{LB}$.*

1: # parallel for *(each rolling stock group g)*

2: ***Step 1 (Initialization):***

3: • *Create the initial label $l_{e_{ori}(g),t_0}$ at the depot source $(e_{ori}(g), t_0)$ with cost $RC_{l_{e_{ori}(g),t_0}} = c_g^1 - \pi_g^1$.*

4: • *Insert this initial label into the deque $LabelDeque$.*

5: ***Step 2 (Label extension and update):***

6: ***while*** *there exists an extendable label in $LabelDeque$* ***do***

7: $l_{e,t} \leftarrow LabelDeque$.POP( )

8: ***for*** *each $e' \in \mathcal{E}_e^+$ and each $t' \in \mathcal{T}_{e'}$* ***do***

9: • *Check turn-around time*

10: ***if*** $T_{e,e'}^{\min} \leq t' - t \leq T_{e,e'}^{\max}$ ***then***

11: • *Check branching rules*

12: ***if*** $\overline{x}_{e,t} = 0$ *or* $\overline{x}_{e',t'} = 0$ *or* $\overline{z}_{g,e,e'} = 0$ ***then***

13: ***Continue***

14: ***end if***

15: • *Compute $RC'_{e',t'} = RC_{e,t} + \overline{c}_{e,e',t,t'}$ using Eqs.* (27)–(28).

16: • *Extend the path and generate a new label:* $l_{e',t'} \leftarrow l_{e,t}$.EXTEND$(e', t', RC'_{e',t'})$.

17: • $LabelDeque$.PUSHBACK$(l_{e',t'})$

18: ***end if***

19: ***end for***

20: ***end while***

21: ***Step 3 (Feasible column filtering):***

22: • *For rolling stock group g, extract all feasible paths from $(e_{ori}(g), t_0)$ to $(e_{des}(g), t_{|\mathcal{T}|})$ with reduced costs below $z_{UB} - z_{LB}$, and add them to the enumerated column pool $\overline{\mathcal{P}}^E$.*

E. Small-scale case study

We consider a small-scale urban rail line, shown in Fig. 18, consisting of three stations (S_1–S_3) and two depots (D_1 and D_2) connected to stations S_1 and S_3, respectively. Running times are labeled on the physical segments in Fig. 18, and turn-around times between adjacent trips are set within [5 min, 30 min]. Within the planning interval [0 min, 60 min], 8 or 10 trips are covered by rolling stock units. Two rolling stock groups are considered in the small-scale case study, i.e., rolling stock units in group 1 start from depot D_1 and end at depot D_1, rolling stock units in group 2 start from depot D_2 and end at depot D_2. Following Assumption 1, the number of rolling stock units in each group ($|\mathcal{K}_g|$) is non-binding. For computational convenience in solving the primal model M5 (see Appendix B) with the CPLEX solver, we set the maximum number of rolling stock units $\frac{|\mathcal{R}|}{2}$ to ensure the model feasibility, i.e., one unit covers two trips in opposite directions while starting and ending at the same depot. Following the parameter settings adopted of the real-world case study in Section 6, the fixed cost associated with assigning a rolling stock unit is set to ¥1,000. In addition, the variable costs for running/dwelling operations and turn-around operations are specified as ¥10 per minute and ¥5 per minute, respectively.

For convenience, we consider two types of passenger arrival rate distributions at all stations $i \in \mathcal{I}$ and all periods $s \in \mathcal{S}$: a linear function $\lambda_{i,s}(t) = -\frac{5}{3}(t-60)$ over $t \in [0, 60]$ in Fig. 19(a) and a quadratic function $\lambda_{i,s}(t) = -\frac{1}{9}t(t-60)$ in Fig. 19(b) over $t \in [0, 60]$. Then, we represent the same OD proportions (i.e., parameter $\gamma_{s,i,j}$ from stations i to j in period s) for all periods $s \in \mathcal{S}$ by the 3×3 matrix $\Upsilon = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{pmatrix}$. Based on these rolling stock and passenger flow inputs, the corresponding path-based model M1 and primal model M5 are formulated.

E.1. Experimental results

In this section, we consider a set of 16 instances and conduct a total of 64 experiments under different settings. These instances are solved using both the CPLEX solver and the proposed BDCG algorithm. For convenience, we introduce a notation with the format $(P, |\mathcal{S}|, |\mathcal{R}|, |\mathcal{E}^T|)$ to denote the experiment features, in which "P" indicates the passenger arrival rate function type ("L" for the linear function in Fig. 19(a), and "Q" for the quadratic function $\lambda_{i,s}(t)$ in Fig. 19(b)), $|\mathcal{S}|$ is the number of periods, $|\mathcal{R}|$ is the number of trips, $|\mathcal{E}^T|$ is the number of available events allowing turn-around operations. The latter significantly affects problem complexity due to constraints such as Eq. (40) in primal model M5. Lastly, we set the termination conditions with a maximum of 60 min (3600s) computing time.

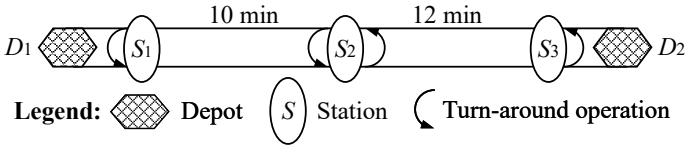


Figure 18 Urban rail line

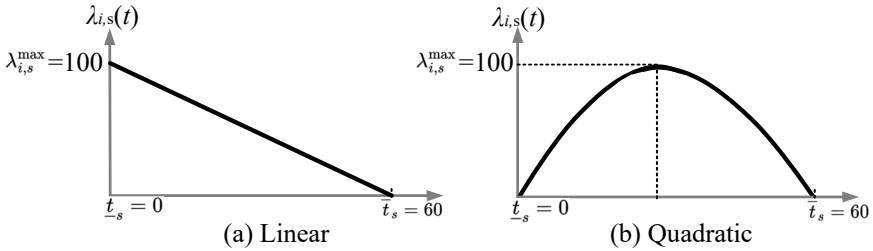


Figure 19 Passenger arrival rate distributions

Table 5 Computational results with different techniques in small-scale instances

ID	$(P, \|\mathcal{S}\|, \|\mathcal{R}\|, \|\mathcal{E}^T\|)$	CPLEX				B&B				CER					All techniques						
		LB	UB	Gap (%)	CT(s)	LB	UB	Gap (%)	CT(s)	#CG	LB	UB	Gap (%)	CT(s)	#CG	#CE	#CR	LB	UB	Gap (%)	CT(s)
1	(L,2,8,16)	7070	7070	0.0	6.6	7070	7070	0.0	2.4	38	7070	7070	0.0	2.4	14	1153	144	7070	7070	0.0	1.5
2	(L,2,8,24)	7070	7070	0.0	161.7	6630	7070	6.2	73.3	790	6848	7230	5.3	492.3	59	15713	1964	6822	7070	3.5	87.1
3	(L,2,10,20)	6693	8545	21.7	3600.0	8545	8545	0.0	5.0	32	8545	8545	0.0	4.4	18	0	0	8545	8545	0.0	2.6
4	(L,2,10,30)	6205	-	-	3600.0	7665	8105	5.4	475.1	1765	7885	8325	5.3	675.7	101	24551	2455	7938	8105	2.1	107.2
5	(L,3,8,16)	8067	8867	9.0	3600.0	8867	8867	0.0	1.4	20	8867	8867	0.0	1.3	14	0	0	8867	8867	0.0	1.0
6	(L,3,8,24)	6210	8780	29.3	3600.0	7553	8960	15.7	179.7	1563	7627	8780	13.1	243.7	125	15713	1964	7627	8780	13.1	58.7
7	(L,3,10,20)	7387	10342	28.6	3600.0	10342	10342	0.0	5.0	32	10342	10342	0.0	4.3	18	0	0	10342	10342	0.0	2.6
8	(L,3,10,30)	7202	-	-	3600.0	8588	15422	44.3	3600.0	6656	8588	9948	13.7	2563.9	201	24381	2455	8662	9548	9.3	377.7
9	(Q,2,8,16)	6845	6845	0.0	9.0	6845	6845	0.0	1.4	20	6845	6845	0.0	1.2	14	0	0	6845	6845	0.0	0.9
10	(Q,2,8,24)	5235	6845	23.5	3600.0	6405	6845	6.4	66.4	571	6789	6845	0.8	112.0	100	15617	1964	6845	6845	0.0	39.8
11	(Q,2,10,20)	6895	8320	17.1	3600.0	8320	8320	0.0	5.0	32	8320	8320	0.0	4.3	18	0	0	8320	8320	0.0	2.7
12	(Q,2,10,30)	5981	8185	26.9	3600.0	7440	11060	32.7	482.4	1122	7660	8870	13.6	419.8	121	24381	2455	7660	8185	6.4	111.7
13	(Q,3,8,16)	10938	10938	0.0	607.5	10938	10938	0.0	1.4	20	10938	10938	0.0	1.3	14	0	0	10938	10938	0.0	0.9
14	(Q,3,8,24)	8006	9871	18.9	3600.0	9260	9871	6.2	147.7	1557	9413	9871	4.6	179.9	161	15665	1964	9431	9871	4.5	88.2
15	(Q,3,10,20)	12396	12413	0.1	1069.17	12413	12413	0.0	5.0	32	12413	12413	0.0	4.3	18	0	0	12413	12413	0.0	2.7
16	(Q,3,10,30)	9006	12458	27.7	3600.0	10466	12319	15.0	1777.03	4469	10466	12319	15.0	1988.1	188	24491	2455	10466	11346	7.8	305.0
Ave.		7575.4	-	-	2590.9	8584.2	9562.0	8.3	426.8	1169.9	8663.5	9095.5	4.5	418.7	74.0	10104.1	1113.8	8674.4	8943.1	2.9	74.4

Notes for Table 5:

(1) ID: Instance ID; LB: Lower bound; UB: Upper bound; Gap = $\frac{\text{UB - LB}}{\text{UB}} \times 100\%$; CT: Computing times until incumbent bounds are obtained with no further improvement; #CG: Number of generated columns; #CE: Number of enumerated columns; #CR: Number of removed columns.

(2) Passenger arrival rate function (P): Linear function $\lambda_{i,s}(t)$ in Fig. 19(a) (L); Quadratic function $\lambda_{i,s}(t)$ in Fig. 19(b) (Q).

(3) B&B: BDCG algorithm with B&B technique; CER: BDCG algorithm with both B&B and CER techniques; All techniques: BDCG algorithm with B&B, CER, and acceleration techniques. Solution status: Feasible solution cannot be found within the maximum computing time ("-"). Ave.: Average performance for instances.

The computational results are presented in Table 5 for all instances, including different solution approaches, lower bounds, upper bounds, gaps, and incumbent computing times. Specifically, the BDCG algorithm with all techniques clearly outperforms the CPLEX solver and can obtain optimal or near-optimal solutions for more instances. With the same termination condition, the average lower bound and incumbent computing time by using the BDCG algorithm with all techniques (i.e., 8674.4 and 74.4 s in column "All techniques") are improved by 14.51% and 97.13% compared to those obtained by the CPLEX solver (i.e., 7575.4 and 2590.9 s in column "CPLEX"), respectively. To further explore the computational performance of the proposed algorithms, we test the instances in Table 5 using the BDCG algorithm with the B&B and CER techniques. Specifically, the average lower bound, gap, and incumbent solution computing time by using the CER technique (i.e., 8663.5, 4.5%, and 418.7 s in column "CER") are improved by 0.92%, 45.78%, and 1.90% with respect to the traditional B&B technique in the BDCG algorithm (i.e., 8584.2, 8.3%, and 426.8 s in column "B&B"). Furthermore, with all techniques in the BDCG algorithm, the average lower bound, gap, and incumbent solution computing time (i.e., 8674.4, 2.9%, and 74.4 s in column "All techniques") are improved by 1.05%, 65.06%, and 82.57% with respect to the traditional B&B method in the BDCG algorithm (i.e., 8584.2, 8.3%, and 426.8 s in column "B&B").

Given the similar performance trends across the 16 instances, we illustrate the convergences under different BDCG enhancements for instance 16 (as an example), as shown in Fig. 20. The results indicate that the three acceleration techniques primarily improve Benders decomposition and the B&B process through dynamic constraint generation, inheritance of columns and Benders cuts, and parallel computing. Furthermore, the heuristic algorithm quickly produces feasible solutions early in the solution process, tightening upper bounds and reducing gaps more rapidly.

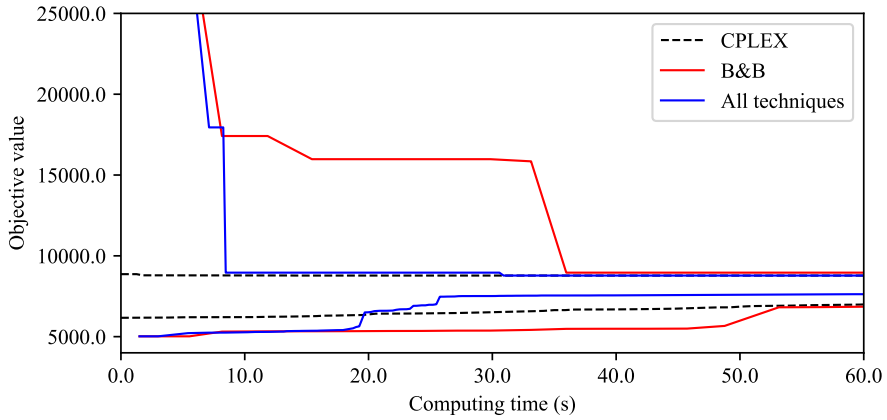


Figure 20 Illustrations of the convergence for instance 16 (as an example)

Next, we will elaborate on the performance analyses from the perspectives of passenger-oriented rolling stock operations and different passenger flow evolution periods.

E.2. Performance analysis of passenger-oriented rolling stock operations

In the passenger-oriented rolling stock scheduling problem, both the optimal number of trips and short-turn operations depend on the specific passenger flow distribution. For example, under a linear passenger flow distribution with a total of 3000 passenger flow volume, scheduling 8 trips in instance 2, where 24 events allow turn-around operations at stations S_1, S_2, and S_3, results in all trips being operated as full-length services with no short-turn operations at station S_2. Consequently, instance 2 achieves the same objective value (i.e., 7070.0) as instance 1, which has only 16 events allowing turn-around operations at stations S_1 and S_3. This outcome occurs because total trip transport capacity is lower than passenger demand, making it optimal to operate all trips as full-length services to maximize coverage, regardless of short-turn availability. By contrast, when 10 trips are scheduled (exceeding passenger demand), instance 4 (with 24 events allowing turn-around operations at stations S_1, S_2, and S_3) incorporates some short-turn operations at station S_2, yielding a lower objective value (i.e., 8105.0) than instance 3 (with 16 events allowing turn-around operations only at stations S_1 and S_3), which records a higher value of 8545.0. Similar results are observed in other instance pairs, such as instances 9–12.

E.3. Performance analysis of passenger flow evolution periods

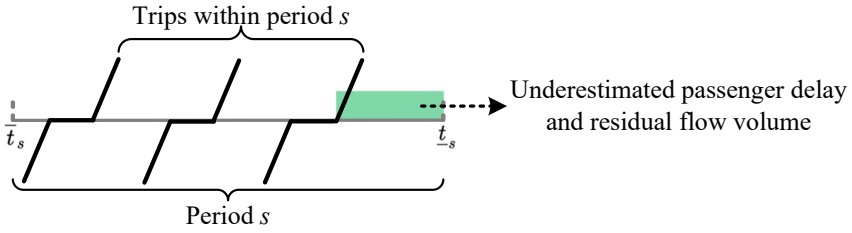


Figure 21 Illustration of underestimated passenger flow related costs

In the period-based passenger flow method, the length of passenger flow evolution periods plays a critical role in determining passenger flow control and assignment results. To examine this effect, we conduct 20 experiments in Table 6 using different period lengths. Specifically, experiments 1–11 and 12–22 partition the passenger flow evolution process into 2, 3, $\cdots$, and 60 periods, corresponding to period lengths of 30 min, 15 min, $\cdots$, and 1 min, respectively. The results reveal that the total objective value, passenger delay cost, and residual passenger penalty cost roughly increase as the number of periods grows. This is primarily because shorter period lengths reduce the underestimation of passenger delay and residual passenger flow volumes. In instances with longer periods, if passengers arrive after the last trip in a period but before the period ends (i.e., the green region in Fig. 21), they must wait until the trips in the next period. However, constraints (6) and (8) only enforce that the boarding passenger flow volume is less than the total passenger arrival flow volume within a period, not at each timestamp, allowing some passengers to be assigned to trips departing before their actual arrival time. As the period length decreases, the "green region" in Fig. 21

shrinks, leading to more accurate (i.e., higher) estimates of the total objective value, passenger delay cost, and residual passenger penalty cost. In contrast, the passenger control cost shows a consistent downward trend with more periods, reflecting that shorter periods enable more precise passenger flow control schemes and thus lower costs.

Table 6 Computational results with different passenger flow evolution periods

ID	$(P, \|\mathcal{S}\|, \|\mathcal{R}\|, \|\mathcal{E}^T\|)$	Obj.	PDC	PCC	RPC	ID	$(P, \|\mathcal{S}\|, \|\mathcal{R}\|, \|\mathcal{E}^T\|)$	Obj.	PDC	PCC	RPC
1	(L,2,8,16)	7070.0	900.0	270.0	0.0	12	(Q,2,8,16)	6845.0	675.0	270.0	0.0
2	(L,3,8,16)	8866.7	1466.7	166.7	1333.3	13	(Q,3,8,16)	10937.8	1777.8	148.9	3111.1
3	(L,4,8,16)	7116.9	337.5	129.4	750.0	14	(Q,4,8,16)	8631.6	487.5	119.1	2125.0
4	(L,5,8,16)	8989.3	2265.6	103.7	720.0	15	(Q,5,8,16)	12026.8	2663.0	87.7	3376.0
5	(L,6,8,16)	9799.2	2316.7	82.5	1500.0	16	(Q,6,8,16)	11950.0	2422.2	72.2	3555.6
6	(L,10,8,16)	9149.2	1879.2	50.0	1320.0	17	(Q,10,8,16)	11419.4	2099.5	43.9	3376.0
7	(L,12,8,16)	9899.6	2458.3	41.3	1500.0	18	(Q,12,8,16)	12263.2	2771.5	36.1	3555.6
8	(L,15,8,16)	10025.7	2519.5	32.9	1573.3	19	(Q,15,8,16)	12514.6	2850.9	28.5	3735.1
9	(L,20,8,16)	10456.8	2822.4	24.4	1710.0	20	(Q,20,8,16)	12886.5	2960.5	21.0	4005.0
10	(L,30,8,16)	10656.1	2886.7	16.1	1853.3	21	(Q,30,8,16)	13337.9	3143.3	13.7	4280.9
11	(L,60,8,16)	11044.8	3133.5	8.0	2003.3	22	(Q,60,8,16)	13820.7	3351.9	6.7	4562.1

Notes for Table 6:

(1) ID: Instance ID; Obj: Objective value; PDC: Passenger delay cost; PCC: Passenger control cost; RPC: Residual passenger penalty cost.

(2) Passenger arrival rate function (P): Linear passenger arrival rate function $\lambda_{i,s}(t)$ in Fig. 19(a) (L); Quadratic passenger arrival rate function $\lambda_{i,s}(t)$ in Fig. 19(b) (Q).