Bi-level model predictive control for metro networks
Integration of timetables, passenger flows, and train speed profiles
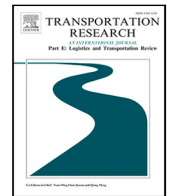
Liu, Xiaoyu; Dabiri, Azita; Xun, Jing; De Schutter, B.H.K.

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Bi-level model predictive control for metro networks: Integration of timetables, passenger flows, and train speed profiles

Xiaoyu Liu [a,*], Azita Dabiri [a], Jing Xun [b], Bart De Schutter [a]

[a] *Delft University of Technology, Delft Center for Systems and Control, Netherlands*
[b] *Beijing Jiaotong University, State Key Laboratory of Advanced Rail Autonomous Operation, China*

ARTICLE INFO

ABSTRACT

This paper deals with the train scheduling problem for metro networks taking into account time-dependent passenger origin–destination demands and train speed profiles. The aim is to adjust train schedules online according to time-dependent passenger demands so that passenger satisfaction and operational costs are jointly optimized. An extended passenger absorption model that explicitly includes time-dependent passenger origin–destination demands is developed, where the term "absorption" refers to passengers boarding trains. Then, the passenger absorption model is extended to a bi-level framework, where passenger demands and rolling stock availability are considered at the higher level, and detailed timetables and train speed profiles are included at the lower level. A bi-level model predictive control (MPC) approach is developed for the integrated problem. The optimization problems of both levels of the bi-level MPC approach can be converted into mixed-integer linear programming (MILP) problems, which enables us to solve them with existing MILP solvers. We then show that the recursive feasibility of both the higher-level and the lower-level optimization problems can be guaranteed. In this way, we can achieve real-time train scheduling for the metro system. Numerical experiments, based on real-life data from the Beijing metro network, illustrate the effectiveness of the extended passenger absorption model and the proposed bi-level MPC approach.

1. Introduction

As a safe, efficient, and eco-friendly transportation mode, the metro system plays a prominent role in public transportation. Real-time train scheduling is recognized as a valuable method for improving passenger satisfaction and energy efficiency under infrastructure limitations. As metro systems continue to expand to large-scale and networked systems, it becomes increasingly challenging to achieve real-time train scheduling while taking time-varying passenger flows and operational costs into account (Wang et al., 2015b; Hou et al., 2019).

Generally speaking, there are three key elements for train operation in metro networks, i.e., passenger flows, timetables, and train speeds. Some data-driven approaches can be applied to predict the near future passenger flow information in real time, which can be represented by time-dependent origin–destination (OD) matrices, thereby facilitating timetable scheduling (Noursalehi et al., 2022). An efficient passenger-oriented timetable should properly address time-dependent passenger OD demands (Wang et al., 2015b). Train speeds are closely related to operation time and energy consumption (Yin et al., 2017; Luan et al., 2018). As train speed control between two stations is usually conducted under the guidance of a recommended train speed profile, a well-designed speed

---

profile is crucial for effective train speed control (Hou et al., 2019). The integration of timetables, passenger flows, and train speed profiles is desired to generate efficient timetables that can jointly consider passenger satisfaction and operational costs in metro networks.

Real-time train scheduling considering passenger flows and train speed profiles is challenging due to its complexity and scale. Many studies include passenger flows in train scheduling problems while also considering stopping patterns of trains (Cacchiani et al., 2020), short-turning (Zhu and Goverde, 2019), and rolling stock circulation (Haahr et al., 2016; Zhao et al., 2023), but without time-dependent passenger origin–destination demands. Furthermore, train speed profiles are not included in these studies, and thus train speed-related objectives, e.g., the energy consumption of trains, cannot be directly included in the passenger-oriented train scheduling problem. Several papers consider the integration of timetables, passenger flows, and train speeds (Wang et al., 2015a,b; Mo et al., 2020; Yin et al., 2017). However, most existing studies that consider both passenger OD demands and train speed profiles, are limited to a single line because of the computational complexity issues arising from the integrated problem. This paper therefore focuses on the integration of timetables, passenger flows, and train speed problems for metro networks.

In order to reduce the computational burden of including many microscopic details of the network, some studies develop macroscopic models to handle passenger OD demands by optimizing departure frequencies (Higgins and Kozan, 1998; Canca et al., 2016; Li et al., 2018). The train departure frequency (i.e., the number of trains departing from a platform per time unit) is crucial for passenger satisfaction since it determines the maximum transport capacity of each line. The departure frequency should be adjusted properly to match time-varying passenger flows, e.g., compared with off-peak hours, higher departure frequencies are required during peak hours to address the large passenger demands. Furthermore, the departure frequency should be linked with specific departure and arrival times for a practically implementable timetable. Therefore, effective model formulations and control approaches are required to integrate train departure frequencies and train timetables in metro networks.

This paper contributes to the state of the art as follows.

1. An extended passenger absorption model (Liu et al., 2022) is developed, by including rolling stock circulation and the case that different lines share the same platform. The model allows determining train departure frequencies in metro networks considering time-dependent passenger OD demands.
2. A bi-level model predictive control (MPC) approach is proposed for real-time train scheduling considering passenger flows, rolling stock circulations, and train speed profiles. Passenger flows are included at the higher level based on the novel extended passenger absorption model, and detailed timetables and train speed profiles are incorporated at the lower level taking into account the detailed rolling stock circulation. The MPC optimization problems of both levels are exactly converted to mixed-integer linear programming problems, and we show that the recursive feasibility of both levels can be guaranteed.

The remaining part of the paper is arranged as follows: Section 2 reviews the related works. Section 4 introduces the developed passenger absorption model and the corresponding bi-level modeling framework. Section 5 introduces the developed bi-level MPC approach. Section 6 shows the effectiveness of the developed approach through numerical experiments, and conclusions are provided in Section 7.

2. Literature review

2.1. Passenger-oriented real-time timetable scheduling

There exists a considerable body of research on passenger-oriented timetable scheduling problems. Cury et al. (1980) presented an analytical model to describe the movement of trains and passengers; then, the optimal schedule is generated considering operational costs and the average delay of passengers. Wang et al. (2015a) developed an iterative algorithm to reduce the total passenger travel time on a metro line while considering the energy efficiency of trains, where train speeds in each segment were simplified via three stages, i.e., acceleration stage, cruising stage, and deceleration stage. Wang et al. (2018) realized real-time train scheduling for a metro line by integrating passenger demands and rolling stock circulation, and the aim is to ensure service quality while reducing operational costs. Hou et al. (2019) considered unexpected disturbances in a metro system and solved an MILP problem to reduce train delays, energy consumption, and the number of stranded passengers, where train speeds were also limited to a finite set of different speed levels. Considering train loading capacity constraints, Mo et al. (2020) formulated an MILP problem to maximize the utilization of regenerative energy, where rolling stock circulation was also incorporated into the resulting train scheduling problem. However, these studies do not include passenger origin–destination (OD) demands, indicating the possibility of further improving passenger satisfaction.

Real-time train scheduling with detailed passenger OD demands has received much attention in recent years. Niu et al. (2015) formulated a mixed-integer nonlinear programming (MINLP) problem for train scheduling in a rail corridor to reduce passenger waiting time taking into account time-dependent passenger demands. A space–time network was used by Yin et al. (2017) to describe the movement of trains on a metro line, where the train operation in a segment is considered for different speed levels; a Lagrangian relaxation-based method was then presented to optimize the total passenger waiting time and operational costs. Bešinović et al. (2022) integrated passenger flow control and train rescheduling under disruptions, and applied an iterative matheuristic approach to reduce the passenger waiting time and the time of recovering from disruptions. Nevertheless, these papers only include passenger OD demands on a single railway line, and further research is still required for the railway network.

Considering passenger OD demands in railway networks, Wang et al. (2015b) presented an event-based model that explicitly includes time-dependent passenger OD demands. Train arrival, train departure, and passenger arrival rate changes were formulated

as three different classes of events to describe the movement of passengers and trains. Yin et al. (2021) formulated a graph-based model to describe feasible passenger travel paths in a metro network; then, a decomposition-based adaptive large-neighborhood search approach is presented to minimize station crowdedness. Zhu and Goverde (2019) developed a timetable rescheduling approach for disruptions in a railway network, where passenger OD demands and passenger paths are included and used to determine weights of different objectives. Corman (2020) investigated the interactions between train schedules and passenger route choices, and presented a game theory-based approach to investigate the equilibrium point between them. Luan and Corman (2022) formulated the train schedules and passenger routing process in an integrated model, and the resulting MINLP formulation is reformulated as an MILP formulation to minimize passenger disutility (i.e., the number of stranded passengers, the passenger delays, and the passenger travel time) and the total train delay. However, these studies typically encounter computational issues because more details about passenger demands and railway networks should be included. Therefore, efficient model and solution approaches are required for passenger-oriented train scheduling.

2.2. Passenger-oriented train departure frequency optimization

The studies introduced in Section 2.1 aim to build elaborate models for detailed passenger dynamics and infrastructure information. These studies can generate directly implementable arrival and departure times of trains; however, the computational burden increases as many details related to passenger dynamics are included using such detailed microscopic models. In order to obtain a balanced trade-off between model accuracy and computational efficiency, another research direction develops macroscopic models to handle passenger OD demands by optimizing departure frequencies (Canca et al., 2016; Li et al., 2018; Liu et al., 2022), considering the periodic characteristic of train departures.

Optimizing the departure frequency determines the maximum transport capacity and is essential for handling passenger demands in urban public transport systems, e.g., city bus systems (Leurent et al., 2014) and metro systems (Higgins and Kozan, 1998). In general, higher departure frequencies typically result in higher operational costs while providing a better chance of boarding trains for passengers. The metro system, however, is quite different from other urban public transport systems, e.g., the braking distance of trains is relatively long, and the signaling system imposes an upper bound on the line frequency. Thus, effective departure frequency control approaches are required for metro networks to address time-dependent passenger OD demands considering operational costs and infrastructure constraints. Canca et al. (2016) solved an MINLP problem to optimize train capacities and line frequencies for each line of metro networks, where train capacities were considered as soft constraints. Li et al. (2018) developed a bi-level strategy to optimize the train departure frequencies at the upper level while a passenger assignment problem was considered at the lower level to balance operational cost and service quality. These studies aim to generate static and published train departure frequencies and schedules at the tactical planning stage based on periodic passenger flows, leaving an open gap in optimizing departure frequencies online based on real-time observed passenger demand.

Adjusting departure frequency online is also regarded as an effective way to accommodate time-dependent passenger demand Gkiotsalitis and Cats (2022). Pu and Zhan (2021) developed a two-stage method for railway line planning problems where the first stage generates a line plan with deterministic passenger demands and the second stage adjusts the line plan to accommodate real-life passenger demands. Liu et al. (2022) presented a passenger flow model to determine departure frequencies of metro systems in real time. However, that paper does not lead to a directly implementable timetable, i.e., specific arrival and departure times are not considered, and the case when different lines use the same physical track and/or physical platforms is also not involved. In summary, the above-mentioned studies only optimize the departure frequency of trains, which does not directly lead to practically executable timetables. Moreover, more detailed passenger flows, rolling stock circulation plans, and operational costs can be included to further improve operational performance.

2.3. MPC for real-time railway train scheduling

The studies introduced in Sections 2.1 and 2.2 are summarized in Table 1 based on the railway network details, passenger demands, and objectives. The train scheduling problem is a typical control problem with input and state constraints. From Sections 2.1 and 2.2, we can conclude that efficient modeling frameworks and control approaches for the integration of timetables, passenger flows, and train speeds in metro networks are urgently needed to achieve passenger-oriented train scheduling.

Model predictive control (MPC) is regarded as an efficient control methodology for real-time control of constrained systems (Mayne et al., 2000). MPC has also been implemented in real-time train scheduling problems. van den Boom and De Schutter (2006) applied MPC to minimize the delay of trains and the costs of changing train orders and braking connections based on a switching max-plus-linear model. Caimi et al. (2012) applied the MPC framework and proposed a scheduling assistant method for complex station areas considering infrastructure constraints and passenger satisfaction. Li et al. (2017) proposed a state space model to represent the dynamics of the train capacity and departure times on a metro line and an MPC approach was then developed to minimize the headway and timetable deviations by adjusting timetables and train capacity. Cavone et al. (2022) applied MPC to address disruptions and disturbances in railway networks, where an MILP problem is formulated under a bi-level structure using macroscopic and mesoscopic models. Wang et al. (2022) introduced a hierarchical MPC framework to integrate railway delay management and train control, which can realize real-time control and reduce delays effectively. Liu et al. (2023) applied MPC to passenger-oriented urban metro networks to adjust a given timetable according to real-time passenger demands. The successful applications of the aforementioned methods have motivated us to design an efficient MPC approach to realize real-time train scheduling.

Table 1
Summary of the relevant studies on passenger-oriented timetable scheduling.

Publications	Infrastructure	Passenger demands	Train capacity	Rolling stock circulation	Train order	Train speed	Objective (s)
Cury et al. (1980)	bi-directional line	OD-independent	no	no	no	no	minimize passenger delays and total number of trains
Niu et al. (2015)	uni-directional line	OD-dependent	hard constraint	no	no	no	minimize the total passenger waiting time at stations
Wang et al. (2015a)	uni-directional line	OD-independent	hard constraint	no	no	continuous speed	minimize train energy consumption and total passenger travel time
Wang et al. (2015b)	network	OD-dependent	hard constraint	no	no	continuous speed	minimize total passenger travel time and train energy consumption
Canca et al. (2016)	network	OD-dependent	soft constraint	yes	no	no	minimize total passenger travel time and operational costs
Yin et al. (2017)	bi-directional line	OD-dependent	hard constraint	no	no	speed levels	minimize total passenger waiting time and train energy consumption
Li et al. (2018)	uni-directional line	OD-dependent	hard constraint	no	no	no	optimizing departure frequency to balance operational cost and service quality
Wang et al. (2018)	bi-directional line	OD-independent	soft constraint	yes	no	no	minimize load factor variation, headway variation, and entering depot operations
Hou et al. (2019)	uni-directional line	OD-independent	hard constraint	no	no	speed levels	minimize train delays, energy consumption, and number of stranded passengers
Zhu and Goverde (2019)	network	OD-dependent	no	yes	yes	no	minimize passenger delays and impacts of cancelling trains and skipping stops
Mo et al. (2020)	bi-directional line	OD-independent	hard constraint	yes	no	no	maximize utilization of regenerative energy
Corman (2020)	network	OD-dependent	no	no	yes	no	analyze equilibrium point between train schedules and passenger route choices
Pu and Zhan (2021)	uni-directional line	OD-dependent	hard constraint	no	no	no	minimize operational costs and total passenger travel time
Yin et al. (2021)	network	OD-dependent	hard constraint	no	no	no	minimize station crowdedness
Bešinović et al. (2022)	bi-directional line	OD-dependent	hard constraint	yes	no	no	minimize passenger waiting time and deviation from original timetable
Luan and Corman (2022)	network	OD-dependent	hard constraint	no	yes	no	minimize passenger disutility and total train delay
Liu et al. (2022)	network	OD-dependent	hard constraint	no	no	no	minimize total passenger travel time
Current paper	network	OD-dependent	hard constraint	yes	yes	speed levels	minimize total passenger travel time and train energy consumption

We therefore develop a bi-level MPC approach for real-time train scheduling while considering time-dependent passenger OD demands and train speed profiles in metro networks. A bi-level model is developed to reduce the computational complexity of the integrated problem, and then the corresponding bi-level MPC approach is proposed. The higher-level controller is conducted with relatively slow dynamics to optimize departure frequencies (i.e., the number of trains departing from a platform per time unit), while the lower-level controller calculates detailed timetables with fast dynamics considering train scheduling constraints. The MPC optimization problems of both levels are transformed exactly into MILP problems, which enables us to solve them with existing MILP solvers.

3. Problem statement and assumptions

3.1. Problem statement

In metro systems, train schedules should be adjusted throughout the day to accommodate time-varying passenger flows while taking operational costs into account. A pre-determined timetable cannot include time-dependent passenger demands information and, in general, may be far from optimal. This paper focuses on adjusting train schedules online based on time-dependent passenger origin–destination demands while taking into account train capacity, rolling stock circulation, train speed profiles, and train orders. As discussed in Section 2, the time-dependent passenger-oriented train scheduling problem typically has computational issues. We therefore handle the problem in a bi-level framework to achieve a balanced trade-off between model accuracy and computational burden.

The general idea of the bi-level framework is illustrated in Fig. 1. The train departure frequency determines the upper bound of the transport capacity and is included at the higher level to address the time-dependent passenger OD demands based on the developed passenger absorption model. As the departure frequency is restricted by the availability of rolling stock, the rolling stock circulation is also considered at the higher level. The lower level focuses on generating a practically implementable timetable to fulfill the departure frequency while considering detailed rolling stock circulation, train speed profiles, and train orders.
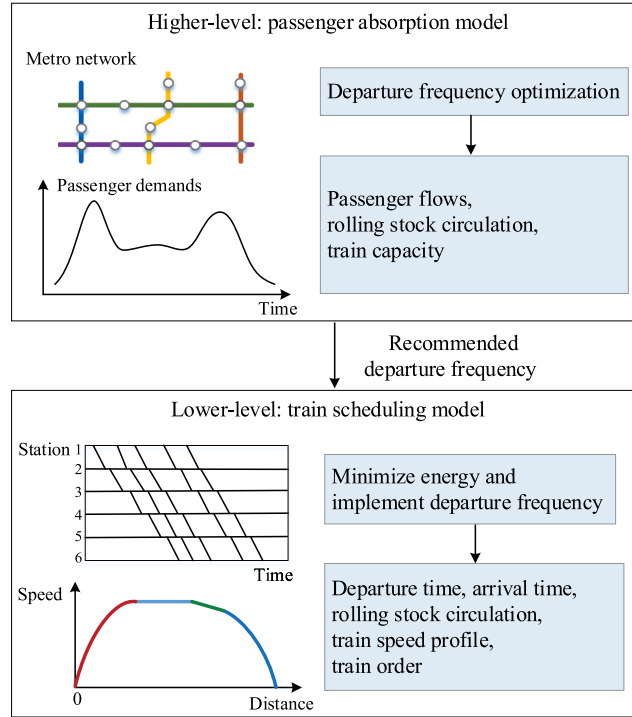
Fig. 1. Illustration of the bi-level framework.

3.2. Explanations and assumptions

Some general explanations and assumptions about the problem are listed as follows.

(1) A line in the metro network is typically defined as the route of one certain class of train services; these train services thus visit identical stations in each run. The assigned platforms for trains of each line are fixed.

(2) Passenger OD demands (i.e., the number of passengers choosing the metro for their travel, their origins, and their destinations) are not influenced by the departure frequencies. Time-dependent passenger OD demands are approximated as piece-wise constant functions.

(3) As passenger route choices observed from metro data collection systems typically exhibit consistent patterns (Noursalehi et al., 2022), we assume that the fractions of passengers choosing each route are given a priori, and that passengers do not change their route once they have entered the metro network.

(4) As we assume that passengers do not change their routes once they have entered the metro network, we define a lower bound for the departure frequency, so that the time interval between the departures of two consecutive trains is always shorter than a given threshold. In this way, the maximum waiting time for passengers should still be acceptable in case the departure frequency and/or departure times change with respect to the original timetable.

4. Mathematical model

Based on the bi-level framework, a bi-level model is presented for the passenger-oriented train scheduling problem, where (1) a macroscopic model, i.e., passenger absorption model, is included at the higher level considering time-independent passenger OD demands, rolling stock circulation, and train departure frequencies, and (2) a train scheduling model is included at the lower level considering the detailed timetable, detailed rolling stock circulation, train speed profiles, and train orders. In this section, we first provide the notations for the mathematical models. Then, the passenger absorption model and the train scheduling model are introduced respectively.

4.1. Notations

Tables 2, 3, and 4 respectively list the indices and input parameters, decision variables, and output variables for the model formulations. Noting that in Table 3 the decision variables for the higher level are the departure frequency $u_\ell(k)$ for all lines while the arrival time $a_{i,p}$, departure times $d_{i,p}$, $d_{i,\ell}^{\text{depot}}$, and speed profile option $x_{i,p,b}$ for all trains at all line platforms are the decision variables for the lower level.

Table 2
Indices and input parameters.

Notations	Definition
i, j	Index of trains
p, q	Index of line platforms, $p \in \mathscr{P}$, $q \in \mathscr{P}$, $\mathscr{P}$ is the set of line platforms
ℓ	Index of lines, $\ell \in \mathscr{L}$, $\mathscr{L}$ is the set of lines
s, e	Index of stations, $s, e \in \mathscr{S}$, $\mathscr{S}$ denotes the set of stations, s_p is the station corresponding to line platform p
z	Index of depots, $z \in \mathscr{Z}$, $\mathscr{Z}$ denotes the set of depots
k	Index of phases
T	Length of a phase
$\mathrm{p}_\ell^{\mathrm{tra}}(i)$	Preceding train of train i at line ℓ
$\mathrm{p}^{\mathrm{pla}}(p)$	Preceding line platform of line platform p
$\rho_{s,e}^{\mathrm{station}}(k)$	Passenger arrival rate at station s with destination e during phase k
$\rho_{p,e}(k)$	Passenger arrival rate at line platform p with destination e during phase k
$\lambda_{s,p,e}(k)$	Proportion of passengers at station s that are assigned to line platform p for their travel to destination e during phase k
$\alpha_{p,e}(k)$	Fraction of passengers absorbed by trains at line platform p with destination e during phase k
C_{train}	Maximum capacity of a train
$\chi_{p,q,e}$	Proportion of passengers transferring from line platform p to q with destination e
$\mathrm{cop}(p)$	The set of line platforms located at the identical station as line platform p
$\mathrm{in}(z)$	The set of platforms related to the entering link of depot z
$\mathrm{out}(z)$	The set of lines corresponding to the output link of depot z
N_z^{train}	The number of available trains at depot z
$t_{p,q}^{\mathrm{transfer}}$	Average time for passengers transferring from line platform p to line platform q
h_p^{min}	Minimum departure–arrival headway at line platform p
τ_p^{min}	Minimum dwell time of train at line platform p
τ_p^{max}	Maximum dwell time of train at line platform p
r_p^{min}	Minimum running time of train from line platform p to its succeeding line platform
r_p^{max}	Maximum running time of train from line platform p to its succeeding line platform
$\mathscr{B}_{i,p}$	Set of speed profile options for train i from line platform p to its succeeding line platform
$r_{i,p,b}$	Running time of train i from line platform p to its succeeding line platform with speed profile b, $b \in \mathscr{B}_{i,p}$
$\sigma_{p,p'}$	Binary parameter; if line platforms p and p' correspond to the same physical platform, $\sigma_{p,p'} = 1$; otherwise, $\sigma_{p,p'} = 0$

Table 3
Decision variables.

Notations	Definition
$u_\ell(k)$	The departure frequency from the depot corresponding to line ℓ during period k
$a_{i,p}$	Arrival time of train i at line platform p
$d_{i,p}$	Departure time of train i at line platform p
$d_{i,\ell}^{\mathrm{depot}}$	Departure time of train i from the depot corresponding to line ℓ
$x_{i,p,b}$	Binary variable indicating whether train i from line platform p selects speed profile b

4.2. Passenger absorption model

This section presents a macroscopic model to determine train departure frequencies based on the time-dependent passenger OD demands. In the passenger absorption model, the planning time window is divided into several phases, and in each phase, the time-dependent passenger demands at each platform are considered to be constant. The train departure frequency during each phase can be optimized while taking into account passenger OD demands. The variables and parameters related to the number of passengers for the passenger absorption model are listed in Table 4. To illustrate the above variables, a general overview of these variables is presented in Fig. 2, which features a station with two line platforms, i.e., line platform p and line platform q. More details about the variables are introduced below.

A matrix is typically used to describe time-dependent passenger OD demands. Each entry of the matrix is represented by $\rho_{s,e}^{\mathrm{station}}(t)$ where s and e are the origin and destination stations, respectively, and t represents time. Generally, $\rho_{s,e}^{\mathrm{station}}(t)$ is a nonlinear time-varying function, and it would significantly increase the computational complexity of including passenger flows in train scheduling problems. Considering the periodic characteristic of passenger flows in metro systems, the planning time window is divided into a sequence of phases with length T, and each phase has constant passenger demands. The illustration for approximating time-dependent passenger arrival rates in the passenger absorption model is given in Fig. 3.

In metro networks (especially in large cities, such as London, Barcelona), different lines may use the same physical track and/or the same physical platforms to maximize the utilization of the infrastructure. To distinguish platforms for different lines and different directions, we introduce the definition of "(virtual) line platform", where each line platform is exclusively linked with one direction of one line. For example, in Fig. 4, Line 1 and Line 2 share the same physical platform B, and we regard platform B as two different line platforms. The safe operation at the line platforms is ensured by constraints (17), (28)–(31) below.

The arrival rate $\rho_{p,e}(k)$ for passengers at line platform $p \in \mathscr{P}$ with destination station $e \in \mathscr{S}$ in phase k is computed by

$$\rho_{p,e}(k) = \lambda_{s_p,p,e}(k)\,\rho_{s_p,e}^{\mathrm{station}}(k), \tag{1}$$

Table 4
Output variables.

| Notations	Definition
$\tau_{i,p}$	Dwell time of train i at line platform p
$r_{i,p}$	Running time of train i from line platform p to its succeeding line platform
$\bar{r}_p$	Average running time of trains from line platform p to its succeeding line platform
$\gamma_p(k)$	Average time for a train from the first line platform to line platform p at phase k
$\beta_p(k)$	The largest integer less than or equal to $\frac{\gamma_p(k)}{T}$
$\phi_p(k)$	The remainder of $\frac{\gamma_p(k)}{T}$
$n_{p,e}(k)$	Number of passengers at line platform p with destination station e at the start of phase k
$n_{p,e}^{\text{absorb}}(k)$	Number of passengers absorbed by trains at line platform p with destination station e during phase k
$C_p(k)$	Total remaining capacity of trains visiting line platform p during phase k
$n_p^{\text{want}}(k)$	Total number of passengers who want to board trains at line platform p during phase k
$n_{p,e}^{\text{on-board}}(k)$	Number of passengers on board when trains arrive at line platform p with destination e during phase k
$n_{p,e}^{\text{alight}}(k)$	Number of passengers alighting from trains at line platform p with destination station e during phase k
$n_{p,q,e}^{\text{transfer}}(k)$	Number of passengers transferring from line platform p to line platform q with destination e during phase k
$n_{p,e}^{\text{trans,arrive}}(k)$	Number of transfer passengers arriving at line platform p with destination station e during phase k
$n_{p,e}^{\text{depart}}(k)$	Number of passengers departing from line platform p with destination station e during phase k
$f_p(k)$	Number of trains departing from line platform p during phase k
$\theta_z(k)$	The total number of trains available at depot z at the end of phase k
$y_{i,j,\ell,p}$	Binary variable; if train j departs from line platform p before train i departs from the depot related to line ℓ, $y_{i,j,\ell,p} = 1$; otherwise, $y_{i,j,\ell,p} = 0$
$\xi_{i,i',p,p'}$	Binary variable; if train i arrives at line platform p earlier than train i' at line platform p', $\xi_{i,i',p,p'} = 1$; otherwise, $\xi_{i,i',p,p'} = 0$
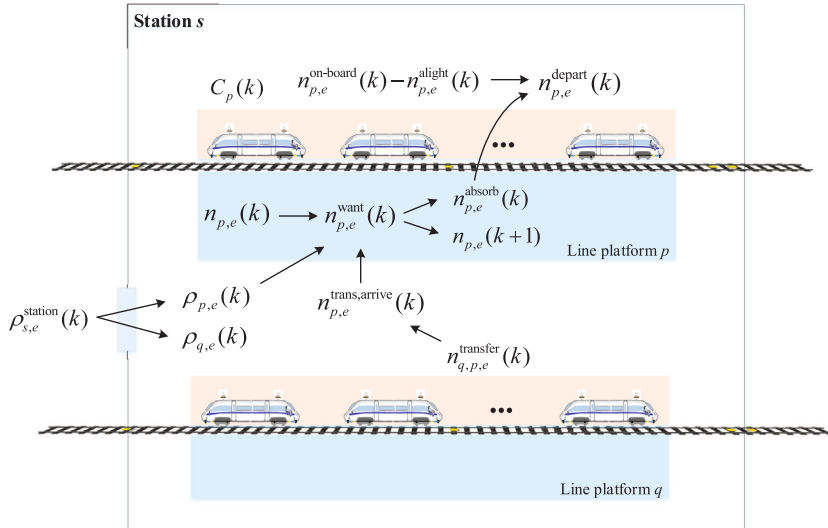
Fig. 2. Variables for the model during phase k.

where s_p represents the station corresponding to line platform p; note that each line platform p is corresponding to only one station s_p; $\lambda_{s_p,p,e}(k)$ denotes the splitting rate of passengers at station s_p who choose line platform p for their travel to destination e; $\rho_{s_p,e}^{\text{station}}(k)$ denotes passenger origin–destination demand at phase k with s_p and e as the origin and destination stations, respectively; $\mathcal{P}$ represents the set collecting all line platforms; $\mathcal{S}$ is the set collecting all stations in the network.

At each line platform, the number of passengers evolves as:

$$n_{p,e}(k+1) = n_{p,e}(k) + \rho_{p,e}(k)T + n_{p,e}^{\text{trans,arrive}}(k) - n_{p,e}^{\text{absorb}}(k), \tag{2}$$
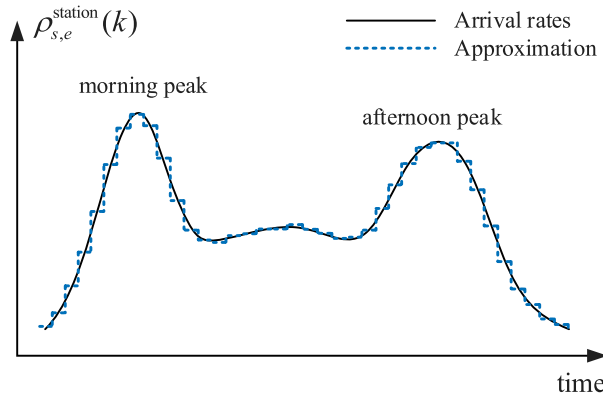
Fig. 3. Illustration for approximating time-dependent passenger arrival rates.
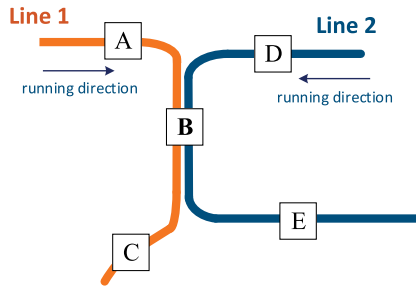

Fig. 4. Illustration for the line platform concept.

where $n_{p,e}(k)$ denotes the number of passengers stranded at line platform p with destination e at the start of phase k; $n_{p,e}^{\mathrm{trans,arrive}}(k)$ is the number of transfer passengers arriving at line platform p with destination e during phase k; $n_{p,e}^{\mathrm{absorb}}(k)$ denotes the number of passengers absorbed by trains at line platform p with destination e during phase k.

The variable $n_{p,e}^{\mathrm{absorb}}(k)$ is estimated by

$$n_{p,e}^{\mathrm{absorb}}(k) = \alpha_{p,e}(k) n_p^{\mathrm{absorb}}(k), \tag{3}$$

where $\alpha_{p,e}(k)$ is the relative fraction of passengers boarding trains at line platform p during phase k in order to reach their destination station e, and $\alpha_{p,e}(k)$ can be estimated through the historical data; $n_p^{\mathrm{absorb}}(k)$ denotes the total number of passengers boarding trains at line platform p during phase k, and we have

$$n_p^{\mathrm{absorb}}(k) = \min\left(C_p(k),\ n_p^{\mathrm{want}}(k)\right), \tag{4}$$

where $C_p(k)$ denotes the total remaining capacity provided by trains that visit line platform p during phase k, $n_p^{\mathrm{want}}(k)$ is the total number of passengers that want to board trains at line platform p during phase k. Thus, we have

$$n_p^{\mathrm{want}}(k) = n_p(k) + \rho_p(k)T + n_p^{\mathrm{trans,arrive}}(k), \tag{5}$$

with

$$n_p(k) = \sum_{e \in \mathscr{S}} n_{p,e}(k), \quad \rho_p(k) = \sum_{e \in \mathscr{S}} \rho_{p,e}(k), \quad n_p^{\mathrm{trans,arrive}}(k) = \sum_{e \in \mathscr{S}} n_{p,e}^{\mathrm{trans,arrive}}(k). \tag{6}$$

The total remaining capacity of trains $C_p(k)$ at line platform p during phase k is determined by the maximum capacity of the trains, the number of passengers already on board the train, and the number of passengers alighting from the trains:

$$C_p(k) = f_p(k) \cdot C_{\mathrm{train}} - \sum_{e \in \mathscr{S}} n_{p,e}^{\mathrm{on-board}}(k) + \sum_{e \in \mathscr{S}} n_{p,e}^{\mathrm{alight}}(k), \tag{7}$$

where $f_p(k)$ denotes the number of trains departing from line platform p during phase k, and $f_p(k)$ is the decision variable of the absorption model; C_{train} represents the maximum capacity of a train; $n_{p,e}^{\mathrm{on-board}}(k)$ denotes the number of passengers with destination station e already on board the train when trains arrive at line platform p during phase k; $n_{p,e}^{\mathrm{alight}}(k)$ represents the number of passengers with destination station e alighting from trains at line platform p during phase k.

We define $\mathrm{p^{pla}}(p)$ as the preceding line platform of line platform p, and $\bar{r}_{\mathrm{p^{pla}}(p)}$ as the mean running time for trains from line platform $\mathrm{p^{pla}}(p)$ to p. Then, the variable $n_{p,e}^{\mathrm{on-board}}(k)$ in (7) is the number of passengers transported by trains from line platform

$p^{\mathrm{pla}}(p)$ to p during phase k with destination station e. As the length of the time step for the absorption model is T, and passengers departing from line platform $p^{\mathrm{pla}}(p)$ require time $\bar{r}_{p^{\mathrm{pla}}(p)}$ to reach line platform p, we have

$$n_{p,e}^{\mathrm{on-board}}(k) = \frac{T - \bar{r}_{p^{\mathrm{pla}}(p)}}{T} n_{p^{\mathrm{pla}}(p),e}^{\mathrm{depart}}(k) + \frac{\bar{r}_{p^{\mathrm{pla}}(p)}}{T} n_{p^{\mathrm{pla}}(p),e}^{\mathrm{depart}}(k-1), \tag{8}$$

where $n_{p^{\mathrm{pla}}(p),e}^{\mathrm{depart}}(k)$ represents the number of passengers departing from line platform $p^{\mathrm{pla}}(p)$ with destination e during phase k, and T and $\bar{r}_{p^{\mathrm{pla}}(p)}$ are parameters of the model. As the developed model aims to address passenger demands within a relatively long time, we typically set $T \gg \bar{r}_{p^{\mathrm{pla}}(p)}$. Note that if p is the first line platform of the line, we set $n_{p,e}^{\mathrm{on-board}}(k) = 0$, which means the train is empty when arriving at the first line platform of a line.

The number of passengers $n_{p,q,e}^{\mathrm{transfer}}(k)$ transferring from line platform p to line platform q with destination e during phase k, is calculated by

$$n_{p,q,e}^{\mathrm{transfer}}(k) = \chi_{p,q,e}\, n_{p,e}^{\mathrm{on-board}}(k), \forall q \in \mathrm{cop}(p) \setminus \{p\}, \tag{9}$$

where $\mathrm{cop}(p)$ defines a set collecting all line platforms located at the identical station as line platform p, $\chi_{p,q,e}$ refers to the proportion of passengers[1] transferring from line platform p to line platform q with destination e, which can be estimated according to the historical data, and $\sum_{q \in \mathrm{cop}(p)} \chi_{p,q,e} = 1$.

At each line platform, passengers that either have transfer connections or have reached their destinations will alight from trains. Thus, the number of alighting passengers $n_{p,e}^{\mathrm{alight}}(k)$ is computed by

$$n_{p,e}^{\mathrm{alight}}(k) = \begin{cases} \displaystyle\sum_{q \in \mathrm{cop}(p)} n_{p,q,e}^{\mathrm{transfer}}(k), & \text{if } e \in \mathscr{S} \setminus \{\mathrm{sta}(p)\}, \\ n_{p,e}^{\mathrm{on-board}}(k), & \text{if } e = \mathrm{sta}(p), \end{cases} \tag{10}$$

where $\mathrm{sta}(p)$ refers to the station corresponding to line platform p.

The number of departing passengers $n_{p,e}^{\mathrm{depart}}(k)$ is computed by

$$n_{p,e}^{\mathrm{depart}}(k) = n_{p,e}^{\mathrm{on-board}}(k) - n_{p,e}^{\mathrm{alight}}(k) + n_{p,e}^{\mathrm{absorb}}(k), \tag{11}$$

which means that, at each line platform, some passengers will alight from trains while passengers waiting at the platform will board the trains before the trains depart from the platform.

As the basic time unit of the model is T, and the transfer passengers require time $t_{q,p}^{\mathrm{transfer}}$ to reach line platform p, the number of transfer passengers arriving at line platform p. Then, $n_{p,e}^{\mathrm{trans,arrive}}(k)$ can be computed by

$$n_{p,e}^{\mathrm{trans,arrive}}(k) = \sum_{q \in \mathrm{cop}(p) \setminus \{p\}} \left(\frac{T - t_{q,p}^{\mathrm{transfer}}}{T} n_{q,p,e}^{\mathrm{transfer}}(k) + \frac{t_{q,p}^{\mathrm{transfer}}}{T} n_{q,p,e}^{\mathrm{transfer}}(k-1) \right), \tag{12}$$

where $t_{q,p}^{\mathrm{transfer}}$ denotes the mean time of transferring from line platform q to line platform p.

In this paper, we address the train scheduling problem without disruptions. Thus, for each line, all trains will visit every predetermined station along the line with the same stopping pattern. Let us define $\gamma_p(k)$ as the mean time of trains from a depot to line platform p. Define $\lfloor x \rfloor$ as the greatest integer less than or equal to x; then, we define

$$\beta_p(k) = \left\lfloor \frac{\gamma_p(k)}{T} \right\rfloor, \tag{13}$$

$$\phi_p(k) = \gamma_p(k) - \beta_p(k)T, \tag{14}$$

where $\phi_p(k)$ denotes the remainder of $\frac{\gamma_p(k)}{T}$ with $0 \le \phi_p(k) < T$. In this context, $\beta_p(k) \ge 0$ determines the number of phases required for trains from the depot to line platform p.

The departure frequency $f_p(k)$ of line platform p is determined by the departure frequency from the output link of the depot. As trains typically depart from depot and require $\gamma_p(k)$ to reach line platform p, $f_p(k)$ is determined by

$$f_p(k) = \frac{T - \phi_p(k)}{T} u_\ell\big(k - \beta_p(k)\big) + \frac{\phi_p(k)}{T} u_\ell\big(k - \beta_p(k) - 1\big), p \in \mathscr{P}_\ell, \tag{15}$$

where $u_\ell(k)$ defines the departure frequency from the depot corresponding to line ℓ during period k; $\mathscr{P}_\ell$ denotes set of line platforms of line ℓ.

The departure frequency determines the time interval between the departure times of two consecutive trains, thereby influencing the maximum waiting time of passengers. We define a lower bound for the departure frequency:

$$f_p(k) \ge f_{\min}, \tag{16}$$

where $f_{\min}$ represents the minimum departure frequency. In this way, the time interval between the departures of two consecutive trains is always shorter than a given threshold. Thus, the maximum waiting time for passengers should still be acceptable in case the departure frequency and/or departure time change.

---

[1] $\chi_{p,p,e}$ represents the proportion of passengers remaining on trains at platform p.

Remark 4.1. We assume that rolling stock resource is such that the minimum departure frequency constraint can always be satisfied. However, in case this assumption is dropped and the rolling stock resource is so limited that the minimum departure frequency constraint can be violated, then we can turn the minimum departure frequency constraint into a soft constraint.

To ensure safe operation, the number of trains departing from line platform p during phase k is constrained by

$$\sum_{p' \in \text{phy}(p)} f_{p'}(k)\left(h_p^{\min} + \tau_p^{\min}\right) \leq T, \tag{17}$$

where phy(p) represents the set of line platforms using the same physical platform as line platform p; $h_p^{\min}$ and $\tau_p^{\min}$ are the minimum departure–arrival headway and the minimum dwell time at line platform p, respectively.

The rolling stock circulation determines the availability of trains for each line, which should be included in the optimization of train departure frequencies. In this paper, we only consider the case that the depot is located at the end of each line, and the constraints for rolling stock circulation are

$$\theta_z(k) = \theta_z(k-1) + \sum_{p \in \text{in}(z)} f_p(k) - \sum_{\ell \in \text{out}(z)} u_\ell(k), \forall z \in \mathscr{Z} \tag{18}$$

$$\theta_z(k) \geq 0, \forall z \in \mathscr{Z}, \tag{19}$$

where z is the depot index, $\mathscr{Z}$ is the set of depots, $\theta_z(k)$ represents the total number of trains available at depot z at the end of phase k, $\sum_{p \in \text{in}(z)} f_p(k)$ calculates the total number of trains entering depot z during phase k, in(z) defines the set of line platforms corresponding to the entering link of depot z, $\sum_{\ell \in \text{out}(z)} u_\ell(k)$ calculates the total number of trains leaving depot k during phase k, and out(z) defines the set of lines corresponding to the output link of depot z, $\theta_z(0) = N_z^{\text{train}}$ is a parameter representing the number of trains available at depot z.

Remark 4.2. Note that if $\theta_z(k) = 0$, depot z may need to wait for new arrivals. This effect is not included in the higher-level problem and may thus result in suboptimality for the final solution produced by the lower-level optimization problem.

4.3. Train scheduling model

As indicated before, the upper level of the proposed bi-level framework determines the number of trains departing from the lines in the metro network. However, the exact departure and arrival times should be determined to obtain a practically implementable timetable. Therefore, a train scheduling model is introduced for the detailed timetable (including departure/arrival time and train orders), detailed rolling stock circulation, and train speed profiles.

There are typically three groups of constraints corresponding to the train operation, i.e., departure/arrival constraints, rolling stock circulation constraints, running time constraints, and headway constraints.

4.3.1. Departure/arrival constraints
The departure time $d_{i,p}$ of train i at line platform p is determined by:

$$d_{i,p} = a_{i,p} + \tau_{i,p}, \tag{20}$$

where $a_{i,p}$ and $\tau_{i,p}$ respectively denote arrival time and dwell time of train i at line platform p, and $\tau_{i,p}$ should satisfy

$$\tau_p^{\min} \leq \tau_{i,p} \leq \tau_p^{\max}, \tag{21}$$

where $\tau_p^{\min}$ and $\tau_p^{\max}$ denote the minimum and the maximum dwell times for trains at line platform p, respectively.

Define $\text{p}^{\text{pla}}(p)$ as the preceding line platform of line platform p, the arrival time $a_{i,p}$ of train i at line platform p is determined by:

$$a_{i,p} = d_{i,\text{p}^{\text{pla}}(p)} + r_{i,\text{p}^{\text{pla}}(p)}, \tag{22}$$

where $d_{i,\text{p}^{\text{pla}}(p)}$ denotes the departure time of train i at line platform $\text{p}^{\text{pla}}(p)$, $r_{i,\text{p}^{\text{pla}}(p)}$ is the running time of train i from line platform $\text{p}^{\text{pla}}(p)$ to line platform p.

Remark 4.3. If p is the first line platform of the line, for completeness, we set $d_{i,\text{p}^{\text{pla}}(p)} = d_{i,\ell}^{\text{depot}}$, where $d_{i,\ell}^{\text{depot}}$ represents the departure time of train i from the depot corresponding to line ℓ, and $r_{i,\ell}^{\text{depot}}$ is the running time of train i from the depot to the first line platform of the line, $p \in \mathscr{P}_\ell$.

4.3.2. Rolling stock circulation constraints
Before sending a train from a depot, the availability of trains in the depot should be taken into account. Let us define a binary variable $y_{i,j,\ell,p}$ based on the departure time $d_{i,\ell}^{\text{depot}}$ of train i from the depot corresponding to line ℓ:

$$y_{i,j,\ell,p} = \begin{cases} 1, & \text{if} \quad d_{j,p} \leq d_{i,\ell}^{\text{depot}}; \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$
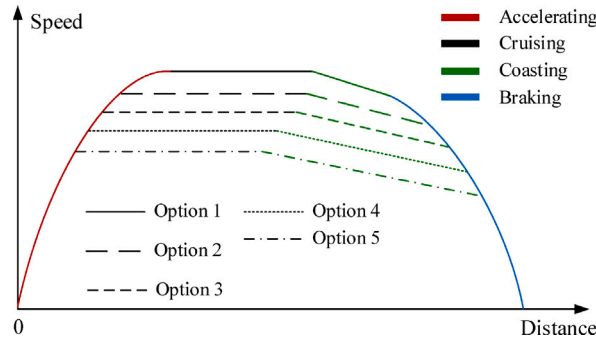
Fig. 5. Illustration of different train speed profile options in a segment.

Then, the rolling stock circulation constraint at the lower level is

$$\sum_{\ell \in \text{out}(z)} \sum_{j \in \mathscr{J}_\ell} y_{i,j,\ell,p} - \sum_{p \in \text{in}(z)} \sum_{j \in \mathscr{I}_p} y_{i,j,\ell,p} \leq N_z^{\text{train}}, \tag{24}$$

where $\mathscr{J}_\ell$ defines the set of trains departing from the output link of the depot corresponding to line ℓ, and $\mathscr{I}_p$ defines the set of trains that depart from line platform p. In (24), the first term represents the total number of trains that have left depot z before train i departs, while the second term accounts for the total number of trains that have entered depot z prior to the departure of train i from the same depot.

4.3.3. Running time constraints

Considering the operational requirement and speed limits, the running time constraint is

$$r_p^{\min} \leq r_{i,p} \leq r_p^{\max}, \tag{25}$$

where $r_p^{\min}$ and $r_p^{\max}$ are the minimum and maximum running times from line platform p to its succeeding line platform, respectively.

In general, $r_{i,p}$ is determined by train running speeds. In real life, train speeds and train running time between two stations are usually adjusted through an on-board train operation system, where different operation levels are defined, and each level corresponds to one speed profile option (Yin et al., 2017). Therefore, we consider different train speed profile options for trains between two stations, and each option is related to a specific running time and a value of energy cost. In this context, the running time $r_{i,p}$ for train i is determined by

$$r_{i,p} = \sum_{b \in \mathscr{B}_{i,p}} x_{i,p,b} \, r_{i,p,b}, \tag{26}$$

where b denotes the train speed profile option index, $\mathscr{B}_{i,p}$ represents the set of speed profile options for train i at line platform p (for example, speed profile options in Fig. 5); $r_{i,p,b}$ denotes the running time corresponding to speed profile option b; $x_{i,p,b}$ represents a binary variable indicating whether a speed profile is selected, i.e., $x_{i,p,b} = 1$ if speed profile option b is selected for train i at line platform p, otherwise, $x_{i,p,b} = 0$.

In order to ensure only one option can be selected, $x_{i,p,b}$ should satisfy

$$\sum_{b \in \mathscr{B}_{i,p}} x_{i,p,b} = 1. \tag{27}$$

In this paper, different speed profiles can be calculated offline, and we only need to select one speed profile among different speed profile options in real time.

4.3.4. Headway constraints

Headway is crucial for the safety of two consecutive trains, and for trains in the same line (see Fig. 6(a)) we have:

$$a_{i,p} \geq d_{p_\ell^{\text{tra}}(i),p} + h_p^{\min}, \tag{28}$$

where $p_\ell^{\text{tra}}(i)$ represents the preceding train of train i at line ℓ, and $h_p^{\min}$ represents the minimum departure–arrival headway at line platform p.

In metro networks (especially in large cities, such as London, Barcelona), different lines may use the same physical track and/or physical platforms to maximize the utilization of infrastructure (see Fig. 6(b)). In this context, headway constraints for trains on different lines are required. We use a binary variable $\xi_{i,i',p,p'}$ to represent the order of trains from different lines:

$$\xi_{i,i',p,p'} = \begin{cases} 1, & \text{if } a_{i,p} \leq a_{i',p'}; \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$
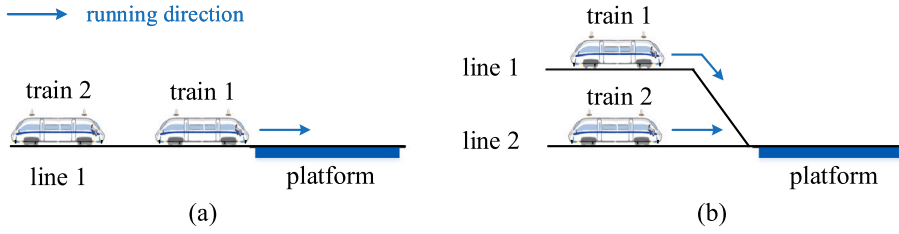
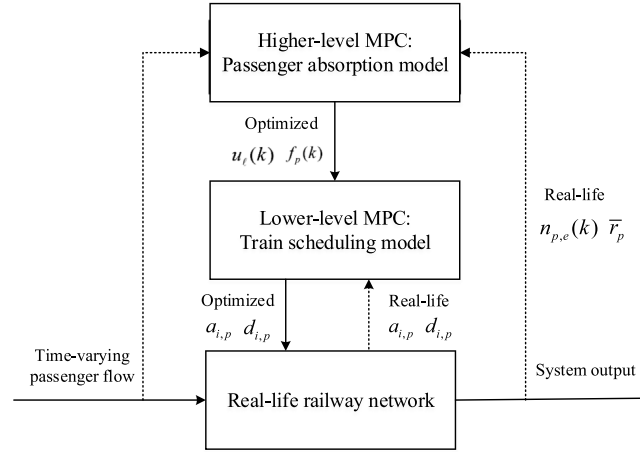**Fig. 6.** Different lines may use the same physical platform.

Fig. 7. Bi-level control structure for the integrated problem.

Then, the headway constraint for train i and train i' can be represented as

$$a_{i,p} - d_{i',p'} \geq h_p^{\min} - M_a(1 - \sigma_{p,p'} + \xi_{i,i',p,p'}), \tag{30}$$

where M_a represents a sufficiently large positive value. Eq. (30) represents the headway constraint of trains i and i' when line platforms p and p' are associated with the same physical platform, i.e., $\sigma_{p,p'} = 1$; otherwise, $\sigma_{p,p'} = 0$, then (30) holds automatically.

Furthermore, the order of trains should also satisfy

$$\xi_{i,i',p,p'} + \xi_{i',i,p',p} = 1, \tag{31}$$

which is employed to keep train order variables consistent, i.e., either $\xi_{i,i',p,p'} = 1$ or $\xi_{i',i,p',p} = 1$.

5. Bi-level MPC for train scheduling

MPC is an efficient real-time model-based control approach where finite-horizon optimization procedures are conducted repeatedly in a receding horizon scheme (Mayne, 2014). By dividing the long planning time window into several short time windows, MPC solves the problem with a short time window in a receding horizon manner to reduce the computational burden, while taking into account the real-time information of the metro network. A bi-level MPC approach is proposed to achieve real-time timetable scheduling in this section. The general introduction and the bi-level structure are introduced in Section 5.1. Then, the MPC approaches for both levels are presented in Section 5.2 and Section 5.3, respectively.

5.1. Bi-level MPC for the integrated problem

The bi-level control scheme is illustrated in Fig. 7 where passenger flow control and train scheduling are addressed at two different levels.

As shown in Fig. 7, the higher level aims to address time-dependent passenger origin–destination (OD) demands by determining the number of trains departing from each line platform during each phase. The higher-level controller uses the passenger absorption model of Section 4.2. As we approximate time-dependent passenger OD demands as piecewise constants, the higher-level controller can be handled at every phase. Therefore, the higher-level controller can be conducted in relatively slow dynamics. Once the higher-level MPC optimization problem is solved, the optimized decision variables $f_p^*(k)$ are sent to the lower level. At the lower level, the train scheduling problem is solved to obtain the optimized arrival time $a_{i,p}^*$ and departure time $d_{i,p}^*$ for each train taking train speed

profiles into account. The lower-level controller should be addressed with fast dynamics for real-time train scheduling so that the obtained arrival times, departure times, and train speed profiles can be implemented into the practical metro network.

In the bi-level MPC scheme, at the end of the control interval of the lower-level controller, the planning time window at the lower level will be moved for one step, and the train scheduling problem is resolved for the next step according to the collected real-life arrival and departure times ($a_{i,p}$ and $d_{i,p}$). At the end of the control interval of the higher-level controller (i.e., one phase), the planning time span for the higher level will be shifted for one phase, and the control problem will be solved again for the next phase based on the realized $\bar{r}_p$ and $n_{p,e}(k)$.

5.2. Higher-level MPC: departure frequency optimization

The time-dependent passenger OD demands can be addressed by a centralized MPC approach based on the model presented in Section 4.2. As passenger flows usually change periodically, the control time interval of the higher-level controller is equal to the length of a phase. The decision variable at the higher level will be the number of trains departing from each line platform during each phase.

The total travel time for passengers during phase k is represented by

$$J^{\text{pass}}(k) = \sum_{p \in \mathcal{P}} \left(n_p(k)T + n_p^{\text{depart}}(k)\bar{r}_p + n_p^{\text{trans,arrive}}(k)t_p^{\text{transfer}} \right), \tag{32}$$

where p defines a set collecting all line platforms in the network; $n_p(k)T$ represents the passenger waiting time at line platform p during phase k; $n_p^{\text{depart}}(k)\bar{r}_p$ denotes the total running time for passengers departing from line platform p during phase k; $n_p^{\text{trans,arrive}}(k)t_p^{\text{transfer}}$ is the total transfer time for passengers at line platform p during phase k, and t_p^{transfer} denotes the average time for passengers transferring to line platform p.

Although scheduling more trains, running with the minimum headway, can help to minimize $J^{\text{pass}}(k)$, it is typically not acceptable to use too many trains in real life, as it would significantly increase the total energy consumption. Thus, a penalty term corresponding to train energy consumption is included in the cost function. Then, the MPC optimization problem for passenger flow control at phase k_0 can be represented by

$$\min_{\boldsymbol{u}(k)} J^{\text{high}} = \sum_{k=k_0}^{k_0+N-1} \left(J^{\text{pass}}(k) + \eta \sum_{p \in \mathcal{P}} f_p(k)\bar{E}_p \right) + L_N(k_0) \tag{33}$$

$$\text{s.t.} \quad (1) - (12), (15) - (19),$$

where N denotes the number of phases in the prediction time span; η represents a weight; $\bar{E}_p$ denotes the average energy consumption for a train running from the line platform p to its succeeding line platform, since the higher level does not know which speed profile will be selected at the lower level when solving the high-level optimization problem, we use the average value among all speed profile options in the high-level optimization problem; and $\boldsymbol{u}(k)$ collects the independent decision variables, i.e., the departure frequency at the depot corresponding to each line $u_\ell(k)$; $L_N(k_0)$ is a penalty term for the passengers that can not board trains at the end of the prediction window, and in this paper we set $L_N(k_0) = \sum_{p \in P} n_p(k_0 + N)T$. As stated in (15), the departure frequencies of other line platforms are determined by $u_\ell(k)$.

In each MPC step, problem (33) is a nonlinear nonconvex optimization problem. By using the following properties in Williams (2013), we can convert the nonconvex term (4) into linear constraints.

Transformation property 5.1: If we introduce a binary variable $\delta_{k,p}^{\text{absorb}}$ and an auxiliary real variable $f_{k,p}^{\text{absorb}}$ with $f_{k,p}^{\text{absorb}} = n_p^{\text{want}}(k) - C_p(k)$. Then, if we define m_p and M_p as the minimum and the maximum values of $f_{k,p}^{\text{absorb}}$, respectively, the expression $\delta_{k,p}^{\text{absorb}} = 1 \Leftrightarrow f_{k,p}^{\text{absorb}} \leq 0$ is equivalent to

$$\begin{cases} f_{k,p}^{\text{absorb}} \leq M_p \left(1 - \delta_{k,p}^{\text{absorb}} \right), \\ f_{k,p}^{\text{absorb}} \geq \varepsilon + \left(m_p - \varepsilon \right) \delta_{k,p}^{\text{absorb}}, \end{cases} \tag{34}$$

where ε represents a sufficiently small number. Then, (4) can be replaced by

$$n_p^{\text{absorb}}(k) = \delta_{k,p}^{\text{absorb}} n_p^{\text{want}}(k) + \left(1 - \delta_{k,p}^{\text{absorb}} \right) C_p(k). \tag{35}$$

Transformation property 5.2: The multiplication of real variable $\tilde{y}$ and logical variable $\tilde{\delta}$ can be replaced by an auxiliary real variable $\tilde{g} = \tilde{y} \cdot \tilde{\delta}$. Then, $\tilde{g} = \tilde{y} \cdot \tilde{\delta}$ can be exactly transformed into

$$\begin{cases} \tilde{g} \leq M_{\tilde{y}} \tilde{\delta}, \\ \tilde{g} \geq m_{\tilde{y}} \tilde{\delta}, \\ \tilde{g} \leq \tilde{y} - m_{\tilde{y}}(1 - \tilde{\delta}), \\ \tilde{g} \geq \tilde{y} - M_{\tilde{y}}(1 - \tilde{\delta}), \end{cases} \tag{36}$$

where $M_{\tilde{y}}$ and $m_{\tilde{y}}$ respectively represent the maximum and minimum values of $\tilde{y}$.

Table 5
Numbers of variables and constraints in problem (37).

Variables or constraints	Maximal possible total number
Continuous variables	$(7 \cdot |\mathscr{S}| + 6) \cdot N \cdot |\mathscr{P}|$
Binary variables	$N \cdot |\mathscr{P}|$
Auxiliary continuous variables	$3 \cdot N \cdot |\mathscr{P}|$
Constraints	$(8 \cdot |\mathscr{S}| + 16) \cdot N \cdot |\mathscr{P}|$

By using the above transformations, problem (33) can be exactly converted to an MILP problem with the following form:

$$
\begin{aligned}
\min_{\substack{\boldsymbol{x}(k),\boldsymbol{u}(k) \\ \boldsymbol{\delta}(k),\boldsymbol{z}(k)}} J^{\mathrm{high}} = \sum_{k=k_0}^{k_0+N-1} & \left(J^{\mathrm{pass}}(k) + \eta \sum_{p\in\mathscr{P}} f_p(k)\bar{E}_p \right) + L_N(k_0) \\
\text{s.t.} \quad & \boldsymbol{x}(k+1) = A_k\boldsymbol{x}(k) + B_{1,k}\boldsymbol{u}(k) + B_{2,k}\boldsymbol{\delta}(k) + B_{3,k}\boldsymbol{z}(k), \\
& D_{2,k}\boldsymbol{\delta}(k) + D_{3,k}\boldsymbol{z}(k) \leq D_{1,k}\boldsymbol{u}(k) + D_{4,k}\boldsymbol{x}(k) + D_{5,k}, \\
& k = k_0, \ldots, k_0 + N - 1,
\end{aligned}
\tag{37}
$$

where $\boldsymbol{x}(k)$ collects the output variables in phase k; $\boldsymbol{\delta}(k)$ and $\boldsymbol{z}(k)$ collect the auxiliary binary and auxiliary continuous variables in phase k, respectively; $\boldsymbol{x}(k+1) = A_k\boldsymbol{x}(k) + B_{1,k}\boldsymbol{u}(k) + B_{2,k}\boldsymbol{\delta}(k) + B_{3,k}\boldsymbol{z}(k)$ includes all equality constraints in (1)–(12), (15), and (18); $D_{2,k}\boldsymbol{\delta}(k) + D_{3,k}\boldsymbol{z}(k) \leq D_{1,k}\boldsymbol{u}(k) + D_{4,k}\boldsymbol{x}(k) + D_{5,k}$ includes all inequality constraints.

Remark 5.1. *(Complexity Analysis)* There are three categories of variables in (37), i.e., continuous variables, binary variables, and auxiliary continuous variables. The constraints include linear and nonlinear constraints. The total numbers of variables and constraints are listed in Table 5, where $\mathscr{S}$, $\mathscr{P}$, and $\mathscr{L}$ are the set of stations, line platforms, and lines, respectively, and $|\cdot|$ denotes the cardinality of a set.

It can be observed from Table 5 that the number of variables depends on the scale of the considered metro network and the prediction horizon N. The MILP problem is an NP-hard problem, and the computation time for solving the problem typically increases rapidly when the number of integer variables increases (Garey and Johnson, 1979). In this problem, the number of binary variables is determined by the number of lines $|\mathscr{L}|$, the number of line platforms $|\mathscr{P}|$, and the prediction horizon N. A large prediction horizon N can include more information in the train departure frequency optimization, while the computational burden increases. Therefore, for a given metro network, choosing a proper prediction horizon is important to balance the computation time versus the performance.

Solving problem (37) results in a series of decision variables from phase k_0 to k_0+N-1, and according to the MPC paradigm, only the variables for phase k_0 are applied. In the next phase, the prediction time span is shifted for one phase, and a new optimization problem can be obtained.

Lemma 5.1. *(Recursive Feasibility)* If problem (37) is feasible at phase k_0 with initial state $\boldsymbol{x}(k_0)$, then the feasibility of problem (37) at phase $k_0 + 1$ can also be ensured.

Proof. The proof is based on finding a feasible solution for phase $k_0 + 1$. At phase k_0 with initial state $\boldsymbol{x}(k_0)$, problem (37) can be solved and the optimized decision variables are collected in $\boldsymbol{U}(k_0)$ with

$$
\boldsymbol{U}(k_0) = [(\boldsymbol{u}^*(k_0))^{\mathrm{T}}, (\boldsymbol{u}^*(k_0+1))^{\mathrm{T}}, \ldots, (\boldsymbol{u}^*(k_0+N-1))^{\mathrm{T}}]^{\mathrm{T}},
\tag{38}
$$

where $\boldsymbol{u}^*(k_0)$ is the optimized value of $\boldsymbol{u}(k_0)$ for solving problem (37). By implementing the first decision variable $\boldsymbol{u}^*(k_0)$, we get

$$
\boldsymbol{x}^*(k_0+1) = A_{k_0}\boldsymbol{x}(k_0) + B_{1,k_0}\boldsymbol{u}^*(k_0) + B_{2,k_0}\boldsymbol{\delta}^*(k_0) + B_{3,k_0}\boldsymbol{z}^*(k_0).
\tag{39}
$$

As we only have input constraint (17) at the higher level, and the inequalities constraints introduced in Transformation property 5.1 and Transformation property 5.2 are equivalent transformations for the mixed logical dynamical (MLD) model, a feasible solution for phase $k_0 + 1$ can always be found as

$$
\boldsymbol{U}(k_0+1) = [(\boldsymbol{u}^*(k_0+1))^{\mathrm{T}}, \ldots, (\boldsymbol{u}^*(k_0+N-1))^{\mathrm{T}}, (\boldsymbol{u}(k_0+N))^{\mathrm{T}}]^{\mathrm{T}},
\tag{40}
$$

where $\boldsymbol{u}^*(k_0+1), \ldots, \boldsymbol{u}^*(k_0+N-1)$ are from solution $\boldsymbol{U}(k_0)$ at phase k_0, and $\boldsymbol{u}(k_0+N)$ can be any solution that satisfies (17), e.g., the corresponding value of the regular timetable. Hence, the recursive feasibility of the higher-level MPC problem is guaranteed. □

5.3. Lower-level MPC: train scheduling

Based on the number of trains departing from each line platform obtained from the higher-level controller, the detailed timetable considering the energy consumption can be generated at the lower level. The lower level uses the train scheduling model introduced in Section 4.3, and the decision variables are departure/arrival times and train speed profile options of trains. As the lower-level controller aims to generate a practically implementable timetable considering real-time information of the network, the lower-level controller should be addressed with relatively fast dynamics.

According to Section 4.3, the energy consumption $E_i(p)$ for train i from line platform p to its succeeding line platform is determined by

$$E_i(p) = \sum_{b \in \mathscr{B}_{i,p}} x_{i,p,b} E_{i,b}(p), \tag{41}$$

where $E_{i,b}(p)$ denotes the energy consumption of speed profile option b for train i from line platform p to its succeeding line platform.

Generally, the energy consumption of a train in a segment is highly related to the running time, i.e., a longer running time (and thus a lower speed) typically leads to lower energy consumption. Furthermore, a penalty term has been to ensure consistency between the desired departure frequency and the departure times of trains, promoting an even spread of departures as much as possible. We define ϑ as the index for the control step of the lower level, where the time interval of each step is R. Then, the objective function for the lower-level controller is defined as

$$J^{\text{low}} = \sum_{i \in \mathscr{I}(k,\vartheta)} \sum_{p \in \mathscr{V}_i} \left(E_i(p) + \zeta \left| \frac{T}{u_\ell(k)} - (d_{i,p} - d_{i-1,p}) \right| \right), \tag{42}$$

where $\mathscr{I}(k, \vartheta)$ denotes the set of indices for trains leaving their first line platforms before the end of phase k but have not yet reached their destination at time step ϑ, $\mathscr{V}_i$ denotes the set of line platforms that train i will visit, and ζ is a weighting factor.

The optimization problem for train scheduling at the lower level is

$$\min_{\mathbf{g}(k,\vartheta)} J^{\text{low}} = \sum_{i \in \mathscr{I}(k,\vartheta)} \sum_{p \in \mathscr{V}_i} \left(E_i(p) + \zeta \left| \frac{T}{u_\ell(k)} - (d_{i,p} - d_{i-1,p}) \right| \right),$$

$$\text{s.t. } (20) - (31), (41), \tag{43}$$

where $\mathbf{g}(k, \vartheta)$ collects the decision variables for trains in set $\mathscr{I}(k, \vartheta)$, i.e., $a_{i,p}$, $d_{i,p}$, and $x_{i,p,b}$, $\forall i \in \mathscr{I}(k, \vartheta)$, $p \in \mathscr{V}_i$, $b \in \mathscr{B}_{i,p}$. Problem (43) contains piecewise constant ("if-then") constraints in (29), which can be reformulated by using the property developed in Bemporad and Morari (1999) (see Transformation property 5.3 below). Therefore, Problem (43) can also be transformed into an MILP problem.

Transformation property 5.3: If we define m_a and M_a as the minimum and maximum values of $a_{i,p}$, respectively, then (29) is equivalent to the following inequalities

$$\begin{cases} a_{i,p} - a_{i',p'} \leq \left(1 - \xi_{i,i',p,p'}\right) \left(M_a - a_{i',p'}\right), \\ a_{i,p} - a_{i',p'} \geq \varepsilon + \xi_{i,i',p,p'} \left(m_a - a_{i',p'} - \varepsilon\right). \end{cases} \tag{44}$$

In the MPC scheme, we solve the optimization problem (43) in a receding horizon way, which enables the decision-making process to include real-time information from the metro network. Solving problem (43) leads to a series of decision variables for all trains $i \in \mathscr{I}(k, \vartheta)$ from their current line platforms to their terminal line platforms. Only the decision variables pertaining to the first interval are executed, following which the prediction window is shortened by one step, and a new problem is formulated considering the newly collected information. The procedure is repeated until the last train in set $\mathscr{I}(k, \vartheta)$ arrives at its terminal line platform.

In this paper, the lower-level controller optimizes the timetable of trains that have not yet reached their destination at phase k. As each train operates from its starting line platform to its terminal line platform, the MPC optimization is terminated until the last planned train arrives at its terminal platform. Therefore, the lower-level controller can be solved in a shrinking horizon manner, i.e., the end of the prediction horizon is fixed and equal to the arrival time of the last train in set $\mathscr{I}(k, \vartheta)$ at its terminal line platform.

Lemma 5.2. *(Recursive Feasibility) Given a feasible solution of problem* (43) *at time step ϑ for trains in the set $\mathscr{I}(k, \vartheta)$ and line platforms in the set $\mathscr{V}_i$, a feasible solution for time step $\vartheta + 1$ can always be found.*

Proof. For trains that have not departed from their depot at the current phase, a feasible solution of problem (43) can always be found by keeping trains at the depot. For trains that have already departed from their first line platform, a feasible solution for time step $\vartheta + 1$ can be found by keeping the solutions (i.e., $a_{i,p}$, $d_{i,p}$, $r_{i,p}$, $\forall i \in \mathscr{I}(k, \vartheta), \forall p \in \mathscr{V}_i$) of the time step ϑ unchanged. In this context, the recursive feasibility of lower-level MPC can be guaranteed. $\square$

In the proposed method, both the higher level and the lower level use centralized MPC. We define the first step of the lower-level controller is indexed by $\vartheta_0(k)$ and the procedure of bi-level MPC for the integration of passenger flows, timetables, and train speed profiles is shown in Algorithm 1.

In the developed bi-level MPC approach, the MPC optimization problems of both levels can be transformed into MILP problems by using the methods introduced in Bemporad and Morari (1999) and Williams (2013). Therefore, we can derive an MILP problem at each level that is an exact equivalence of the original optimization problem. Furthermore, with existing MILP solvers, the resulting optimization problems can be solved.

6. Case study

This section involves conducting simulations to demonstrate the efficacy of the proposed passenger absorption model and bi-level control approach. Firstly, we introduce the metro network and the basic setup utilized in the case study. Subsequently, we evaluate the passenger absorption model based on real-life data from the Beijing metro network. Finally, simulations are conducted to assess the performance of the developed bi-level framework and bi-level MPC approach.

Algorithm 1 Bi-level MPC for the integrated problem

Input: $k_{\max}$, $\vartheta_{\max}(k)$; initial estimate for the variables γ_p, $\bar{r}_p$;
Output: optimized values $a_{i,p}$, $d_{i,p}$

1: $k \leftarrow k_0$
2: **repeat**
3:     $\vartheta \leftarrow \vartheta_0(k)$
4:     solve the higher-level problem (37), get $u_\ell(k)$ and $f_p(k)$
5:     **repeat**
6:         solve problem (43), get $a_{i,p}$ and $d_{i,p}$
7:         implement $a_{i,p}$ and $d_{i,p}$ to real-life network
8:         $\vartheta \leftarrow \vartheta + 1$
9:         collect real-life value of $a_{i,p}$, $d_{i,p}$, and $n_{p,e}(k)$
10:     **until** $\vartheta = \vartheta_{\max}(k)$
11:     $k \leftarrow k + 1$
12:     calculate real-life values of γ_p, $\bar{r}_p$
13: **until** $k = k_{\max}$
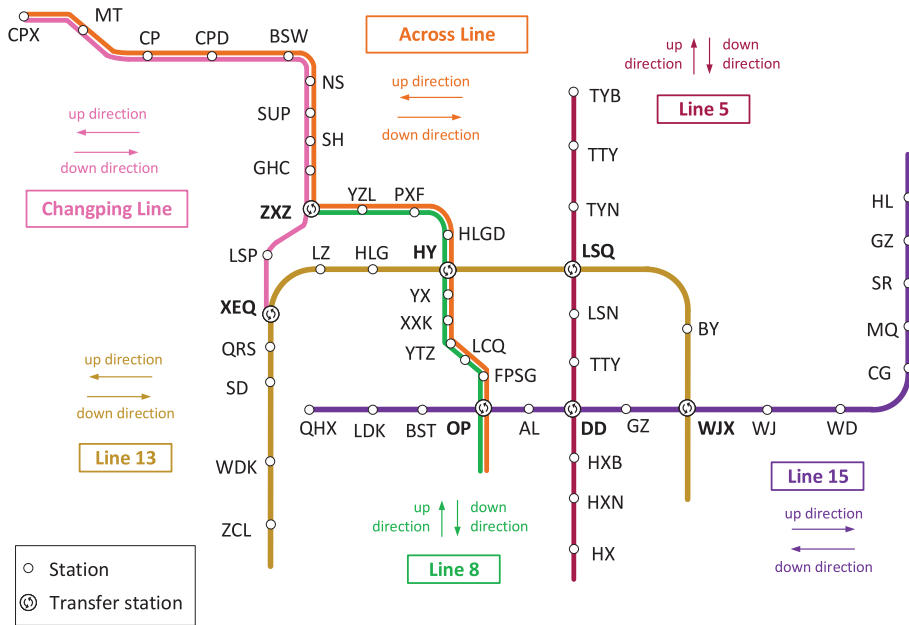
Fig. 8. Layout of the considered metro network.

6.1. Basic setup

In this paper, we carry out the case study based on the real-life passenger flow data from the Beijing metro network. The network is displayed in Fig. 8, which is generated according to the northern part of the Beijing metro network. The network includes six bidirectional lines and 54 stations. Moreover, the network contains seven transfer stations, i.e., Station ZXZ, Station XEQ, Station HY, Station OP, Station WJX, Station LSQ, and Station DD, where passengers can transfer from one line to another to reach their destinations. Transfer passengers are defined as passengers whose route consists of more than one line. Depots are located at both ends of each line.

The across-line operation is one important way to maximize the utilization of infrastructure and to improve passenger satisfaction by reducing the number of transfer activities in the network (especially in big cities like London, Barcelona, and Beijing[2]). Therefore, we add an "Across Line" for the case study to meet the case when different lines use the same physical track and/or platforms.[3] Some passengers at the Across Line (e.g. from CPD to PXF) can use the Across Line to reach their destination and transfer actions

---

Table 6
Parameters for the simulations.

Parameters	Line 8 Changping line across line	Line 13	Line 15
Minimum departure–arrival headway	120 s	120 s	120 s
Regular departure–arrival headway	480 s	180 s	240 s
Maximum dwell time $\tau_p^{\max}$	360 s	360 s	360 s
Minimum dwell time $\tau_p^{\min}$	30 s	30 s	30 s
Regular dwell time $\tau_{i,p}$	60 s	60 s	60 s
Maximum capacity of a train C_{train}	2400 persons	2400 persons	2400 persons
Average transfer time t_p^{transfer}	60 s	60 s	60 s
Phase time T	1800 s	1800 s	1800 s
Number of speed profile options	8 options	8 options	8 options

are not required anymore, so they are not considered to be transfer passengers. There are five lines in Fig. 8, where Changping Line, Line 8, Line 13, and Line 15 are the real-life lines, and the Across Line is added in this paper to simulate the case of cross-line operation. The Across Line uses the same physical platforms as Changping Line from Station CPX to Station GHC, and the same physical platforms as Line 8 from Station ZXZ to Station OP.

The passenger OD data are generated according to the real-life passenger flow data, i.e., the entering and exiting flow data of the Beijing metro network. This information is updated every 30 min. The data we use is for the morning peak hours from 7:00AM. The prediction time window is 1 h. In the case study, we include the case when different lines use the same physical platforms, and the order of trains from different lines at the same physical platform can be adjusted. Table 6 presents the main parameters for the simulation. The parameters are generated based on the real-life timetable of the Beijing metro network. As the Across Line is not yet included in the historical data, in the basic timetable, trains of the Across Line and trains of Changping Line (or Line 8) depart alternately, which means part of the transport capacity that was originally performed by Changping Line (or Line 8) is taken over by the Across Line to reduce the number of transfer actions of passengers, and that change does not affect the total transport capacity or the number of trains needed for the basic timetable. Thus, the original OD demand is divided equally over two lines, so for the basic timetable, half of the departures of the original timetable is then arranged to Changping Line (or Line 8), while the other half is arranged to Across Line. This also means that the total number of trains in the network and the depot does not have to be changed compared with the original timetable. The simulation is coded using MATLAB (R2019b) on an Intel Xeon W2223 CPU (3.60 GHz) with 8 GB RAM. In this paper, we assume passengers' route choices are given a priori, and we consider passengers will choose the route with the shortest travel time for their travel.

As far as we know, there is no well-recognized micro-simulator currently available that includes timetables, passenger OD demands, and train speeds. The model developed by Wang et al. (2015b) is the most elaborate model we noticed in the literature; thus, we use the model of Wang et al. (2015b) as the "accurate model" of the practical passenger dynamics in the railway network. The passenger absorption model combined with the train scheduling model presented in Section 4 are used as prediction models for the train scheduling problem. The basic timetable is generated by using the regular headway and the regular dwell time given in Table 6.

6.2. Assessment of the absorption model

As mentioned in Section 6.1, we select the "accurate model" developed by Wang et al. (2015b) as the benchmark to assess the passenger absorption model. Instead of focusing on the specific times of train arrivals and departures, the passenger absorption model deals with the train departure frequencies in each phase. Thus, we regard the number of passengers as a function of the phase index rather than as a function of time.

The accumulated number of waiting passengers (AWP) and the accumulated number of boarding passengers (ABP) in each line are two main variables in passenger-oriented metro networks. In particular, AWP reflects whether passengers can board trains in time, since if passengers are unable to board trains in the current phase, they should wait for trains in the next phase. ABP reflects the transport capacity of trains.

The simulations are conducted on the network in Fig. 8 based on both the developed model and the "accurate model" of Wang et al. (2015b). We perform the simulation from 7:00 to 15:00 which includes both peak and off-peak hours. We collect the AWP and ABP values in each phase. The required simulation time for the developed model and the accurate model are 2.10 s and 84.24 s, respectively. The relative differences between the absorption model and the "accurate model" for AWP and ABP of each line are displayed in Table 7. The simulation contains 16 phases, and we select the minimum, maximum, and final values of the relative difference among the phases at each line.

It can be observed from Table 7 that Line 8 has the largest average relative difference for AWP, while the largest average relative difference for the ABP value occurs in Line 15. We select Line 8 and Line 15 for visualization, and the corresponding values for AWP and ABP at each time step are respectively depicted in Figs. 9 and 10.

Compared with the accurate model, the passenger flows can be modeled by the absorption model with the largest final error of about 10% and the required simulation time drops with a factor of around 40. Hence, we can conclude that with an acceptable

Table 7
Relative differences of variables for each line.

	Min		Max		Average	
	AWP	ABP	AWP	ABP	AWP	ABP
Changping Line	1.91%	5.12%	7.74%	19.38%	4.76%	7.69%
Line 13	0.50%	0.13%	22.15%	25.79%	7.81%	3.78%
Line 8	1.24%	4.28%	20.48%	33.55%	8.73%	8.66%
Across Line	0.09%	6.32%	17.82%	17.54%	3.53%	9.14%
Line 15	0.61%	0.16%	27.71%	32.19%	5.61%	5.12%
Line 5	0.07%	6.01%	17.12%	19.76%	2.99%	8.96%
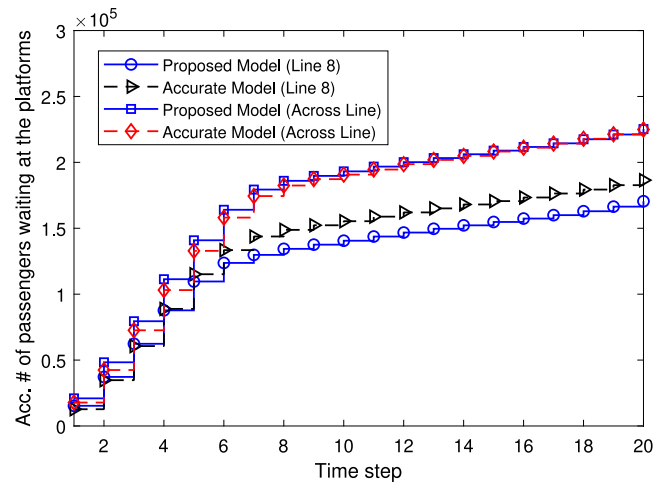
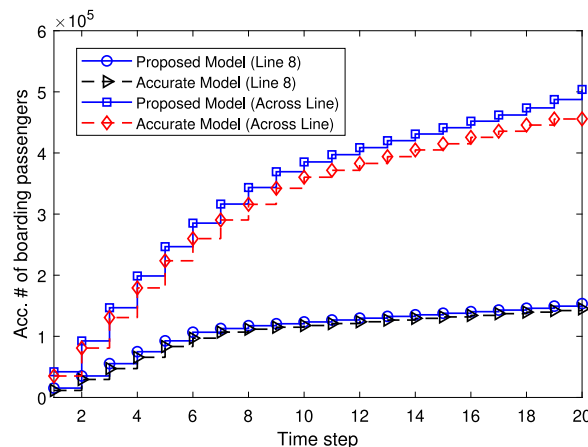**Fig. 9.** Accumulated number of passengers waiting at the platforms in each phase (AWP).

Fig. 10. Accumulated number of boarding passengers in each phase (ABP).

accuracy loss, the absorption model can simulate passenger flows much more efficiently with time-dependent passenger OD demands, which allows more efficient methods for passenger-oriented train scheduling problems. The major loss is that the developed model does not include detailed arrival and departure times of trains, and thus a train scheduling model in the lower level is required to determine the specific departure and arrival times of trains.

6.3. Bi-level optimization based on the absorption model

We first perform simulations of sequentially solving optimization problems at both levels based on the developed model. We also use the single-level optimization approach to solve the integrated problem in a centralized manner. Then, we compare the single-level approach with the proposed bi-level approach based on solution quality and solution time. The single-level optimization

Table 8

Simulation results of different approaches in two cases.

Case	Method	Objective function	CPU time (s)
Unsaturated case	Basic timetable	$8.3925 \cdot 10^3$	–
	Single-level-1-brk	$7.5520 \cdot 10^3$	3106.1
	Single-level-4-brk	$7.5339 \cdot 10^3$	7200.0
	Bi-level approach	$7.5903 \cdot 10^3$	40.5
Over-saturated case	Basic timetable	$9.5186 \cdot 10^3$	–
	Single-level-1-brk	$9.1386 \cdot 10^3$	5250.7
	Single-level-4-brk	$9.1027 \cdot 10^3$	7200.0
	Bi-level approach	$9.1119 \cdot 10^3$	87.0

Table 9

Comparison of different approaches for real-time train scheduling.

Method	Objective function	CPU time (s)	
		t_{avrg}	$t_{\max}$
Basic timetable	$1.4859 \cdot 10^5$	–	–
Single-level MPC	$1.2451 \cdot 10^5$	3181.5	7200.0
Bi-level MPC	$1.1815 \cdot 10^5$	42.4	95.9

problem is a nonlinear nonconvex problem containing integer variables. Compared with the bi-level optimization problem, the single-level counterpart introduces an additional nonlinear term, namely $\frac{T}{u_\ell(k)}$, in (42). The single-level optimization problem can also be converted to an MILP problem by approximating the nonlinear term with linear inequalities using the method of in Williams (2013) (see Appendix C). We use the gurobi to solve all MILP problems. In Appendix C, the nonlinear function is approximated as a piecewise linear function by setting several breakpoints. However, setting more breakpoints can lead to a more accurate approximation of the nonlinear term, while more computation time is required for solving the resulting MILP problem. Therefore, in the case study, we use both one breakpoint and four breakpoints for the approximation of the nonlinear term in the single-level optimization problem, and for simplicity, the corresponding approaches are called single-level-1-brk and single-level-4-brk, respectively.

We evaluate the developed approach in both the over-saturated (i.e., peak hours) and the unsaturated (i.e., off-peak hours) cases. For comparison, both single-level-1-brk and single-level-4-brk are also applied to solve the optimization problem. As our aim is to generate a timetable online, it is required to check whether an approach is real-time implementable. In the case study, the time limit for each method is set to be 7200 s, which is larger than the length of a step (1800 s) because we want each method to have sufficient time to find its solution, and we can compare the relative time of different methods. By using the regular dwell time and departure–arrival headway in Table 6, we can obtain a basic timetable.

The simulation results and CPU times of solving the problem for one step are presented in Table 8. The objective for comparison is the weighted sum of the total passenger travel time and the total energy consumption based on the simulation model. In both the unsaturated case and the over-saturated case, the simulation results indicate that single-level-4-brk performs slightly better than single-level-1-brk with regard to the objective function value. However, the CPU time of single-level-4-brk increases significantly as more integer variables are introduced when adding more breakpoints. As real-time feasibility is important for real-time train scheduling, single-level-1-brk is more suitable for real-life applications than single-level-4-brk.

Compared to the basic timetable, the single-level-1-brk approach, single-level-4-brk approach, and bi-level approach exhibit a performance improvement of 10.01%, 10.23%, and 9.56%, respectively, in the unsaturated case, while the improvement for the over-saturated case is 3.99%, 4.37%, and 4.27%, respectively. The bi-level approach can find its optimal solution very quickly. The CPU times of single-level-1-brk and single-level-4-brk are much larger than the bi-level approach, which implies that single-level optimization may not be a suitable option for real-time train scheduling of large-scale metro networks. The results thus show that the bi-level optimization approach can achieve a balanced trade-off between the solution quality and the computation time.

6.4. Bi-level MPC for real-time train scheduling

In this section, we conduct the case study under the MPC scheme to illustrate the closed-loop performance and the real-time feasibility of the developed approach. The prediction time window of MPC is one hour.

As shown in Section 6.3, the single-level-1-brk approach requires less computation time than single-level-4-brk with an acceptable sacrifice of performance. Considering the real-time feasibility of approaches, we select the single-level-1-brk approach to solve the optimization problems of single-level MPC. The maximum solution time for the MPC optimization problem in each step is set to be 7200 s. The simulation results of single-level MPC and bi-level MPC are displayed in Table 9 and Fig. 11, where the objective function value means the accumulated objective function value for all included simulation times. The performance of the basic timetable is also given for comparison.

The simulation results indicate that, compared with the basic timetable, bi-level MPC can improve the overall performance, i.e., the objective function value, by 20.49%, while the improvement of single-level MPC is 16.21%. The average computation time
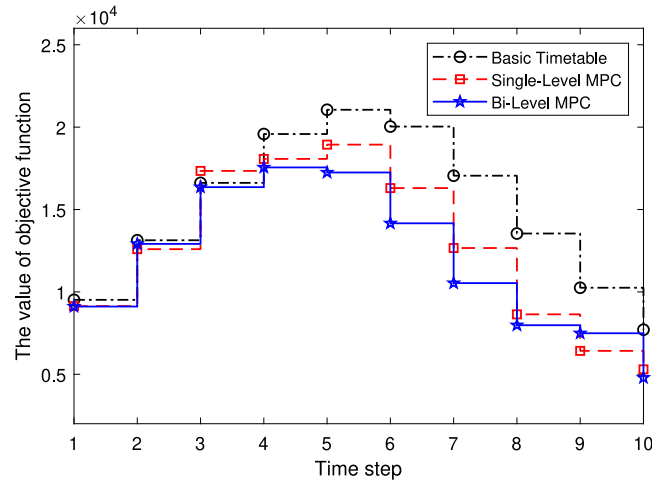
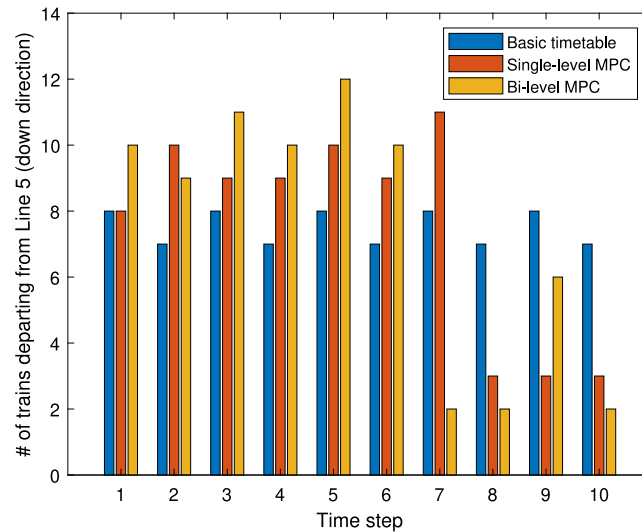**Fig. 11.** Comparison of different approaches for real-time train scheduling.

Fig. 12. Number of trains departing from the first line platform of Line 5 (down direction) at each time step.

for single-level MPC is 3181.5 s. Due to the time limit, single-level MPC cannot always obtain its optimal solution within the given maximum solution time in every MPC step, which influences the solution quality of single-level MPC. The average and maximum solution times of bi-level MPC are 42.4 s and 95.9 s, respectively. Simulation results indicate that bi-level MPC can compute its optimal solution within an acceptable time. However, single-level MPC is not efficient in terms of computation time, and as a result, single-level MPC may not be suitable for real-time implementation in large-scale metro networks.

For further illustration, the number of trains departing from the first line platform of Line 5 (down direction) is shown in Fig. 12 as an example. As time steps 1–6 correspond to the morning peak hours from 7:00AM to 10:00AM, compared with the basic timetable, more trains are scheduled with the single-level and the bi-level MPC approaches to address the large passenger demand, which indicates that bi-level MPC is able to optimize the number of trains departing from each line according to the time-dependent passenger demands.

We select Line 5 (down direction) as a representative line to show the timetables generated by different approaches. The basic timetable of the morning peak hour from 8:00AM to 9:00AM is shown in Fig. 13. The timetables generated by single-level MPC and bi-level MPC from 8:00AM to 9:00AM are respectively exhibited in Fig. 14 and Fig. 15. The time window 8:00AM to 9:00AM corresponds to time steps 3 and 4 in Fig. 13. The above simulation results indicate that the bi-level MPC approach based on the absorption model can generate practically implementable timetables online, which means the bi-level MPC approach can be implemented for real-time train scheduling of metro networks. Furthermore, the line thickness now indicates the number of passengers on board the current train. Then, it can be observed from Figs. 13, 14, and 15 that compared with the basic timetable the optimized timetables allow more trains to transport more passengers so that passenger satisfaction can be improved.
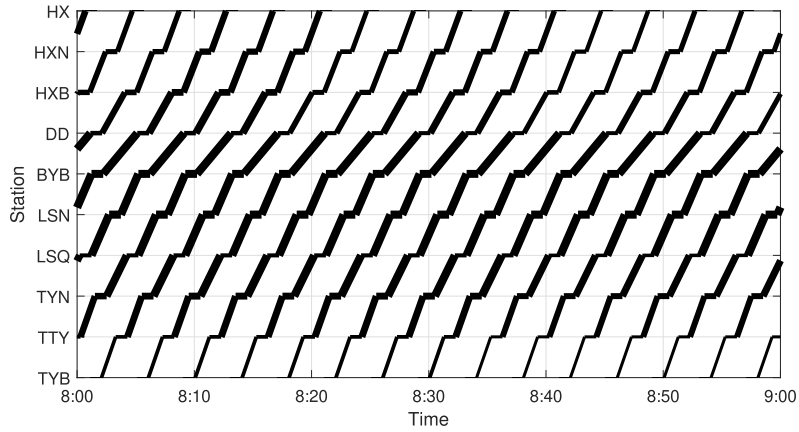
Fig. 13. Basic timetable from station TYB to HX (the line thickness represents the number of passengers on board the train).
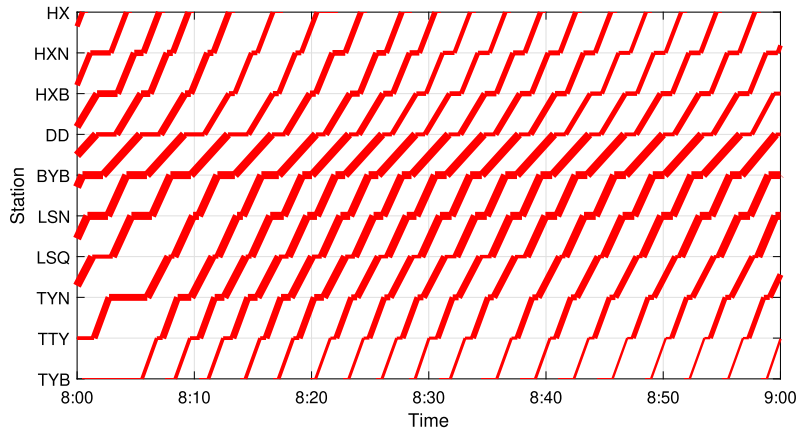

Fig. 14. Timetable obtained by single-level MPC from station TYB to HX (the line thickness represents the number of passengers on board the train).

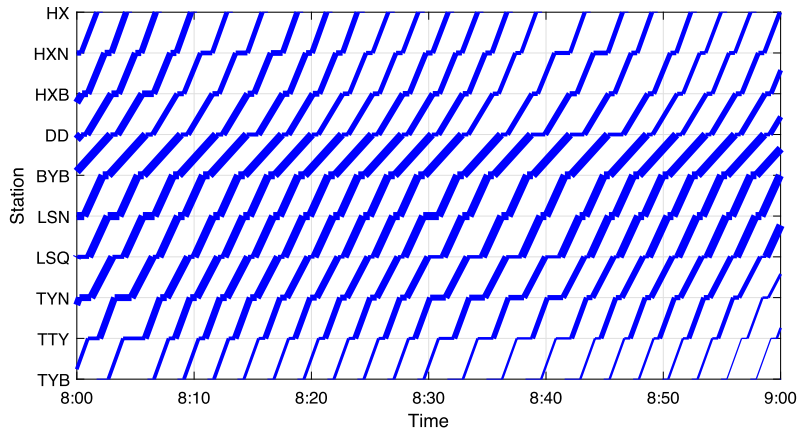**Fig. 15.** Timetable obtained by bi-level MPC from station TYB to HX (the line thickness represents the number of passengers on board the train).

7. Conclusions

In this paper, we have investigated the real-time train scheduling problem considering time-dependent passenger OD demands and train speed profiles in metro networks. We have proposed an extended passenger absorption model to handle time-dependent passenger OD demands and rolling stock circulation in metro networks. The planning time window is divided into several phases,

where the train departure frequency of each platform during each phase is considered. The passenger absorption model has been extended to a bi-level model where detailed timetables, detailed rolling stock circulation, train speed profiles, and train orders are also included. A bi-level MPC approach has been developed for real-time train scheduling of metro networks. The MPC optimization problems in both levels have been transformed into small-scale MILP problems, which enables us to solve them with existing MILP solvers. Numerical experiments show that the developed bi-level MPC approach yields a balanced trade-off between computation time and solution quality, which indicates that the developed model and the proposed bi-level MPC approach can be implemented for real-time train scheduling of metro networks.

The future work includes extending the bi-level framework to include more details of the metro system, e.g., flexible coupling of trains, regenerative braking, etc. Furthermore, uncertain passenger origin–destination demands and stochastic control approaches to deal with these uncertainties will also be a topic of future research. As the current paper only considers time-varying passenger demands, the dynamic interactions between departure frequencies and passenger route choices still ask for further research. Moreover, some learning-based approaches, that integrate learning-based strategies to learn integer variables, can also be studied to solve the resulting optimization problem efficiently while ensuring constraint satisfaction.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix A. Complete mathematical formulation of problem (33)

$$\min_{\boldsymbol{u}(k)} J^{\text{high}} = \sum_{k=k_0}^{k_0+N-1} \left(J^{\text{pass}}(k) + \eta \sum_{p \in \mathscr{P}} f_p(k) \bar{E}_p \right) + L_N(k_0), \tag{45a}$$

subject to

$$\rho_{p,e}(k) = \lambda_{s,p,e}(k) \rho_{s,e}^{\text{station}}(k), \tag{45b}$$

$$n_{p,e}(k+1) = n_{p,e}(k) + \rho_{p,e}(k)T + n_{p,e}^{\text{trans,arrive}}(k) - n_{p,e}^{\text{absorb}}(k), \tag{45c}$$

$$n_{p,e}^{\text{absorb}}(k) = \alpha_{p,e}(k) n_p^{\text{absorb}}(k), \tag{45d}$$

$$n_p^{\text{absorb}}(k) = \min\left(C_p(k), \, n_p^{\text{want}}(k) \right), \tag{45e}$$

$$n_p^{\text{want}}(k) = n_p(k) + \rho_p(k)T + n_p^{\text{trans,arrive}}(k), \tag{45f}$$

$$n_p(k) = \sum_{e \in \mathscr{S}} n_{p,e}(k), \quad \rho_p(k) = \sum_{e \in \mathscr{S}} \rho_{p,e}(k), \quad n_p^{\text{trans,arrive}}(k) = \sum_{e \in \mathscr{S}} n_{p,e}^{\text{trans,arrive}}(k), \tag{45g}$$

$$C_p(k) = f_p(k) \cdot C_{\text{train}} - \sum_{e \in \mathscr{S}} n_{p,e}^{\text{on-board}}(k) + \sum_{e \in \mathscr{S}} n_{p,e}^{\text{alight}}(k), \tag{45h}$$

$$n_{p,e}^{\text{on-board}}(k) = \frac{T - \bar{r}_{\text{p}^{\text{pla}}(p)}}{T} n_{\text{p}^{\text{pla}}(p),e}^{\text{depart}}(k) + \frac{\bar{r}_{\text{p}^{\text{pla}}(p)}}{T} n_{\text{p}^{\text{pla}}(p),e}^{\text{depart}}(k-1), \tag{45i}$$

$$n_{p,q,e}^{\text{transfer}}(k) = \chi_{p,q,e} \, n_{p,e}^{\text{on-board}}(k), \forall q \in \text{cop}(p) \backslash \{p\}, \tag{45j}$$

$$n_{p,e}^{\text{alight}}(k) = \begin{cases} \sum_{q \in \text{cop}(p)} n_{p,q,e}^{\text{transfer}}(k), & \text{if } e \in \mathscr{S} \backslash \{\text{sta}(p)\}, \\ n_{p,e}^{\text{on-board}}(k), & \text{if } e = \text{sta}(p), \end{cases} \tag{45k}$$

$$n_{p,e}^{\text{depart}}(k) = n_{p,e}^{\text{on-board}}(k) - n_{p,e}^{\text{alight}}(k) + n_{p,e}^{\text{absorb}}(k), \tag{45l}$$

$$n_{p,e}^{\text{trans,arrive}}(k) = \sum_{q \in \text{cop}(p) \backslash \{p\}} \left(\frac{T - t_{q,p}^{\text{transfer}}}{T} n_{q,p,e}^{\text{transfer}}(k) + \frac{t_{q,p}^{\text{transfer}}}{T} n_{q,p,e}^{\text{transfer}}(k-1) \right), \tag{45m}$$

$$f_p(k) = \frac{T - \phi_p(k)}{T} u_\ell\left(k - \beta_p(k)\right) + \frac{\phi_p(k)}{T} u_\ell\left(k - \beta_p(k) - 1\right), \tag{45n}$$

$$\sum_{p' \in \text{phy}(p)} f_{p'}(k)\left(h_p^{\min} + \tau_p^{\min}\right) \leq T, \tag{45o}$$

$$\theta_z(k) = \theta_z(k-1) + \sum_{p\in\mathrm{in}(z)} f_p(k) - \sum_{\ell\in\mathrm{out}(z)} u_\ell(k), \forall z \in \mathscr{Z}, \tag{45p}$$

$$\theta_z(k) \geq 0, \forall z \in \mathscr{Z}, \tag{45q}$$

$$w_p(k) = n_p^{\mathrm{want}}(k) - n_p^{\mathrm{absorb}}(k), \tag{45r}$$

$$k = k_0, k_0 + 1, \ldots, k_0 + N - 1,$$

Appendix B. Complete mathematical formulation of problem (43)

$$\min_{\mathbf{g}(k,\vartheta)} J^{\mathrm{low}} = \sum_{i\in\mathscr{I}(k,\vartheta)} \sum_{p\in\mathscr{V}_i} \left(E_i(p) + \zeta \left| \frac{T}{f_{\mathrm{fst}(p)}(k)} - (d_{i,p} - d_{i-1,p}) \right| \right), \tag{46a}$$

subject to

$$d_{i,p} = a_{i,p} + \tau_{i,p}, \tag{46b}$$

$$\tau_p^{\min} \leq \tau_{i,p} \leq \tau_p^{\max}, \tag{46c}$$

$$a_{i,p} = d_{i,\mathrm{p^{pla}}(p)} + r_{i,\mathrm{p^{pla}}(p)}, \tag{46d}$$

$$y_{i,j,\ell,p} = \begin{cases} 1, & \text{if} \quad d_{j,p} \leq d_{i,\ell}; \\ 0, & \text{otherwise.} \end{cases} \tag{46e}$$

$$\sum_{\ell\in\mathrm{out}(z)} \sum_{j\in\mathscr{I}_\ell} y_{i,j,\ell,p} - \sum_{p\in\mathrm{in}(z)} \sum_{j\in\mathscr{I}_p} y_{i,j,\ell,p} \leq N_z^{\mathrm{train}}, \tag{46f}$$

$$r_p^{\min} \leq r_{i,p} \leq r_p^{\max}, \tag{46g}$$

$$r_{i,p} = \sum_{b\in\mathscr{B}_{i,p}} x_{i,p,b}\, r_{i,p,b}, \tag{46h}$$

$$\sum_{b\in\mathscr{B}_{i,p}} x_{i,p,b} = 1, \tag{46i}$$

$$a_{i,p} \geq d_{\mathrm{p^{tra}}(i),p} + h_p^{\min}, \tag{46j}$$

$$\xi_{i,i',p,p'} = \begin{cases} 1, & \text{if } a_{i,p} \leq a_{i',p'}; \\ 0, & \text{otherwise.} \end{cases} \tag{46k}$$

$$a_{i,p} - d_{i',p'} \geq h_p^{\min} - M_{\mathrm{a}}(1 - \sigma_{p,p'} + \xi_{i,i',p,p'}), \tag{46l}$$

$$\xi_{i,i',p,p'} + \xi_{i',i,p',p} = 1, \tag{46m}$$

$$E_i(p) = \sum_{b\in\mathscr{B}_{i,p}} x_{i,p,b} E_{i,b}(p). \tag{46n}$$

Appendix C. Transformation of inverse proportionality functions of real variables

A piecewise affine function can be used to approximate the inverse proportionality function of the real variable $h(y) = 1/y$:

$$h_{\mathrm{PWA}}(y) = \begin{cases} \alpha_1 y + \beta_1 & \text{if } y \leq Y_1, \\ \alpha_2 y + \beta_2 & \text{if } y > Y_1, \end{cases} \tag{47}$$

where α_1, α_2, β_1, and β_2 are parameters that can be computed by the least squares approach; Y_1 is the breakpoint of the subdomain. It is worth noting that the approximation can be conducted by only concentrating on the effective section of the domain where the value of y can be taken in real life so that we can reduce the approximation error. Moreover, we can also reduce the approximation error by adding more breakpoints in (47).

Appendix D. Sensitivity analysis

To show the influence of the train departure frequency and the train speed profile, we have performed a sensitivity analysis for the following four cases: (1) both the departure frequency and the train speed profile are changed, (2) only the departure frequency is changed, (3) only the train speed profile is changed, and (4) both the departure frequency and the train speed profile are fixed. The simulation results are shown in Table 10.

It can be observed from Table 10 that compared with case 1 only changing the speed profile (i.e., case 2) can reduce the total energy consumption by 9.07% while sacrificing the total passenger travel time of 0.84%. Thus, including the train speed profiles in the train scheduling problem can help to reduce energy consumption with a limited sacrifice of the passenger travel time. Furthermore, only changing the departure frequency (i.e., case 3) can reduce the total energy consumption by 19.94% while also reducing the total passenger travel time by 19.46%. By optimizing the train departure frequency, more trains are scheduled in peak hours to transport more passengers while fewer trains are used in off-peak hours to reduce energy consumption; hence, both the total energy consumption and the total passenger travel time can be reduced. Optimizing both the departure frequency and the train speed profile (i.e., case 4) can reduce the total energy consumption by 25.32% while also reducing the total passenger travel time by 18.71%, which yields the best overall performance and still has an acceptable online computation time.

Table 10
Sensitivity analysis for real-time train scheduling.

	Departure frequency	Speed profile	Total energy consumption (kWh)	Total travel time (h)	CPU time (s)	
					t_{avrg}	t_{max}
Case 1	Fixed	Fixed	$1.0778 \cdot 10^5$	$6.0943 \cdot 10^5$	–	–
Case 2	Fixed	Changeable	$9.8008 \cdot 10^4$	$6.1456 \cdot 10^5$	4.6	5.7
Case 3	Changeable	Fixed	$8.6288 \cdot 10^4$	$4.9081 \cdot 10^5$	34.8	68.4
Case 4	Changeable	Changeable	$8.0492 \cdot 10^4$	$4.9540 \cdot 10^5$	42.4	95.9

References

Bemporad, A., Morari, M., 1999. Control of systems integrating logic, dynamics, and constraints. Automatica 35 (3), 407–427.

Bešinović, N., Wang, Y., Zhu, S., Quaglietta, E., Tang, T., Goverde, R.M., 2022. A matheuristic for the integrated disruption management of traffic, passengers and stations in urban railway lines. IEEE Trans. Intell. Transp. Syst. 23 (8), 10380–10394.

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

Caimi, G., Fuchsberger, M., Laumanns, M., Lüthi, M., 2012. A model predictive control approach for discrete-time rescheduling in complex central railway station areas. Comput. Oper. Res. 39 (11), 2578–2593.

Canca, D., Barrena, E., De-Los-Santos, A., Andrade-Pineda, J.L., 2016. Setting lines frequency and capacity in dense railway rapid transit networks with simultaneous passenger assignment. Transp. Res. B 93, 251–267.

Cavone, G., van den Boom, T., Blenkers, L., Dotoli, M., Seatzu, C., De Schutter, B., 2022. An MPC-based rescheduling algorithm for disruptions and disturbances in large-scale railway networks. IEEE Trans. Autom. Sci. Eng. 19 (1), 99–112.

Corman, F., 2020. Interactions and equilibrium between rescheduling train traffic and routing passengers in microscopic delay management: A game theoretical study. Transp. Sci. 54 (3), 785–822.

Cury, J., Gomide, F., Mendes, M., 1980. A methodology for generation of optimal schedules for an underground railway system. IEEE Trans. Automat. Control 25 (2), 217–222.

Garey, M.R., Johnson, D.S., 1979. Computers and Intractability, Vol. 174. freeman San Francisco.

Gkiotsalitis, K., Cats, O., 2022. Optimal frequency setting of metro services in the age of COVID-19 distancing measures. Transp. A: Transp. Sci. 18 (3), 807–827.

Haahr, J.T., Wagenaar, J.C., Veelenturf, L.P., Kroon, L.G., 2016. A comparison of two exact methods for passenger railway rolling stock (re) scheduling. Transp. Res. E: Logist. Transp. Rev. 91, 15–32.

Higgins, A., Kozan, E., 1998. Modeling train delays in urban networks. Transp. Sci. 32 (4), 346–357.

Hou, Z., Dong, H., Gao, S., Nicholson, G., Chen, L., Roberts, C., 2019. Energy-saving metro train timetable rescheduling model considering ATO profiles and dynamic passenger flow. IEEE Trans. Intell. Transp. Syst. 20 (7), 2774–2785.

Leurent, F., Chandakas, E., Poulhès, A., 2014. A traffic assignment model for passenger transit on a capacitated network: Bi-layer framework, line sub-models and large-scale application. Transp. Res. C 47, 3–27.

Li, S., Dessouky, M.M., Yang, L., Gao, Z., 2017. Joint optimal train regulation and passenger flow control strategy for high-frequency metro lines. Transp. Res. B 99, 113–137.

Li, C., Ma, J., Luan, T.H., Zhou, X., Xiong, L., 2018. An incentive-based optimizing strategy of service frequency for an urban rail transit system. Transp. Res. E: Logist. Transp. Rev. 118, 106–122.

Liu, X., Dabiri, A., De Schutter, B., 2022. Timetable scheduling for passenger-centric urban rail networks: Model predictive control based on a novel absorption model. In: 2022 IEEE Conference on Control Technology and Applications. CCTA, IEEE, pp. 1147–1152.

Liu, X., Dabiri, A., Wang, Y., De Schutter, B., 2023. Modeling and efficient passenger-oriented control for urban rail transit networks. IEEE Transactions on Intelligent Transportation Systems 24 (3), 3325–3338.

Luan, X., Corman, F., 2022. Passenger-oriented traffic control for rail networks: An optimization model considering crowding effects on passenger choices and train operations. Transp. Res. B 158, 239–272.

Luan, X., Wang, Y., De Schutter, B., Meng, L., Lodewijks, G., Corman, F., 2018. Integration of real-time traffic management and train control for rail networks-part 1: Optimization problems and solution approaches. Transp. Res. B 115, 41–71.

Mayne, D.Q., 2014. Model predictive control: Recent developments and future promise. Automatica 50 (12), 2967–2986.

Mayne, D.Q., Rawlings, J.B., Rao, C.V., Scokaert, P.O., 2000. Constrained model predictive control: Stability and optimality. Automatica 36 (6), 789–814.

Mo, P., Yang, L., D'Ariano, A., Yin, J., Yao, Y., Gao, Z., 2020. Energy-efficient train scheduling and rolling stock circulation planning in a metro line: a linear programming approach. IEEE Trans. Intell. Transp. Syst. 21 (9), 3621–3633.

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. Transp. Res. B 76, 117–135.

Noursalehi, P., Koutsopoulos, H.N., Zhao, J., 2022. Dynamic origin-destination prediction in urban rail systems: A multi-resolution spatio-temporal deep learning approach. IEEE Trans. Intell. Transp. Syst. 23 (6), 5106–5115.

Pu, S., Zhan, S., 2021. Two-stage robust railway line-planning approach with passenger demand uncertainty. Transp. Res. E Logist. Transp. Rev. 152, 102372.

van den Boom, T.J., De Schutter, B., 2006. Modelling and control of discrete event systems using switching max-plus-linear systems. Control Eng. Pract. 14 (10), 1199–1211.

Wang, Y., D'Ariano, A., Yin, J., Meng, L., Tang, T., Ning, B., 2018. Passenger demand oriented train scheduling and rolling stock circulation planning for an urban rail transit line. Transp. Res. B 118, 193–227.

Wang, Y., Ning, B., Tang, T., van den Boom, T.J., De Schutter, B., 2015a. Efficient real-time train scheduling for urban rail transit systems using iterative convex programming. IEEE Trans. Intell. Transp. Syst. 16 (6), 3337–3352.

Wang, Y., Tang, T., Ning, B., van den Boom, T.J., De Schutter, B., 2015b. Passenger-demands-oriented train scheduling for an urban rail transit network. Transp. Res. C 60, 1–23.

Wang, Y., Zhu, S., Li, S., Yang, L., De Schutter, B., 2022. Hierarchical model predictive control for on-line high-speed railway delay management and train control in a dynamic operations environment. IEEE Trans. Control Syst. Technol. 30 (6), 2344–2359.

Williams, H.P., 2013. Model Building in Mathematical Programming. John Wiley & Sons.

Yin, J., D'Ariano, A., Wang, Y., Yang, L., Tang, T., 2021. Timetable coordination in a rail transit network with time-dependent passenger demand. European J. Oper. Res. 295 (1), 183–202.

Yin, J., Yang, L., Tang, T., Gao, Z., Ran, B., 2017. Dynamic passenger demand oriented metro train scheduling with energy-efficiency and waiting time minimization: Mixed-integer linear programming approaches. Transp. Res. B 97, 182–213.

Zhao, Y., Li, D., Yin, Y., Zhao, X., 2023. Integrated optimization of demand-driven timetable, train formation plan and rolling stock circulation with variable running times and dwell times. Transp. Res. E Logist. Transp. Rev. 171, 103035.

Zhu, Y., Goverde, R.M.P., 2019. Railway timetable rescheduling with flexible stopping and flexible short-turning during disruptions. Transp. Res. B 123, 149–181.