Data-Driven Train Timetabling with Contextual Information

Jiateng Yin[1], Hanxiao Fan[1], D'Ariano Andrea[2], Lixing Yang[2]
1. Beijing Jiaotong University, Beijing, China
2. Rome Tre University, Roma, Italy

Extended abstract submitted for presentation at the 12th Triennial Symposium on Transportation Analysis conference (TRISTAN XII)
June 22-27, 2025, Okinawa, Japan

February 27, 2025

Keywords: (Train timetabling; demand-driven; predict-then-optimize; urban rail systems)

1 INTRODUCTION

In the urban rail systems of large cities like Beijing, Shanghai, and Tokyo, passenger demand fluctuates significantly throughout a day. For example, at Tiantongyuan station, one of the largest metro stations in Beijing, around 84% of the total daily passenger volume arrives between 7:00 AM and 9:00 AM. Consequently, designing a demand-oriented train timetable that adjusts service frequency based on time-varying passenger demand is essential for balancing service quality with operational costs in urban rail systems.

With the increasing availability of passenger travel data, such as from Automated Fare Collection (AFC) systems or mobile phones, the demand-oriented train timetabling problem (DTP) has gained significant research attention in recent years. Barrena *et al.* (2014) developed three mixed-integer linear programming (MILP) formulations for DTP aimed at minimizing the average passenger waiting time. Their results demonstrate that, compared to a constant-frequency timetable, a demand-oriented timetable reduces passenger waiting time by an average of 30%. Using a time-dependent origin-destination (OD) matrix as input, Niu *et al.* (2015) proposed two mixed-integer programming (MIP) formulations with quadratic objective functions and linear constraints for DTP with skip-stop patterns. This model was further extended by Yin *et al.* (2017), incorporating bi-objective functions to minimize both passenger waiting time and system energy consumption. More recently, Liu *et al.* (2023) introduced a two-phase approach for the train scheduling problem that integrates timetables, time-varying passenger flows, and train speed profiles, with the objective of reducing both passenger travel time and train energy consumption.

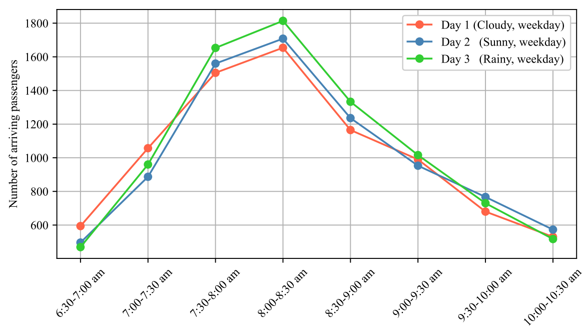

Figure 1 – *Number of arriving passengers at Sihui station, Beijing metro (Aug. 8-10, 2023)*

Overall, the studies mentioned above rely on the average demand over a period of days, often represented by an expected time-dependent OD matrix or arrival rate. However, in practice, metro passenger volumes are uncertain and influenced by various factors such as weather, day of the week (weekday or weekend), and even temperature. Figure 1 presents historical data from Sihui station in the Beijing metro, showing passenger arrival rates over three days with different weather conditions. The data reveal that passenger numbers tend to increase on rainy days (day 3), likely due to greater road congestion caused by the weather.

By analyzing a large-amount of historical data, our study proposes a data-driven approach for DTP considering contextual information. We particularly adopt an integrated "predict-and-optimize" (IPO) scheme for DTP. Our IPO is a variant of the "smart predict-then-optimize" (Elmachtoub & Grigas, 2022), which involves a prediction model to predict the passenger demand of the next day according to the forecast contextual information and a decision model to optimize a train timetable given the predicted passenger demand. The overall aim of IPO is to minimize the expected waiting time of all passengers. To solve the problem, we transform the model into a bi-level programming model and further into a large-scale single-level MILP. Solving the MILP is traceable with commercial solvers but may take a very long computational time. So we propose a column-and-row generation decomposition scheme that iteratively solves small-scale problems. We test our approach on the real-world data of Beijing metro and we compare our approach with traditional stochastic programming approaches, involving sample-average-approximation (SAA) and robust optimization (RO), etc.

2 Problem Description

Our study considers a typical urban rail system consisting of a set of stations, denoted by $S = \{1, 2, \cdots, |S|\}$, with one or multiple depots. The rail company operates a fleet of physical trains (or rolling stock resources), denoted by K, with each train having a passenger loading capacity of C.

A train schedule is described by a sequence of actions over the planning horizon (e.g., 5:00 am to 12:00 am), defining the frequency and explicit arrival/departure times of trains at each station (Schettini *et al.*, 2022). According to the schedule for each operational day, the trains depart from the depot, dwell at the stations $1, \cdots, |S|$ for passengers to board and alight, and finally return to the depot. This entire process is referred to as a *cycle*. After completing a cycle and returning to the depot, each train can be dispatched for a new cycle (Van Lieshout, 2021).

In our study, we collected a large set of historical data samples, denoted by:

$$\mathcal{D} = \{(X^1, Y^1), \cdots, (X^i, Y^i), \cdots, (X^N, Y^N)\} \tag{1}$$

Each tuple (X^i, Y^i) represents the collected data from the i-th day, where $X^i = \{x_1^i, x_2^i, \cdots x_p^i\}$ denotes the contextual information matrix (i.e., features), which includes the date (weekday, weekend, or holiday), weather, and hourly temperature; $Y^i = \{y_{1,1}^i, y_{1,2}^i, \cdots, y_{|S|,|T|}^i\}$ represents the passenger demand throughout the operational period (e.g., 5:00 am to 23:00 pm). Here, the operational period is discretized into a set of equal intervals $T = \{1, 2, \cdots, |T|\}$, with each interval lasting, for instance, 10 seconds. Thus, for any $t \in T$ and $s \in S$, $y_{s,t}^i$ represents the number of arriving passengers at station s during time t.

Different from the existing studies, which rely solely on historical passenger demand data for DTP, our *IPO* approach aims to optimize the train timetable for the next operational day by leveraging forecasted contextual information (e.g., weather, temperature). The objective is to improve service quality by aligning with actual demand distribution and minimizing the expected passenger waiting time.

3 Methodology

Our IPO approach involves a decision model that generates an optimal train timetable, based on predicted demand from a prediction model that uses contextual feature information as input.

3.1 Decision model

Given the predicted demand $\hat{Y} = \{y_{s,t} \geq 0 | s \in S, t \in T\}$, the decision model for DTP is formulated as follows:

$$
\begin{aligned}
f(\hat{Y}) = \min \quad & \sum_{s \in S} \sum_{t \in T} t \left(\sum_{\tau \in T_t} (y_{s,\tau} - \beta_{i\tau}) \right) \\
\text{s.t.} \quad & \sum_{\tau \in T_t^H} z_\tau \leq 1, && \forall t \in T \\
& \sum_{\tau \in T_t^C} z_\tau \leq K, && \forall t \in T \\
& \sum_{\tau \in T_t} (\beta_{i,\tau} - y_{s,t}) \leq 0, && \forall t \in T, s \in S \\
& \beta_{s,t} \leq C \sum_{\tau \in T_{s,t}} z_\tau, && \forall t \in T, s \in S \\
& z_t \in \{0,1\}, \forall t \in T, \beta_{s,t} \geq 0, \forall t \in T, s \in S
\end{aligned}
\tag{2}
$$

In the above formulation, we define binary decision variables $z_t \in \{0, 1\}$ to represent the departure of a train at time t, where $z_t = 1$ indicates a train departure; continuous variables $\beta_{s,t}$ to represent the volume of boarding passengers at station s at time t. The objective function $f(\hat{Y})$ corresponds to the waiting time of passengers, where $T_t = \{\tau \in T | \tau < t\}$. The first constraint ensures the safety headway between trains, where $T_t^H = \{\tau \in T \mid t < \tau \leq t + t_{\text{safe}}\}$, and t_{safe} is the minimum headway time required to maintain safe train following distances. The second constraint restricts the rolling stock circulations, where $T_t^C = \{\tau \in T \mid t < \tau \leq t + t_C\}$, with t_C representing the circulation time. The third and fourth constraints deal with passenger alighting and boarding processes, which will be further detailed in the conference.

3.2 Decision with prediction

Model (2) relies on the predicted passenger demand $\hat{Y}$. To achieve this, we adopt a MIMO linear regression model that predicts the demand using contextual information $\hat{X}$ of the next day. Specifically, let $\boldsymbol{A} \in \mathbb{R}^{N \times T}$ represent the matrix of coefficients and $\boldsymbol{\theta} \in \mathbb{R}^{1 \times T}$ represent the vector of bias, both of which are to be determined.

Traditionally, $\boldsymbol{A}$ and $\boldsymbol{\theta}$ are trained directly using historical data $\mathcal{D}$, with the objective of minimizing prediction error (via the least square method). Once the optimal matrix $\boldsymbol{A}$ and vector $\boldsymbol{\theta}$ are determined, we can use $\hat{Y} = \boldsymbol{A}^T \hat{X} + \boldsymbol{\theta}$ to generate the train timetable by solving model (2). This approach constitutes a *predict then optimize* (PTO) method.

The key idea is that while PTO minimizes prediction error, it does not necessarily minimize decision error, i.e., the expected passenger waiting time. Therefore, our IPO formulation integrates model (2) with the prediction model, resulting in a bi-level optimization formulation:

$$
\begin{aligned}
\text{(B-IPO)} \ \min_{\boldsymbol{A},\boldsymbol{\theta}} \quad & \frac{1}{N} \sum_{i=1}^{N} \left| f(\boldsymbol{A}^T X^i + \boldsymbol{\theta}) - f(Y^i) \right| \\
& \boldsymbol{A} \in \mathbb{R}^{N \times T}, \boldsymbol{\theta} \in \mathbb{R}^{1 \times T}
\end{aligned}
\tag{3}
$$

The objective function of (3) minimizes the difference between the passenger waiting time based on the predicted demand (i.e., $f(\boldsymbol{A}^T X^i + \boldsymbol{\theta})$) and the waiting time based on realized demand (i.e., $f(Y^i)$) over the historical data set $\mathcal{D}$.

Model (3) is a typical bi-level optimization problem, which can be challenging to solve. We reformulate it into a large-scale MILP and develop an iterative method based on column-and-row generation to efficiently solve it. Further details will be presented at the conference.

4 Results

We conduct numerical experiments using field data from the Beijing Metro Batong Line, collected between August 1 and August 30, 2023 (i.e., $N = 30$). We also construct a total of 20 instances as the testing dataset, with varying values of D, T, and K. We compare our IPO approach with four benchmarks: 1) **AVE**, which uses average passenger demand to optimize DTP, a common method in both the literature Yin *et al.* (2017) and practice; 2) **SAA**, which optimizes DTP under stochastic scenarios generated from historical data $\mathcal{D}$; 3) **RO**, which also adopts discrete scenarios but minimizes under the worst-case scenario; and 4) **PTO**, as introduced earlier.

Table 1 – *Performance comparison on the testing data set*

	Parameters			Methods ($\times 10^7$)				
Instances	$\|D\|$	$\|T\|$	$\|K\|$	AVE	SAA	RO	PTO	IPO
I-1	5	300	18	2.63	2.51	2.68	2.32	2.25
I-2	7	300	18	2.56	2.53	2.62	2.47	2.32
I-3	10	300	18	2.82	2.64	2.65	2.68	2.52
I-4	12	200	18	3.52	3.33	3.55	3.23	3.12
I-5	10	200	18	3.10	3.11	3.21	3.11	2.96
Average	-	-	-	2.93	2.82	2.94	**2.76**	**2.63**

Table 1 presents (a subset of) the performance comparison between our IPO and the four benchmarks. Both PTO and IPO significantly outperform the other three methods, highlighting the importance of incorporating contextual information into DTP. More importantly, our IPO further reduces passenger waiting time by approximately 5% compared to PTO. Additional results and analysis will be presented at the conference.

References

Barrena, Eva, Canca, David, Coelho, Leandro C, & Laporte, Gilbert. 2014. Exact formulations and algorithm for the train timetabling problem with dynamic demand. *Computers & Operations Research*, **44**, 66–74.

Elmachtoub, Adam N, & Grigas, Paul. 2022. Smart "predict, then optimize". *Management Science*, **68**(1), 9–26.

Liu, Xiaoyu, Dabiri, Azita, Xun, Jing, & De Schutter, Bart. 2023. Bi-level model predictive control for metro networks: Integration of timetables, passenger flows, and train speed profiles. *Transportation Research Part E: Logistics and Transportation Review*, **180**, 103339.

Niu, Huimin, Zhou, Xuesong, & Gao, Ruhu. 2015. Train scheduling for minimizing passenger waiting time with time-dependent demand and skip-stop patterns: Nonlinear integer programming models with linear constraints. *Transportation Research Part B: Methodological*, **76**, 117–135.

Schettini, Tommaso, Jabali, Ola, & Malucelli, Federico. 2022. Demand-driven timetabling for a metro corridor using a short-turning acceleration strategy. *Transportation Science*, **56**(4), 919–937.

Van Lieshout, Rolf N. 2021. Integrated periodic timetabling and vehicle circulation scheduling. *Transportation Science*, **55**(3), 768–790.

Yin, Jiateng, Yang, Lixing, Tang, Tao, Gao, Ziyou, & Ran, Bin. 2017. Dynamic passenger demand oriented metro train scheduling with energy-efficiency and waiting time minimization: Mixed-integer linear programming approaches. *Transportation Research Part B: Methodological*, **97**, 182–213.