Theory and Practice of Science and Technology
2025, VOL. 6, NO. 8, 66-70
DOI: 10.47297/taposatWSP2633-456912.20250608

Research on Dynamic Train Scheduling for Shanghai Metro Based on Enhanced PageRank and Data Fusion

Kaihui Han

School of Mathematics and Statistics, Shandong University of Technology, Zibo, Shandong, 255022, China

ABSTRACT

This paper addresses the issues of idle capacity and peak congestion in Shanghai Metro caused by uneven spatiotemporal passenger flow distribution. It proposes a dynamic train scheduling model integrating an improved PageRank algorithm. The study pioneers a three-part data fusion architecture combining "section flow data, timetable data, and transfer station observations." By employing the Expectation-Maximization (EM) algorithm to back-propagate dynamic origin-destination (OD) matrices from publicly available section flow data, it effectively resolves the challenge of obtaining complete passenger flow data. At the algorithmic level, an innovative dual-mechanism PageRank model is designed: a time-sensitive factor (λ =0.7) quantifies passenger flow variation rates, while capacity-adaptive weights (β=1.2) reflect station load states, precisely capturing the dynamic characteristics of station importance. Based on this, a multi-objective spatiotemporal network optimization model is constructed, embedding station importance parameters into decision variable constraints such as headway intervals and train formation quantities. Experimental results demonstrate: Compared to fixed timetables, this model reduces average passenger waiting time during peak hours by 22.4% (measured from 158s to 123s), decreases train empty-run rates during off-peak periods by 31.7%, and achieves a 17.3% improvement in multi-objective Pareto solution sets over the genetic algorithm benchmark. Field validation at People's Square Station (sample size: 300 passengers) demonstrated a waiting time prediction error of only 3.2 seconds (MAPE=2.1%) and an overall reliability of 92.7%, providing an innovative solution for dynamic scheduling in urban rail transit.

KEYWORDS

Dynamic train scheduling; PageRank algorithm; Data fusion; Metro scheduling; Multi-objective optimization

1 Introduction

As one of the world's largest urban rail transit networks, Shanghai Metro has reached an operational mileage of 831 kilometers (as of 2024) with a daily ridership exceeding 12 million passengers. However, its operations face severe challenges: during peak hours, sections of core lines like Line 1 and Line 9 experience load factors exceeding 140%, leading to passenger congestion at stations. Conversely, during off-peak hours, suburban lines operate at less than 40% capacity utilization, resulting in significant resource wastage. This temporal and spatial mismatch between transport capacity allocation and passenger demand not only degrades the travel experience but also substantially increases operational costs. The 2024 operational report of the Shanghai Metro Group indicates that energy waste resulting from inefficient scheduling reaches as high as 120 million yuan annually.

Current subway train scheduling research primarily employs static optimization methods. Liu Zhaoyi et al [1]. optimized station layouts using genetic algorithms, Dong Shengwei [2] predicted short-term passenger flows with BP neural networks, and Shi Jungang et al [3]. optimized train stop plans through integer linear programming. While these approaches have achieved some success, they exhibit clear limitations: they struggle to respond in real-time to sudden passenger flow fluctuations and rely on complete historical origin-destination (OD) data, which is often difficult to obtain due to privacy protection requirements. More critically, existing research fails to effectively balance conflicting objectives such as passenger experience (minimizing waiting time), operational costs (minimizing energy consumption), and system efficiency (balancing load factors).

The PageRank algorithm, as a tool for quantifying the influence of network nodes, has demonstrated powerful analytical capabilities in recent years across social networks [4] and urban planning [5]. This algorithm transmits weights through inter-node linkages, exhibiting topological similarity to the passenger flow connections between stations in subway networks. However, the static nature of traditional PageRank struggles to adapt to the intense fluctuations of subway passenger flows, and its application in transportation scheduling has yet to be thoroughly explored.

This research achieves breakthroughs on three levels:

(1) Data layer: Innovatively proposes a three-component fusion architecture integrating "cross-section flow-schedule-transfer observation." Utilizes the EM algorithm to infer the entire network's dynamic origin-destination (OD) matrix from passenger flows at 10 key cross-sections;

(2) Algorithm level: Designs a dual-mechanism PageRank model incorporating time-sensitive factors (λ =0.7) and

*** Corresponding Author:** Kaihui Han, 2686805496@qq.com

capacity-adaptive weights (β=1.2) to address the challenge of modeling time-varying passenger flow characteristics;

(3) Model Level: Established a multi-objective spatiotemporal network optimization framework, embedding station importance parameters as core constraints in train scheduling decisions.

The study received data support from the Shanghai Municipal Commission of Transport and conducted field observations at People's Square Station, providing a solid foundation for model validation.

2 Research methodology

2.1 Data fusion framework

This study employs a fusion analysis framework utilizing three data sources:

(1) Section Passenger Flow Data: Time-series passenger flow statistics from 10 key sections published in the Shanghai Rail Transit Development Report 2023, including hub sections such as Line 1 Xinzhuang→Xujiahui (8,520 passengers/hour during morning peak) and Line 2 Pudong→Nanjing Road (11,387 passengers/hour during morning peak). These sections represent over 75% of the network's passenger corridors and are highly representative.

(2) Timetable Data: Analyzed the 2024 timetable published on the Shanghai Metro official website to extract key operational parameters including minimum headway (120 seconds), station dwell time (25 – 45 seconds), and train formation type (Type A trains with a capacity of 328 passengers per car).

(3) Transfer Station Observation Data: During the morning peak hours (7:30 – 9:00) from March 12 to 14, 2024, a systematic sampling method was used to record the waiting times of 300 passengers at People's Square Station. The measured average waiting time was 158 seconds (standard deviation 23 seconds). The observation period featured clear weather, eliminating external interference factors.

Due to privacy restrictions preventing access to complete passenger origin-destination (OD) data, this study employed the Expectation-Maximization (EM) algorithm to infer the OD matrix from cross-sectional flow data. The mathematical principle involves maximizing the following likelihood function:

$$L\left(\theta|F\right)=\Pi_{t,k}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(F_k^t-\Sigma_{i,j\in\Gamma_k}Q_{i,j}^t\right)^2}{2\sigma^2}\right)$$

where `$F_k^t$` denotes the observed passenger flow at section k during time period t, `$Q_{i,j}^t$` represents the origin-destination (OD) flow from station i to j, `$\Gamma_k$` is the set of OD pairs covered by section k, and `$\sigma^2$` is the random error variance. The algorithm iteratively executes two steps: the Expectation step (E-step) calculates the expected value of OD flows under current parameters, while the Maximization step (M-step) updates parameters to maximize the likelihood function. After 100 iterations, the algorithm converged with an ODM reconstruction error (RMSE) of 3.8%. This method effectively addresses the issue of missing complete passenger flow data, providing robust data support for subsequent modeling.

2.2 Improved pagerank algorithm design

The traditional PageRank algorithm evaluates webpage importance through hyperlink relationships between pages, with its core formula defined as:

$$PR_i=\left(1-d\right)+d\Sigma_{j\in B_i}\frac{PR_j}{L\left(j\right)}$$

where d=0.85 is the damping factor,B_i denotes the set of webpages pointing to node i, and$L\left(j\right)$ represents the number of outbound links from webpage j. The static nature of this model struggles to adapt to the dynamic passenger flow changes in subway networks.

To overcome this limitation, this study proposes two core improvements:

Time-sensitive factor: quantifies the intensity of passenger flow changes over time at stations

$$\alpha_i^t=\lambda\cdot\frac{Q_i^t-Q_i^{t-\Delta t}}{Q_i^{t-\Delta t}}\,(\lambda=0.7)$$

where Q_i^t represents passenger flow at station i during time period t, andΔt denotes the number of outbound links from station i within a 15-minute time window. Whenα_i^t indicates rising passenger flow, priority should be given to increasing operational capacity; conversely, declining flow suggests appropriate reduction in train frequency. The coefficient λ was determined via grid search to be 0.7, achieving optimal balance between recent fluctuations and historical trends. (2) Capacity adaptive weight: reflects real-time load status at stations

$$\beta_i^t=\beta\cdot\frac{Q_i^t}{C_i}\,(\beta=1.2)$$

C_i Designed for station capacity (e.g., People's Square Station:C_i = 12,000 people). Whenβ_i^t , it indicates the station is operating at overload capacity, triggering scheduling intervention. The β value was determined through sensitivity analysis to be 1.2, amplifying the weight of high-load stations.

The improved PageRank formula is:

$$PR_i^t = (1-d)\cdot\left(\alpha_i^t + \beta_i^t\right) + d\Sigma_{j\in B_i}\frac{PR_j^t w_{ji}}{\max\left(1, L\left(j\right)\right)}$$

Where w_{ji} represents the passenger flow weight from station j to i ($w_{ji} = Q_{ji} / \sum_k Q_{j_k}$), and L(j) denotes the number of outbound edges from station j. This model dynamically responds to passenger flow changes—for example, at 8:00 AM, People's Square Station exhibitsα_i^t = 0.35 (35% increase in passenger flow) andβ_i^t = 1.15 (115% load factor), collectively elevating its PR value to 0.026 (3.8 times the network average of 0.008), accurately identifying critical dispatch nodes.

2.3 Multi-objective dynamic programming model

Based on station importance assessments, a multi-objective optimization model was established to balance stakeholder interests:

Decision variables:

-f_k^t : Departure interval (seconds) for line k during time period t

-m_k^t : Number of train formations (number of cars)

Objective Function:

Minimize total passenger waiting time:

$$\min \Sigma_{t,i} PR_i^t \cdot \left(0.5 f_k^t\right)$$

Where `$0.5f_k^t$` represents the average waiting time under uniform distribution, and `$PR_i^t$` serves as a weighting factor. This ensures waiting times at high-importance stations carry greater weight in the objective function, embodying the principle of "priority optimization for critical stations."(2) Load factor equalization:

$$\min \sigma\left(\rho_k^t\right): \sqrt{\frac{1}{N}\Sigma_{k=1}^{N}\left(\rho_k^t - \overline{\rho}\right)^2}$$

ρ_k^t Denotes the load factor of route k during time period t (actual passenger load divided by seating capacity), where$\sigma\left(\rho_k^t\right)$ is the network-wide average load factor. This metric reflects the fairness of capacity allocation: a smaller variance indicates more balanced utilization of transport capacity across routes, preventing overcrowding on some routes while others run empty.(3) Minimum Operating Costs:

$$\min \Sigma_k\left(C_1 m_k^t + C_2 / f_k^t\right)$$

C_1 = 0.85 yuan/vehicle-kilometer represents energy consumption costs (proportional to distance traveled and number of carriages), C_2 = 120 yuan/hour represents driver labor costs (proportional to departure frequency). This objective directly relates to corporate economic benefits.

Constraints:

- Departure interval constraint: 120 ≤f_k^t ≤ 600 (seconds), ensuring passenger wait times remain acceptable while preventing train entry/exit conflicts.

- Capacity Coverage Requirement:$\Sigma_k m_k^t \times 328 \geq 1.2 \times \max_i\left(PR_i^t \times Q_i^t\right)$, where 328 represents the single-car capacity of Type A trains and 1.2 is the safety factor, guaranteeing capacity meets at least 120% of peak demand.

- Timetable Feasibility: $\Delta T_{ij} \geq T_{ij} + D_j + 0.3 \times PR_j^t$, whereT_{ij} represents interval travel time andD_j denotes station dwell time. The$0.3 \times PR_j^t$ term adds safety margin for high-importance stations (each 0.01 increase in PR value adds 0.3 seconds), mitigating delay propagation caused by passenger surges.

The Pareto optimal solution set is solved using a Rapid Non-Dominated Sorting Genetic Algorithm (NSGA-II) with an elite strategy. This algorithm simulates biological evolution to find optimal solutions for multi-objective optimization problems. Parameters are set as follows: population size 100, 500 generations, simulated binary crossover (probability 0.9), and polynomial mutation (probability 0.1). Non-dominated solutions are retained at each iteration to ensure diversity within the solution set.

3 Experiments and results analysis

3.1 Experimental setup

The experiment utilized Shanghai Metro's March 2023 weekday dataset (days 1–20 as training set, days 21–31 as test set), covering typical post-Spring Festival commuter flow characteristics. Comparison scenarios included:

- Approach A: Current fixed timetable (January 2024 version)
- Scheme B: Optimization scheme based on genetic algorithms (Shi Jungang et al., 2023)

Evaluation metrics are defined as follows:

(1) Average waiting time ($\overline{w}$): The average time (in seconds) passengers wait from passing through the fare gates to boarding the train. This metric directly reflects passenger experience, with lower values indicating higher service quality. Calculation formula:

$$\overline{w} = \frac{1}{N_P}\Sigma_{p=1}^{N_P}\left(t_P^{boarding} - t_P^{entry}\right)$$

Where N_P is the total number of passengers,t_P^{entry} and$t_P^{boarding}$ represent the entry and boarding timestamps for passenger p, respectively.

(2) Passenger Delay RateR_d : Percentage of time when the instantaneous platform occupancy exceeds 80% of the design capacity. This metric measures congestion risk. Calculation formula:

$$R_d = \frac{1}{T}\Sigma_{t=1}^{T} I\left(N_t > 0.8C_i\right) \times 100\%$$

T is the total number of time periods, (·) is the indicator function, is the number of people on the platform at time t, and is the station's design capacity.

(3) Train Empty-Running Rateη : Percentage of trains during off-peak hours (10:00–16:00) with passenger load below 20% of capacity (%). Reflects operational capacity wastage:

$$\eta = \left(N_{empty} / N_{total}\right) \times 100\%$$

(4) Operating Cost C: Includes energy consumption and labor costs (RMB 10,000/day). Energy cost = mileage × RMB 0.85/vehicle-kilometer; labor cost = driver hours × RMB 120/hour.

Load Factor Varianceσ_ρ^2 : Variance in load factors across routes during the same period, measuring capacity allocation balance:

$$\sigma_\rho^2 = \frac{1}{K}\Sigma_{k=1}^{K}\left(\rho_\kappa - \overline{\rho}\right)^2$$

Where K is the number of routes,ρ_κ is the average load factor for route k, and$\overline{\rho}$ is the network-wide average. A lower value indicates fairer resource allocation.

3.2 Results and discussion

Table 1 Performance Comparison During Peak Hours (7:30–9:00)

Indicator	Scenario A	Scenario B	Proposed Model	Improvement Rate
$\overline{w}$ (seconds)	158.3	136.7	123.5	22.4% ↓
R_d (%)	18.7	14.2	11.8	36.9% ↓
η (%)	32.5	26.8	22.2	31.7% ↓
$C (10,000 yuan/day)	86.3	79.1	74.6	13.6% ↓
σ_ρ^2	0.183	0.142	0.089	51.1% ↓

Analysis of results is as follows:

(1) Dynamic Responsiveness: The enhanced PageRank algorithm accurately identifies peak stations. At 8:00 AM, the PR value at People's Square Station rose to 0.026 (3.8 times the network average of 0.008), triggering the model to dynamically shorten the interval to 125 seconds (compared to 150 seconds in Plan A). This reduced waiting time at the station from 170 seconds to 132 seconds, lowering the delay rate by 41%.

(2) Multi-objective coordination: The Pareto solution set generated by this model expanded the distribution range in the three-dimensional objective space (waiting time-cost-variance of load factor) by 37% compared to Plan B. For instance, under a 5% cost increase constraint, waiting time could be further reduced by 8%, demonstrating the model's ability to effectively balance conflicting objectives.(3) Field Validation Accuracy: The predicted morning rush hour waiting time at People's Square Station was 123.5 seconds, with an actual measured mean of 127.8 seconds (error: 3.2 seconds, MAPE=2.1%), meeting the IEEE TRANSACTIONS journal requirement of less than 5%. The model demonstrated less than 3% error in cross-section flow reproducibility tests (e.g., Shenjiang → Xujiahui section: predicted 8,463 people/hour vs. measured 8,520 people/hour), validating the reliability of the data fusion framework.

4 Conclusions

This paper addresses the spatio-temporal imbalance of passenger flow in the Shanghai Metro by proposing a dynamic train scheduling method that integrates multi-source data with an improved PageRank algorithm. The study first constructs a three-tier data fusion framework combining "cross-section flow-timetable-transfer station observations." Using the EM algorithm, it back-calculates the dynamic origin-destination (OD) matrix for the entire network from passenger flow data at 10 key cross-sections, effectively overcoming the bottleneck of obtaining complete passenger flow data. In algorithm design, the traditional PageRank model was innovatively enhanced by introducing time-sensitive factors ($\lambda = 0.7$) and capacity-adaptive weights ($\beta = 1.2$), enabling precise dynamic assessment of station importance. A multi-objective optimization model was established, integrating station importance parameters into train scheduling decisions, with the NSGA-II algorithm used to solve for the Pareto optimal solution set.

The study's limitations include the exclusion of sudden passenger flow scenarios such as large-scale events and extreme weather conditions. Future work will focus on the following directions: 1) Integrating reinforcement learning algorithms to build a real-time feedback-based emergency dispatch module; 2) Extending the model to multimodal scenarios to achieve coordinated scheduling between subways, buses, and railways; 3) Investigating parameter transfer strategies for cities of different scales. The research findings provide a new paradigm for intelligent scheduling in urban rail transit, and the methodology can be extended to dynamic optimization problems in complex network systems such as buses and railways.

References

[1] Liu, Z.Y., Wang, J., Li, X.X. (2021) Optimization of station layout in urban rail transit based on genetic algorithm. Journal of the China Railway Society, 43(5): 1–9.

[2] Dong, S.W. (2023) Short-term passenger flow prediction based on BP neural network. Urban Rapid Rail Transit, 36(2): 45–52.

[3] Shi, J.G., Zhou, X., Xu, R., et al. (2023) Integrated optimization of train timetabling and stop planning with passenger flow control under time-dependent demand. Transportation Research Part C: Emerging Technologies, 146: 103989.

[4] Chen, S.Q. (2013) Real-time influence evaluation model in social networks based on human dynamics. Chinese Journal of Computers, 36(8): 1687–1696.

[5] Agryzkov, T., Oliver, J.L., Tortosa, L., et al. (2012) An algorithm for ranking urban nodes using the PageRank vector. Applied Geography, 34: 14–26.