Transactions on Environment, Energy and Earth Sciences
ISSN: 2960-1924, eISSN: 2960-2300 | Volume 7
GTREE 2026

Abnormal Passenger Flow Forecasting for Rail Transit Based on Multi-Source Big Data Fusion

Kaiwen Ling *, Jinbing Ha and Qiyue Chen

School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China

* Corresponding Author Email: 124107010934@njust.edu.cn

Abstract. To address the problem of forecasting abnormal passenger flow at urban rail transit stations caused by large-scale events, holidays, and weather changes, this paper proposes a multi-source big data abnormal spatio-temporal prediction framework, MBDASPF. Based on historical entry and exit passenger flow, the model incorporates external disturbance factors such as day of week, holidays, weather, official event website information, and Weibo super topic data, and adopts a "branch-transform-fuse" strategy to extract the historical passenger flow correlation and the abnormal disturbance correlation, respectively. The historical branch employs GRU to capture recent proximity, daily periodicity, and weekly tendency, while the disturbance branch employs LSTM to model the influence of external events. Finally, through two-stage fusion and an attention-based output layer, multi-step forecasting for the next 10, 20, and 30 min are realized. Experiments on data from the Olympic Sports Center Station of Nanjing Rail Transit Line 10 show that MBDASPF outperforms ARIMA, SVR, BPNN, and GRU overall in terms of MAE, RMSE, and MAPE, and can more effectively characterize abnormal passenger flow fluctuations during large-scale events.

Keywords: Abnormal passenger flow; passenger flow forecasting; multi-source data fusion; deep learning; urban rail transit.

1. Introduction

Urban rail transit is characterized by large capacity, high punctuality, and limited susceptibility to external interference, and has become an essential component of the public transport system in large cities. Accurate short-term passenger flow forecasting assists operators in formulating flow-control measures, adding extra train services, and designing passenger flow guidance plans, and is of direct significance for safeguarding operational safety during large-scale events. In recent years, short-term passenger flow forecasting research for urban rail transit has gradually shifted from single historical series modeling to deep spatio-temporal modeling [1, 2]. However, during holidays, heavy rainfall, large-scale concerts, or sporting events, station passenger flow exhibits fluctuations that deviate markedly from regular periodic patterns. In particular, large-scale events induce abnormal passenger flows that are short-lived, exhibit pronounced peaks, and lack strong periodicity [3], and relying solely on historical passenger flow series is often unable to capture such variations in a timely manner.

Existing deep learning research mainly proceeds along three lines. The first uses structures such as LSTM, GRU, and CNN to extract the temporal dependence and local fluctuation features of passenger flow [2] [4]; the second employs graph convolution, spatio-temporal attention, or multi-graph structures to characterize the spatial correlations among stations [5, 6, 7, 8]; the third introduces external or multi-source data such as search indices, social media, OD matrices, geographical locations, and real-time entry/exit volumes to enhance the explanatory ability of models for unconventional scenarios [9, 10]. These studies demonstrate that deep models can significantly improve short-term forecasting accuracy; however, in large-scale event scenarios, without explicit modeling of event disturbances and public attention, models may still underestimate sudden peaks or respond with a lag. Although existing research has already paid attention to multi-source data and spatio-temporal deep models, most methods are more suitable for regular periodic passenger flow

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

and are insufficient in characterizing short-term peaks induced by large-scale events, precursor signals from social media, and asymmetric disturbances before and after events.

To address the above issues, this paper focuses on station-level abnormal passenger flow forecasting induced by large-scale events and proposes *the multi-source big data abnormal spatio-temporal prediction framework* (MBDASPF). In line with recent research on event passenger flow forecasting and multi-source fusion [3, 4], this paper further integrates official event website information, weather, holidays, and Weibo super topic indicators into the station-level 10 min passenger flow series. The main contributions are as follows. First, a multi-source feature system oriented to station-level abnormal passenger flow is constructed. Second, a historical passenger flow branch and an abnormal disturbance branch are designed to model regular temporal patterns and the influence of external events, respectively. Third, multi-step forecasting experiments are conducted on real station data, with comparisons against ARIMA, SVR, BPNN, and GRU.

2. Data and Feature Construction

This paper takes the Olympic Sports Center Station of Nanjing Rail Transit Line 10 as the research object. Around the station are the Nanjing Olympic Sports Center and the Jiangsu Grand Theatre, which host concerts, sports events, dramas, and exhibition performances throughout the year, making it a suitable scenario for abnormal passenger flow forecasting under large-scale events. The experiments use the station's entry and exit card-swipe data from January 1 to December 31, 2023, aggregated at a 10 min granularity to form passenger flow series; meanwhile, official event website information, government holiday calendars, weather records, and Weibo super topic indicators are collected as external disturbance data.

Table 1. Multi-source data and feature descriptions.

Data type	Main fields	Role
Passenger flow data	Entry volume, exit volume, timestamp	Provides historical passenger flow series and the forecasting target
Temporal data	Day of week, holiday flag	Characterizes commuting periodicity and holiday differences
Weather data	Air temperature, precipitation, wind speed	Characterizes the impact of weather on travel demand
Event data	Event venue, start/end time, number of events	Characterizes the spatio-temporal disturbance of large-scale events
Weibo super topic data	Original post volume of location-based super topics, discussion volume and original post volume of event super topics	Reflects public attention and event popularity

Historical passenger flow features are constructed from three perspectives: recent proximity, daily periodicity, and weekly tendency. Let y_t denote the passenger flow volume at time t. The recent feature consists of the c time steps preceding time t; the daily periodic feature takes the n time steps around the same time period of the previous day; and the weekly tendency feature takes the m time steps around the same time period of the previous week. The three types of features reflect short-term continuous variation, intra-day repetitive patterns, and weekly periodic regularities, respectively.

$$N_t = [y_{t-1}, y_{t-2}, \ldots, y_{t-c}] \tag{1}$$

$$D_t = [y_{t-d}, y_{t-2d}, \ldots, y_{t-nd}], \quad W_t = [y_{t-w}, y_{t-2w}, \ldots, y_{t-mw}] \tag{2}$$

Abnormal disturbance features include day of week, holidays, weather, large-scale events, and social media popularity. Holidays are encoded with 0/1; weather uses air temperature, precipitation, and

wind speed; event features use the number of events at the venue as well as the numbers of time steps from the current time to the start and end of the event; the Weibo data use the original post volume of location-based super topics, the discussion volume of event super topics, and the original post volume of event super topics, which have relatively strong correlations with passenger flow.

Table 2 . Example correlations between social media indicators and passenger flow(Lag1–Lag4 denote the correlation of passenger flow at lags of [specific time]; * $p < 0.05$, ** $p < 0.01$)

Data source	Metric	Lag1	Lag2	Lag3	Lag4
Location-based super topic	Discussion volume trend	0.482*	0.326*	0.296*	0.326*
Location-based super topic	Original post volume trend	0.621**	0.571*	0.431*	0.472*
Event super topic	Discussion volume trend	0.502*	0.442*	0.452*	0.382*
Event super topic	Original post volume trend	0.696**	0.626*	0.598*	0.480*

3. The MBDASPF Model

MBDASPF adopts the overall paradigm of "branch-transform-fuse". The model first partitions the inputs into a historical passenger flow correlation branch and an abnormal disturbance correlation branch; it then performs feature transformation using networks suited to their respective data structures; finally, referring to the multi-source feature fusion idea [4] [8], it carries out two-stage fusion to obtain a comprehensive representation and outputs passenger flow forecasting for multiple future time steps.

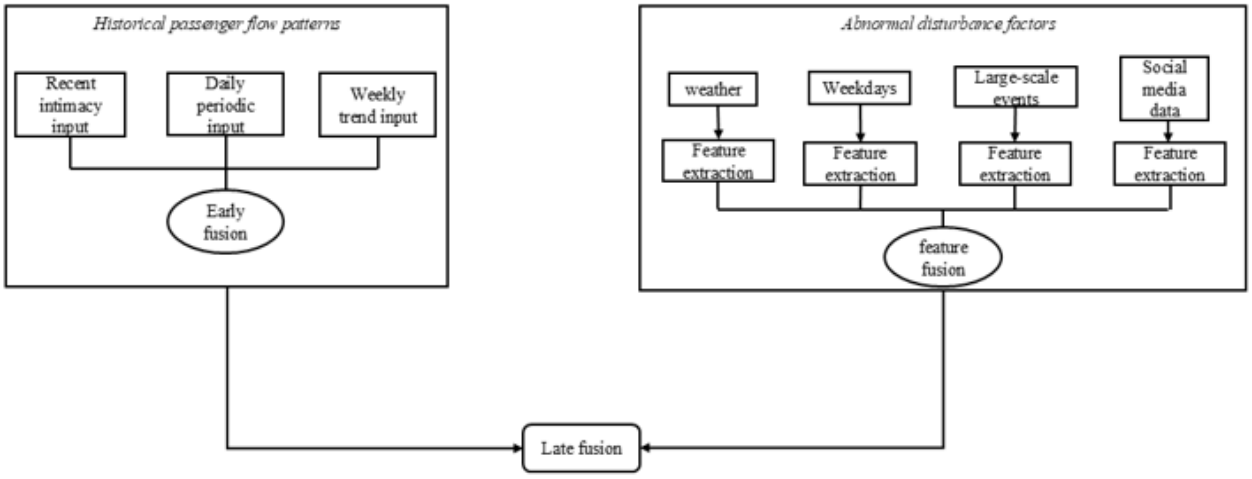


Fig 1. The "branch-transform-fuse" strategy

The historical passenger flow correlation branch uses three GRU sub-networks to process recent proximity, daily periodicity, and weekly tendency, respectively. The outputs of the three sub-networks are first combined by early fusion to learn the importance of different historical patterns for the current forecasting task. Let H_N, H_D, H_D denote the hidden representations of the three types of historical features, respectively. The early fusion result H can be expressed as:

$$H = W_N H_N + W_D H_D + W_W H_W \tag{3}$$

The abnormal disturbance correlation branch uses LSTM to perform temporal modeling of day of week, holidays, weather, event information, and Weibo indicators. This branch is designed to learn the sustained influence of external events before, during, and after large-scale events. Subsequently, the historical branch output H and the disturbance branch output E are concatenated along the feature dimension to obtain the comprehensive feature Z.

$$Z = H \oplus E \tag{4}$$

The output layer adopts an attention mechanism and a multi-task learning structure, treating the forecasting for the next 10, 20, and 30 min as three related tasks. The attention weights are computed by a fully connected layer and Softmax, and are used to highlight key information across different time steps and different feature dimensions. This design is consistent with the recent idea of introducing attention mechanisms into passenger flow forecasting to capture key temporal segments. The loss function adopts mean square error, and the parameters are trained with the Adam optimizer.

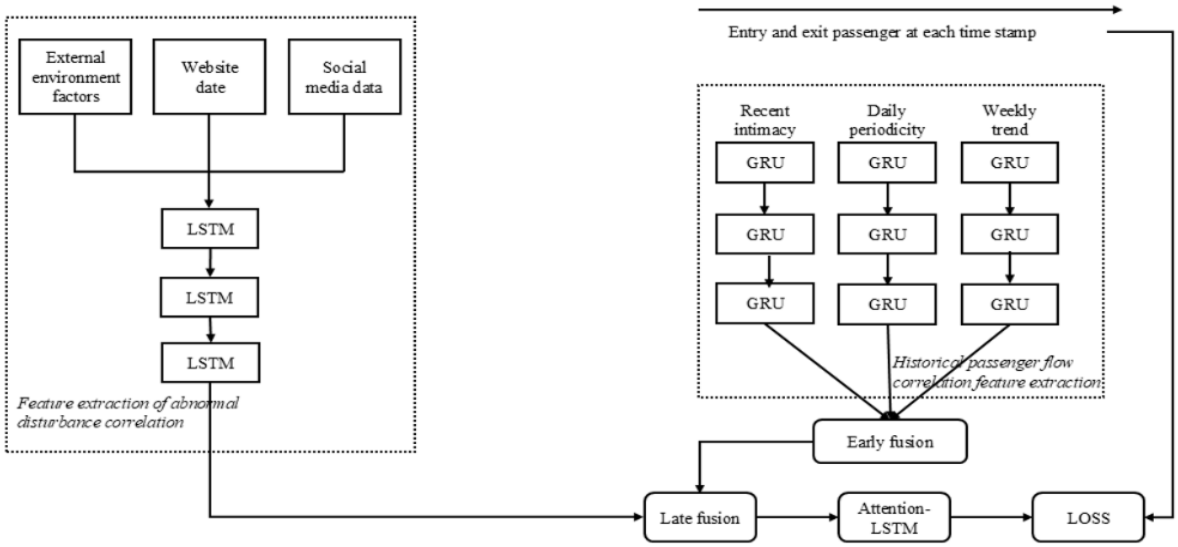


Fig 2. Overall architecture of the MBDASPF model

4. Experimental Setup

The experimental time granularity is 10 min, and the forecasting targets are the entry and exit passenger flows for the next three time steps. All continuous features are mapped to the 0-to-1 interval by min-max normalization, and the model outputs are then de-normalized to obtain actual passenger flow values. MAE, RMSE, and MAPE are selected as evaluation metrics, where smaller values indicate lower forecasting errors.

$$MAE = \frac{1}{K} \Sigma |y_i - \hat{y}_i|, \quad RMSE = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)\text{^}2}{K}} \tag{5}$$

$$MAPE = \frac{1}{K} \Sigma \left| \frac{(y_i - \hat{y}_i)}{y_i} \right| \times 100\% \tag{6}$$

To verify the effectiveness of the model, ARIMA, SVR, BPNN, and GRU are selected as baseline models. Among them, ARIMA represents statistical time series methods, SVR and BPNN represent traditional machine learning and shallow neural network methods, and GRU represents commonly used deep temporal forecasting models. All comparison experiments adopt the same time granularity and forecasting horizon.

5. Results and Analysis

The trained MBDASPF is used to forecast the entry and exit passenger flow during the "A-Mei ASMR MAX World Tour Concert—Nanjing Station" held at Nanjing Olympic Sports Center Stadium on May 11, 2024, and the forecasted curves can well follow the passenger flow peak variations brought by event dispersal. Compared with models that rely only on historical passenger

flow, after introducing event and social media features, the model responds to abnormal peaks more promptly.

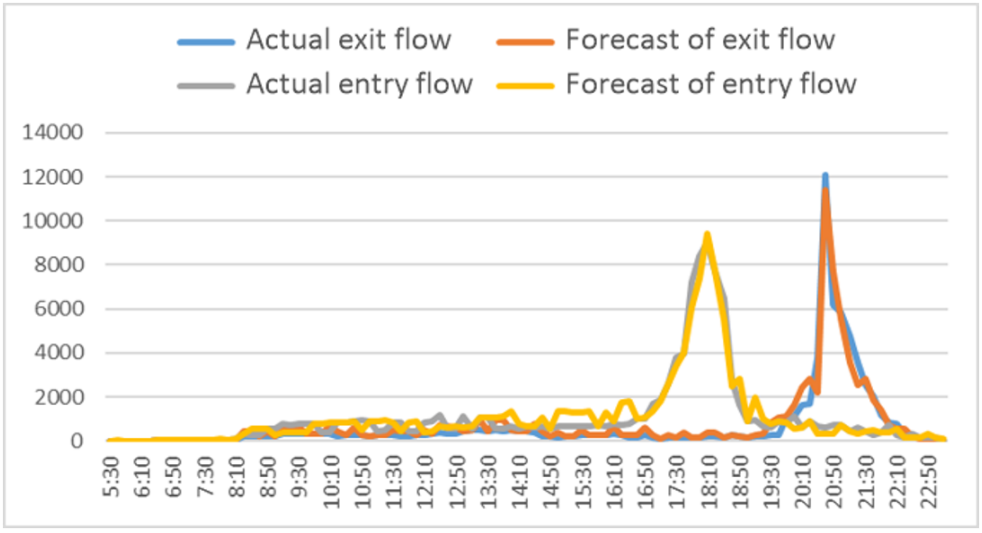


Fig 3. Forecasting results of the MBDASPF model

Table 3 . Performance of different models in entry passenger flow forecasting

Forecasting horizon	Metric	ARIMA	SVR	BPNN	GRU	MBDASPF
10min	MAE	27.89	30.29	26.34	24.58	16.78
10min	RMSE	56.87	47.67	41.76	39.89	27.54
10min	MAPE/%	19.02	23.01	14.21	13.65	12.23
20min	MAE	28.38	39.21	27.89	25.99	17.22
20min	RMSE	56.83	58.79	45.67	42.67	29.76
20min	MAPE/%	20.32	23.98	15.65	14.21	12.98
30min	MAE	28.65	47.67	27.65	26.45	17.88
30min	RMSE	58.98	67.45	52.54	47.58	34.58
30min	MAPE/%	22.43	25.01	16.32	15.67	13.56

Table 4 . Performance of different models in exit passenger flow forecasting

Forecasting horizon	Metric	ARIMA	SVR	BPNN	GRU	MBDASPF
10min	MAE	25.68	29.59	28.84	22.48	17.48
10min	RMSE	58.70	47.94	44.26	37.95	28.94
10min	MAPE/%	18.62	17.61	15.56	14.54	13.43
20min	MAE	26.42	37.71	28.97	26.79	18.24
20min	RMSE	55.34	53.93	47.37	42.73	30.65
20min	MAPE/%	19.72	20.28	17.52	16.18	14.28
30min	MAE	29.58	49.67	30.54	29.65	19.48
30min	RMSE	62.48	72.72	53.64	45.76	36.28
30min	MAPE/%	20.39	24.81	18.23	19.79	16.36

It can be seen from Table 3 and Table 4 that the MAE, RMSE, and MAPE of MBDASPF are generally lower than those of the four baseline models across different forecasting horizons. ARIMA and SVR have weak adaptability to abnormal peaks; BPNN lacks an explicit temporal structure; and GRU is able to capture historical passenger flow variations but does not fully exploit external event information. MBDASPF simultaneously considers regular historical patterns and abnormal disturbance factors, and therefore achieves more stable forecasting performance in large-scale event scenarios.

The ablation experiments show that using the historical branch alone yields a certain level of forecasting capability, but the error increases significantly when facing event passenger flow peaks; using the disturbance branch alone lacks the basic passenger flow regularities, resulting in insufficient forecasting stability. Fusing the two types of features significantly reduces the error, indicating that the historical passenger flow correlation and the abnormal disturbance correlation have complementary effects.

6. Conclusion

This paper addresses the forecasting of abnormal passenger flow at rail transit stations in large-scale event scenarios and proposes the MBDASPF model, which fuses historical passenger flow, weather, holidays, official event website information, and Weibo super topic data. The model learns regular temporal patterns and the influence of external events through the historical passenger flow branch and the abnormal disturbance branch, respectively, and realizes multi-step forecasting via two-stage fusion. Experimental results on data from the Olympic Sports Center Station in Nanjing show that, compared with ARIMA, SVR, BPNN, and GRU, the proposed method achieves lower forecasting errors and effectively characterizes the entry and exit passenger flow peaks during events.

There remain certain limitations in this study: the experiments are mainly based on single-station data and have not yet fully considered the propagation effects of transfer stations and adjacent stations within the rail transit network; information such as large-scale event types, audience scale, and ticket sales can be further refined. Future research may combine multi-station graph structures with richer event semantic features to improve the generalization ability of the model across different lines and cities.

References

[1] CAI Changjun. Study on urban rail transit entry and exit short-term passenger flow forecasting model[J]. Urban Mass Transit, 2021, 24(9): 14-19+24.

[2] Zhang J, Chen F, Cui Z, Guo Y, Zhu Y, Zhu Y. Deep learning architecture for short-term passenger flow forecasting in urban rail transit[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(11): 7004-7014.

[3] Xue G, Liu S, Ren L, Ma Y, Gong D. Forecasting the subway passenger flow under event occurrences with multivariate disturbances[J]. Expert Systems with Applications, 2022, 188: 116057.

[4] Li H, Jin K, Sun S, Jia X, Li Y. Metro passenger flow forecasting through multi-source time-series fusion: An ensemble deep learning approach[J]. Applied Soft Computing, 2022, 120: 108644.

[5] Li S, Wang Q W, Chen Y R, et al. Forecasting of short-time passenger flow on multi-station urban rail based on SAE-ConvLSTM deep learning model[J]. Application Research of Computers, 2022, 39(7): 2025-2031.

[6] Lu W, Zhang Y, Li P, Wang T. Mul-DesLSTM: An integrative multi-time granularity deep learning forecasting method for urban rail transit short-term passenger flow[J]. Engineering Applications of Artificial Intelligence, 2023, 125: 106741.

[7] Li X, Wu J, He D, Teng X, Ren C. Learning spatial-temporal dynamics for short-term passenger flow forecasting in urban rail transit[J]. Transportation Research Record, 2023, 2677(5): 1330-1348.

[8] Cui H, Si B, Wang J, Zhao B, Pan W. Short-term origin-destination flow forecasting for urban rail network: A deep learning method based on multi-source big data[J]. Complex & Intelligent Systems, 2024, 10: 4675-4696.

[9] Ma C, Zhang B, Li S, Lu Y. Urban rail transit passenger flow forecasting with ResCNN-GRU based on self-attention mechanism[J]. Physica A: Statistical Mechanics and its Applications, 2024, 638: 129619.

[10] LUO Weiliang, HOU Wei, SHI Gan, et al. Research on urban rail transit passenger flow forecasting method during dispersal period of large-scale events[J]. Urban Mass Transit, 2026, 29(3): 166-173.