DOI: 10.54254/2753-8818/26/20241060

Short-time subway passenger flow forecast based on the ARIMA-LSTM

Ziyue Chang[1], **Yiquan Ding**[2, 4], **Xinran Guo**[3]

[1]International Education College, Hubei University of Economics, Wuhan, 430205, China
[2]School of Computer Engineering, Nanjing Institute of Technology Nanjing, 211167, China
[3]School of Information and Intelligent Engineering, Zhejiang Wanli University, Ningbo, 315100, China

[4]x00202210309@njit.edu.cn

Abstract. In recent years, with the rapid development of China's economy and the continuous increase in population, the urbanization process has intensified. This travel issue caused by peak traffic congestion affects a significant portion of urban populations. In order to conduct a thorough analysis, this article primarily focuses on the urban rail transit industry's subway system. In this study, three models - LSTM, ARIMA and ARIMA-LSTM were employed to analyze and predict short-term inbound flow data for the Hangzhou subway in January 2019. While the LSTM model used for forecasting, it was found to be ineffective in predicting data with extreme peaks. The ARIMA model is subsequently employed for data prediction, revealing its inadequacy in accurately forecasting unstable and non-patterned data. To overcome this limitation, a combined ARIMA-LSTM model is proposed to mitigate the shortcomings of individual models and achieve superior performance. The implementation of the ARIMA-LSTM model effectively mitigates subway congestion issues, thereby facilitating informed decision-making by subway operators. Moreover, it has a certain degree of robustness, even if the other data are not regular enough. The model can be applied in the real subway passenger flow prediction, which is conducive to the choice of people traveling to work and the management decisions of subway operators.

Keywords: Subway, passenger flow, ARIMA, LSTM, ARIMA-LSTM.

1. Introduction

In recent years, the rapid growth of China's economy and its expanding population have posed a significant challenge in terms of peak congestion during travel, thereby driving the remarkable development of the urban rail transit industry. In the Chinese mainland, a total of 55 cities have implemented urban rail transit systems (including subways, urban high-speed rails, cross-monorail trams, rail rubber wheel systems, electronic guide rubber wheel systems, light rails and maglev traffic), operating 308 lines with a combined length of 10,000 kilometers and reaching 10,287.45 kilometers in total length - an increase of 11.91% year-on-year. In terms of passenger flow volume in 2022, the use of urban rail transit accounted for 45.82% of all public transport passenger transportation - an increase of

2.45 percentage points from the previous year[1]. The gradual increase in the level of urbanization in our country has led to a continuous rise in the passenger traffic volume of rail transit, especially during the peak hours of working days. The contradiction, such as overcrowding of passengers, is becoming more prominent.

Short-term passenger flow prediction primarily serves real-time information and dynamic scheduling, making it one of the main means for urban rail transit to alleviate congestion and improve service levels. Therefore, based on the results of short-term passenger flow prediction, the subway operators can implement scientific and feasible real-time operation mode, adjust passenger transportation organization strategy in time, steer passenger behavior, and implement corresponding measures such as efficiently managing high-volume passenger flows.

The research on short-term passenger flow prediction in urban rail transit, both domestically and internationally, primarily falls into three categories: mathematical statistical models, intelligent algorithm models, and combination prediction models. Traditional time series models cannot consider random events and unexpected events. Prediction methods based on mathematical statistics, such as ARIMA (Autoregressive Integrated Moving Average Model) and SARIMA (Seasonal Autoregressive Integrated Moving Average) models. The ARIMA model treats the traffic flow as a non-stationary stochastic sequence, which has a better prediction effect on the data with strong smoothing but has poor prediction effect on the data with stronger stochasticity and the existence of extreme peaks [2]. The emergence and development of machine learning address the problem of traditional prediction models based on statistical theory being unable to handle large-scale complex data, enabling more accurate and efficient data processing. However, when using prediction models based on intelligent algorithms to handle a large amount of passenger flow data, overfitting or underfitting issues may arise, thereby affecting the accuracy of the prediction models. Roos et al. developed a short-term passenger flow prediction model for the Paris Metro based on dynamic Bayesian networks [3]. The model incorporates the Expectation Maximization algorithm (EM) to estimate the model parameters via maximum likelihood, which allows for prediction even in the presence of missing data. The overall performance of the model is good, but it may degrade when confronted with severe data missingness. Yushan Wang developed a hybrid model integrating high-voltage length and short-term memory (LSTM) neural network while optimizing the relevant parameters,-time prediction of passenger flow at rail transit stations [4]. Yanjun Huang utilized LSTM to establish a real-time prediction model for each time component, enabling the prediction of passenger flow at each station. Additionally, they employed the k-neighbor algorithm to construct a passenger flow structure prediction model for OD (original destination) passengers [5]. Yang Liu extended the urban rail transit prediction model based on CNN-LSTM by introducing an attention mechanism that captures the weight allocation of spatial and temporal features in probability [6]. Huangkun Liu primarily employed LSTM for medium and long-term predictions, while utilizing a ridge regression auxiliary model for short-term predictions when dealing with limited data amounts. This approach demonstrated excellent fitting effects on urban rail transit passenger flow data [7]. A single LSTM model cannot fully capture the spatiotemporal correlation of related event sequences, while fusion models outperform single models in terms of performance and applicability. Lu Saiqun et al. combined the ARIMA model and LSTM model, using a sliding window-based dynamic weighting method to predict highway traffic flow. Compared to single models and equal-weight combination methods, this model has better performance and greater generality [8].

This study focuses on the data collected from Hangzhou Metro in January 2019 as the research subject. The short-term passenger flow prediction of the subway is analyzed through the three regressive Integrated Moving Average Model, LSTM, and ARIMA-LSTM mixed model. The predictive performance of each model is examined in detail, followed by an analysis of the limitations of individual models such as ARIMA or LSTM.

2. Methods

2.1. Data source

This study employs objective and reliable data from the Hangzhou Subway Passenger Data for January 2019 in its analysis.

Table 1. Name and explanation of variables

Full Name	Data Type
TIME_PERIOD	FLOAT
DAY	INT
WEATHER	INT
NUM	INT

The 5 variables utilized in the study are listed in Table 1 along with their full names, data types, and explanations. The official data source for the information is the Hangzhou Subway. As a result, it is more reliable and accurate. The 819,134 pieces of data, which represent each entry or exit from a subway station in a single day, are more than enough to assist the investigation of rail transmit flow and rail transit flow forecast.

2.2. Introduction to the method

2.2.1. LSTM. The long and short-term memory (LSTM) neural network is a unique type of recurrent neural network or RNN. In some recurrent neural networks, LSTM is frequently employed to address long-term reliance issues since it is effective at transferring and representing data across lengthy time series. The gradient disappearance problem (small weight/bias gradient, causing neural network parameter adjustment rate to drop sharply) and the gradient explosion problem (large weight/bias gradient, causing neural network parameter adjustment amplitude to be too high, overcorrect) are two neural network problems that the LSTM can solve for the RNN with ease.

Triple gates—forgotten, input, and output gates—are present in LSTM. This model's key strength is its ability to quickly and effectively collect spatial and temporal data on traffic flow and to forecast various traffic indicators including congestion and traffic flow. The model can be enhanced at the same time, and as time changes, the prediction accuracy will rise steadily. Figure 1. depicts the LSTM neural network structure.

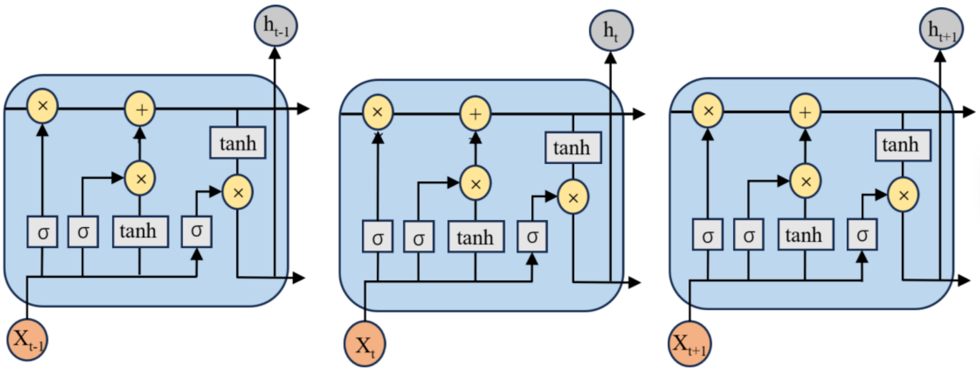

Figure 1. LSTM network

2.2.2. ARIMA. Time-series data is frequently predicted and modeled using ARIMA models. The core of ARIMA is to create a model that uses autoregressive, mobile average, and differential transformation of time series data to describe the data features and then use that model to forecast future data changes. The ARIMA model can fit the data with a limited number of parameters and does a good job of handling the features of many time-series data. It's quite easy to use the ARIMA model. However, the timing data must be stable when using the ARIMA model to predict them. It is impossible to capture the law with unreliable data.

In conclusion, employing LSTM to forecast transient passenger flow can yield good experimental outcomes. With some room for improvement, additional algorithms can be utilized to help concurrently.

2.3. Assessment method

2.3.1. MSE. The mean squared error(MSE) is the mean reflecting the sum of squares of the difference between the predicted data and the actual value [9].

$$\mathrm{MSE} = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} (\mathrm{yi} - \mathrm{f}(\mathrm{x_i}))^2 \, \mathrm{y} \tag{1}$$

2.3.2. RMSE. Root mean square error (RMSE) is a commonly used measure of the difference between model predicted and actual observed values, which is used to assess how well the model fits to a given data.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{Y_i} - \mathrm{f}(\mathrm{x_i}))^2} \tag{2}$$

2.3.3. R^2. The coefficient of determination (R^2)means how much proportion of the predicted value explains the variance of the target variable, and it measures how well the predicted value fits
to the true value [10].

$$R^2 = 1 - \frac{\sum_{\mathrm{i}} (\hat{\mathrm{y}}_{\mathrm{i}} - y_i)^2}{\sum_{\mathrm{i}} (\bar{y}_i - y_i)^2} \tag{3}$$

3. Results and discussion

3.1. Descriptive analysis

These four images, which are histograms of the passenger flow of the Hangzhou subway at different times of the day at half-hourly intervals, are displayed in Figure 2 and use 1,000 people as the unit.

Figure 2. illustrates that the subway passenger traffic on Nos. 8, 9, 10, and 11 peaks every day between 8:00 a.m. and approximately 18:00 and 19:00, coinciding with the daily work and commute schedules of individuals.

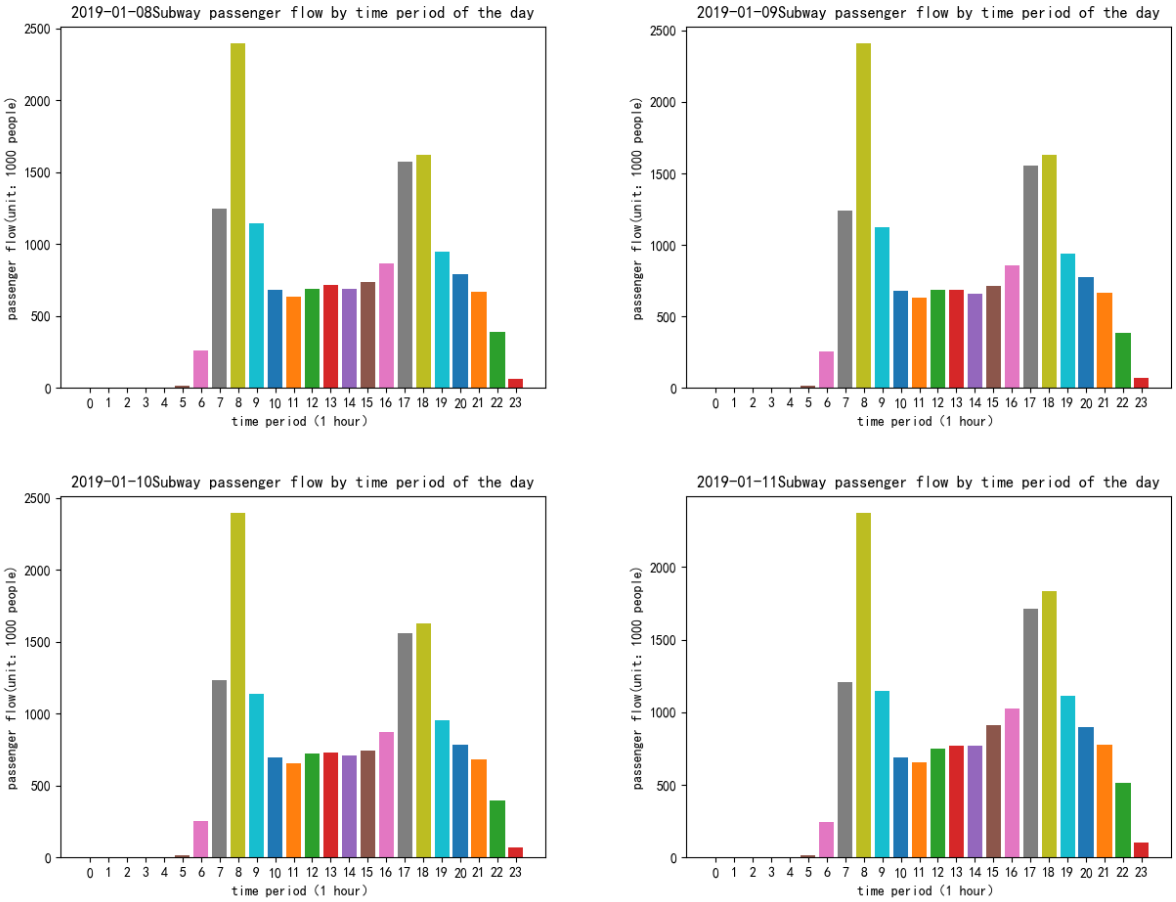

Figure 2. Subway traffic

The data has some smoothness, seasonality, and other characteristics thanks to pre-processing and analysis of the passenger flow parameter num, which is shown in Figure 3. and Figure 4. for the ACF and PACF analysis of num. These characteristics fit the requirements for using the ARIMA model.

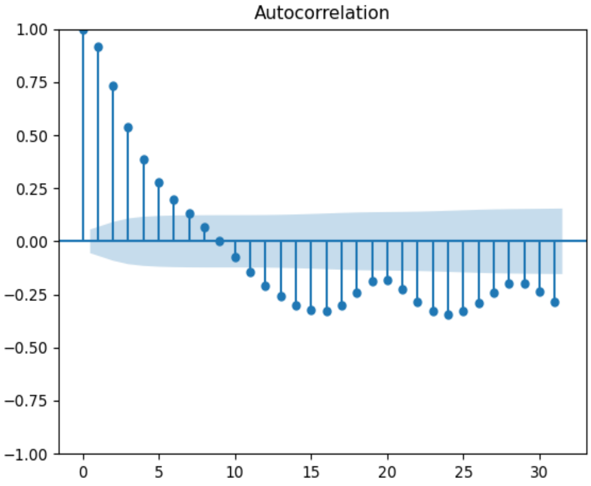

Figure 3. ACF

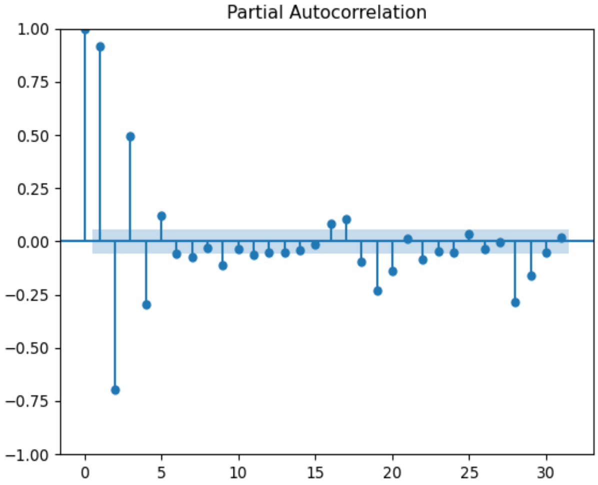

Figure 4. PACF

3.2. *Inferential analysis*

Based on Figure 5, it can be inferred that the LSTM model performs better in the vicinity of more concentrated mean values; nevertheless, at some peaks and valleys, the predicted values deviate significantly from the observed data. After testing, it is determined that the LSTM model needs more development to produce better outcomes, as it is not optimal for prediction alone. Furthermore, the passenger flow data exhibits a greater degree of periodicity, leading to the decision to attempt to include the ARIMA model to increase prediction accuracy.

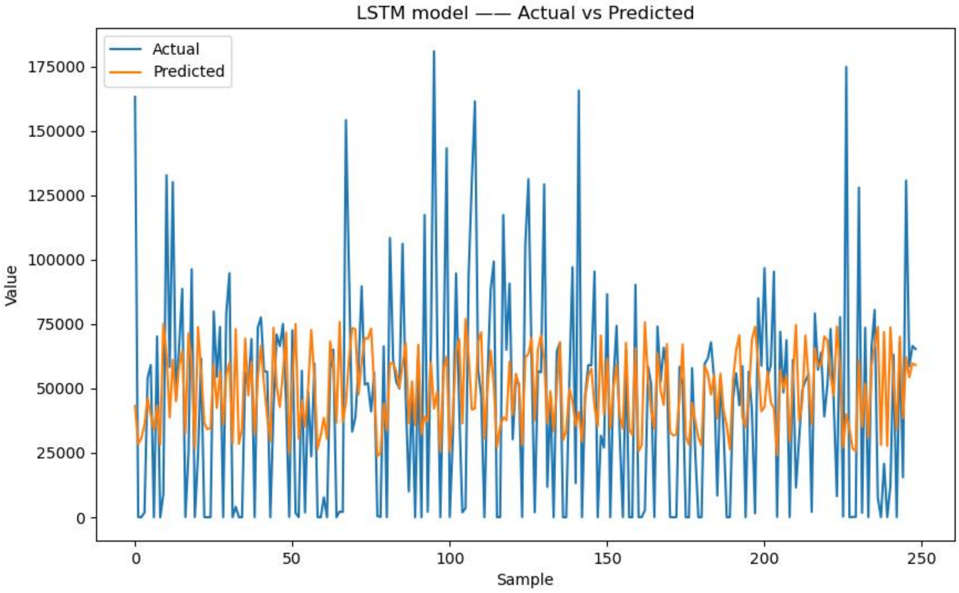

Figure 5. LSTM model result

Using the ARIMA-LSTM hybrid model can effectively avoid the drawbacks of using a single model and achieve a better prediction effect. As seen in Figure 6, the prediction effect of using the ARIMA model alone is very good for data. However, for some data that are unstable or lack an obvious pattern, the prediction effect of ARAMA is poor.

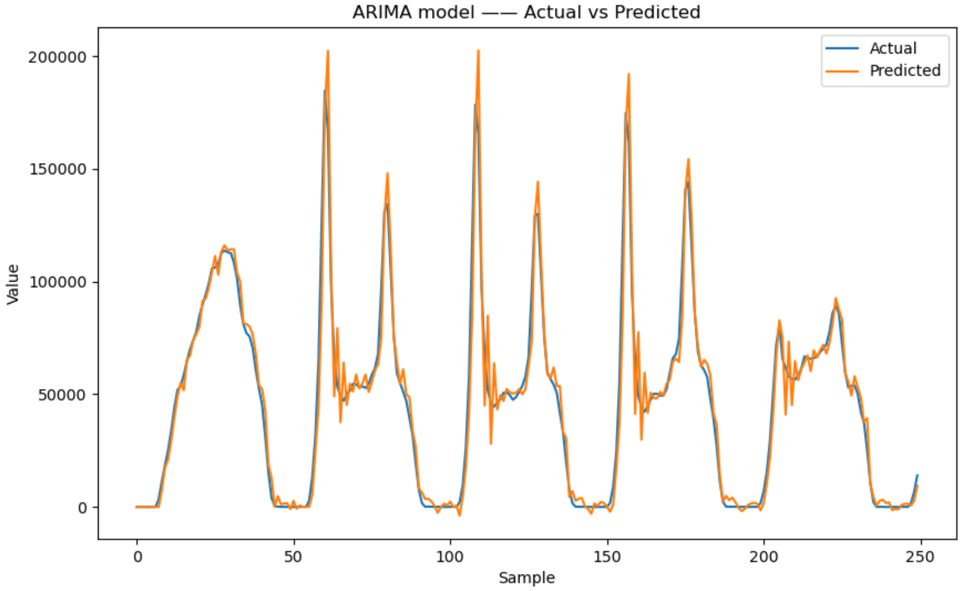

Figure 6. ARIMA model result

Figure 7 illustrates that the ARIMA-LSTM hybrid model fits well and that the model's loss function value lowers after training.

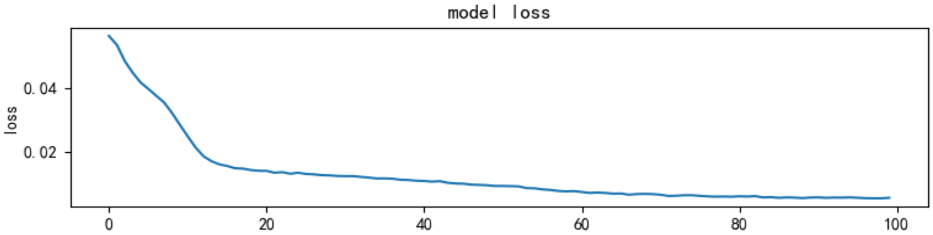

Figure 7. ARIMA-LSTM model loss

The final ARIMA-LSTM hybrid model prediction is shown in Figure 8. In the neighborhood of each value, it is excellent, and the mistake falls inside the permitted range.

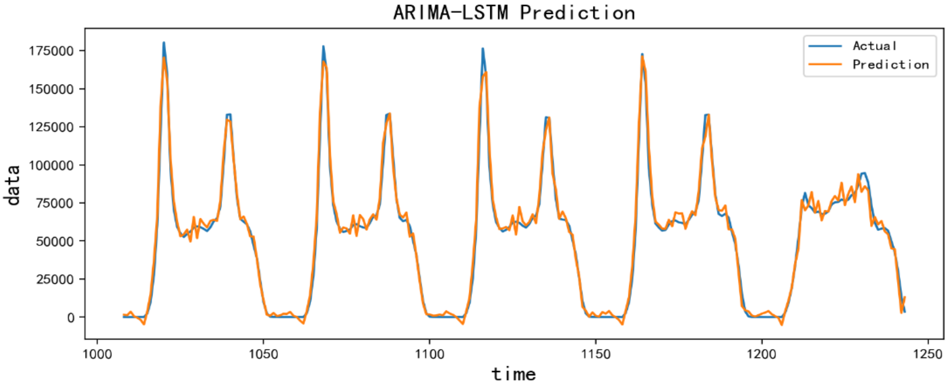

Figure 8. ARIMA-LSTM model result

3.3. Evaluation analysis

Table 2. The predictive performance of each model

Model name	Assessment criteria		
	RMSE	MAE	R^2
ARIMA	0.0718	0.0469	0.9763
LSTM	0.1996	0.1505	0.1556
ARIMA-LSTM	0.0446	0.0267	0.9492

Table 2 illustrates how bad the LSTM single model is, with all three evaluation criteria ranking lower than those of the other models. So you cannot use this model. Despite this, the ARIMA single model's R^2 assessment criteria would be superior to the ARIMA-LSTM hybrid model's R^2 criterion. The RMSE and MAE criteria of the ARIMA-LSTM hybrid model are significantly better than those of the other models, and these criteria further support the ARIMA-LSTM hybrid model's superiority over the other models for this Hangzhou subway prediction. Overall, the criteria of the hybrid model are better than the other models.

4. Conclusion

The passenger flow data of the Hangzhou subway is predicted in this research using the trained ARIMA-LSTM hybrid model, with good results. It demonstrates that it has some degree of stability by comparing the outcomes with those anticipated by a single model. Along with handling the prediction of subway passenger flow with good application in certain scenarios.

By using the ARIMA-LSTM hybrid model's prediction results, it is possible to efficiently lessen or even eliminate the problem of peak-hour subway traffic congestion, giving commuters a positive travel experience and cutting down on the needless wasting of public transportation resources.

The ARIMA-LSTM hybrid model trained in this paper has a small number of parameters. It requires high-quality input data. To further enhance generalization skills, accuracy, and other factors. The model requires the addition of other variables, such as the station's surrounding infrastructure and its precise placement across the entire metro network. The ARIMA-LSTM hybrid model can be further improved using all these measures to get better outcomes.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] A total of 2,022,1,085.17 kilometers of new urban rail transit operation lines will be added China Urban Rail Transit Association

[2] Du J S 2022 Research and system implementation of short-time passenger flow prediction for urban rail transit based on deep learning Spiedigitallibrary p 53

[3] Roos J, Gavin G and Bonnevey S 2017 A dynamic Bayesian network approach to forecast shortterm urban rail passenger flows with incomplete data. Transportation Research Procedia 26 pp 53-61.

[4] Wang Y S 2022 Urban rail transit station Wuhan: Huazhong University of Science and Technology

[5] Yan J H 2021 Urban rail transit network Beijing: Beijing Jiaotong University

[6] Liu Y, Mu C and Zhou P 2022 The Short-Term Passenger Flow Prediction Method of Urban Rail Transit Based on CNN-LSTM with Attention Mechanism International Conference on Mobility, Sensing and Networking (MSN) pp 909-914

[7] Liu H, Ren J and Wang Z 2023 Short-term passenger flow prediction strategy of urban rail transit based on ridge regression and LSTM IEEE International Conference on Control, Electronics and Computer Technology (ICCECT) pp530-534

[8] Lu S 2021 A combined method for short-term traffic flow prediction based on recurrent neural network. Alexandria Engineering Journal 60(1) pp 87-94.

[9] Han M 2021 E-Bayesian estimations of parameter and its evaluation standard: E-MSE (expected mean square error) under different loss functions Communications in Statistics-Simulation and Computation 50(7) p 1971.

[10] Liu Y X 2023 Research on futures price prediction and quantification strategy based on deep learning algorithm Shandong Normal University