Research on Short-term Passenger Flow Prediction Method of Urban Railway Based on Machine Learning

Yang Chen*

Chang'an University, Xi'an, Shaanxi, 710021, China
2020906132@chd.edu.cn

***Abstract*—**With the development of urban rail transit, the improvement of communication transmission technology, and big data processing level, the passenger flow data of urban rail transit continues to grow and can be effectively collected and stored, providing many basic data resources for analysis and passenger prediction research. With the development of artificial intelligence, short-term passenger flow prediction based on machine learning has received widespread attention. For this reason, based on the passenger flow data of Seoul Metro, this paper has carried out experiments and analysis on the existing models RNN-LSTM and ConvLSTM with better performance, providing theoretical support for short-term passenger flow prediction. The method in this paper solves the problem that a single traditional prediction model based on statistical theory is difficult to deal with large-scale complex data. It can complete the data processing task more accurately and efficiently, and has a good application prospect.

Keywords—Short-term, Machine learning, Urban Railway, Passenger Flow Prediction

1. Introduction

As an important part of urban rail transit system management and control, short-term urban rail passenger flow prediction provides decision-making basis for urban rail transit real-time operation and passenger flow organization, and has very important practical significance for improving the level of traffic management services and control ability. Short-term rail passenger flow forecasting is the key to a high match between rail capacity resources and transportation demand by predicting daily passenger flow demand in the next few days [1]. Currently, linear theoretical models, artificial intelligence models, and combined forecasting models are widely used for time series-based deep learning, especially the last two. Therefore, this paper will apply several types of general deep learning models to predict the short-term subway passenger flow and verify the superiority of the model.

On the night of October 29, 2022, a huge number of people went to Itaewon, South Korea to participate in Halloween activities, which led to a crowd surge. The accident occurred near Exit 1 of Itaewon Station on Seoul Metro Line 6. If the short-term passenger flow data in the subway station could be predicted in advance, the relevant departments could evacuate the crowd early, thereby avoiding the occurrence of crowd surge accidents.

In the traffic volume forecasting problem, traditional methods mainly include Zhang [2] using an ARIMA model to forecast short-term passenger flows in urban rail systems. Jiao [3] used an improved Kalman filter model to forecast the flow of iron passengers on a trip. Zhu [4] studied the chaos characteristics of a time series of railroad traffic volumes and used a chaos theory model to predict volume; Xu [5] used a gray forecasting model to predict train passenger flows. In addition, in recent years, artificial intelligence models have been proposed that contain machine learning models, Recurrent Neural Networks (RNN), and deep learning models. In order to support the dynamic adjustment of rail

passenger operation plans, deep learning technology has advanced rapidly in accurate traffic prediction by analyzing historical data from a single urban rail station. For example, single models such as GRU [6], LSTM [7,8], ConvLSTM, and several hybrid models such as CNN-LSTM [9,10], RNN-LSTM. Many studies have shown that deep learning frameworks are better than single models. This paper will apply the different hybrid models to forecast the short-term passenger flow situation and verify the superiority of these models.

This paper conducts research based on the Seoul Metro data in 2021. Firstly, it introduces the data set structure, divides the original data into four sub-datasets GB, GG, XB, and XG, and selects the data of Seoul Station in January and February for visualization and analyzes data characteristics qualitatively from the explanatory, overall dimensions. Secondly, this paper uses RNN-LSTM and ConvLSTM two hybrid models to predict short-term passenger flow, visualizes the prediction results, and quantitatively analyzes the prediction results. Thirdly, the RMSE index is applied to quantitatively evaluate the two models to compare these models and analyze the reasons. Finally, prospect the application of the superior model.

2. Methods

2.1 Data Source and Description

With almost 310 subway stations on 10 lines, Seoul ranks among the largest metropolitan cities in the world. During workdays and weekends, respectively, each line carries about 700,000 and 300,000 passengers. The passenger flow dataset for the Seoul Metro system from https://data.seoul.go.kr was used in this study. This dataset tracks the number of people using 224 Seoul Metro stations on lines 1 to 8 from January to December 2021 in 1-hour intervals. This dataset also contains data regarding how many people board and get off at each subway station throughout various times of the day in 2021. Rapid transit and commuter rail line Seoul Metro line 1 serves the majority of the Seoul Capital Area, which is the subject of this study. The core underground part of this rail line is the oldest subway component in the Seoul Metropolitan Subway system. In addition, because both Seoul Station and Itaewon Station are located in business districts with huge passenger flows, thus the research on the short-term passenger flow forecasting of Seoul Station will help to set a typical example for the passenger flow forecasting and control work of other similar metro stations. Consequently, this paper selects the station lines on Seoul Metro line 1 for analysis, specifically the data on workdays and weekends of Seoul Station in 2021.

Aiming to study the change patterns of passenger flow in a single station, this paper uses the passenger flow of Seoul Station on Seoul Metro Line 1 from January to February 2021 as the research data for 59 days of passengers entering and exiting the station. Since there were 2242 pieces of data throughout this time period, the data are primarily separated into the following four datasets: the number of passengers boarding on workdays (GB), the number of passengers getting off on workdays (GG), the number of passengers boarding on weekends (XB), and the number of passengers getting off on weekends (XG).

2.2 Data Distribution

2.2.1 Explanatory data analysis

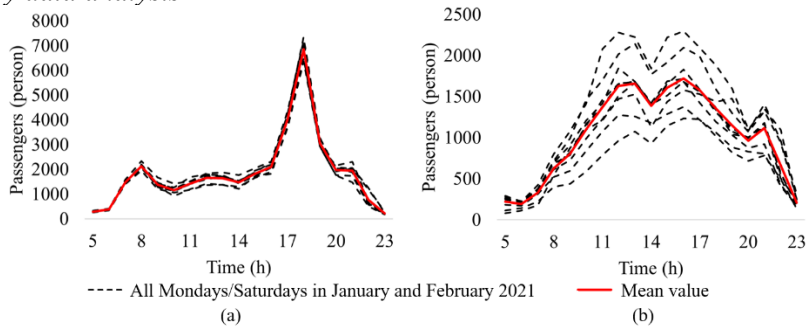

Figure 1. Diagram of the average daily number of passengers boarding the metro

Figure 1 shows the average number of daily passengers boarding the metro on all Mondays and Saturdays in January and February 2021 at Seoul Station on Seoul Metro line 1, with the average values represented by the red line in each figure. Plot (a) in Figure 1 depicts the boarding passenger flow pattern on workdays (GB). It is clear that there are two peaks in each plot, with the first peak occurring between 7 and 10 am and the second peak occurring between 5 and 8 pm. As both peaks are present during rush hours and this railroad is densely crowded, this is consistent with natural intuition. Plot (b) in Figure 1 depicts the boarding passenger flow pattern on weekdays (XB). It demonstrates that there is no peak commute hour on weekends and that passenger flow is often higher during the daytime since fewer people utilize the subway to commute to work on weekends. The weekend pattern is more moderate than the workday pattern.

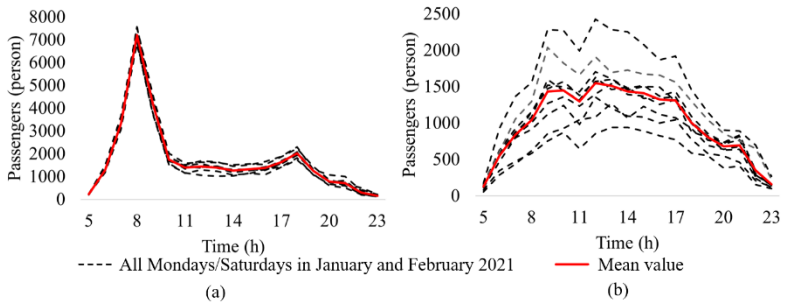

Figure 2. Diagram of the average daily number of passengers getting off the metro

Figure 2 shows the average number of daily passengers getting off the metro on all Mondays and Saturdays in January and February 2021 at Seoul Station on Seoul Metro line 1. Compared with the changing trend of boarding passenger flow shown in Figure 1, although the passenger flow of getting off during the workdays (GG) in Plot (a) also presents two peak periods, during the morning peak period between 7 and 10 am, the passenger flow getting off the metro is much higher than that of boarding the metro. Also, during the evening peak period between 5 and 8 pm, the passenger flow boarding the metro is much higher than that of getting off the metro. Through the comparison between the above two different features, it can be confirmed that Seoul Station is a downtown subway station that mainly serves business districts rather than residential areas. Plot (b) shows the changing trend of getting off passenger flow during the weekends (XG), which is basically similar to the change of XB passenger flow. There is no peak in the morning and evening during the commuting time, and the passenger flow gradually increases in the morning and decreases in the afternoon and evening.

2.2.2 Overall data analysis

This research analyzes and forecasts passenger flows on the four categories of datasets GB, GG, XB, and XG separately due to the significant variances in passenger flow trends between workdays and weekends at this station as shown in Figure 1. The overall change trends of the selected data are depicted in Figures 2 and 3, which mostly represent the passenger flow patterns per weekend or workday through Seoul Station from January to February 2021. Considering data trends, the trend of processed data is shown in Table 2. The Seoul Station serves as an interchange station for three metro lines, including Line 1, Line 4, and Airport Railroad as the hub of the Seoul metro network. Additionally, this station is a subway stop in a typical business area with many commercial buildings, residential areas, educational institutions, etc., thus this subway station has a steady and substantial monthly passenger flow.

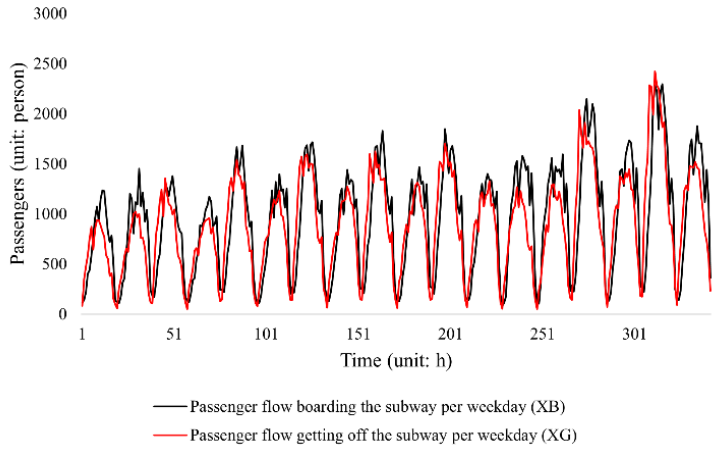

Figure 3. Diagram of Passenger flow per weekend of Seoul Station from January to February 2021

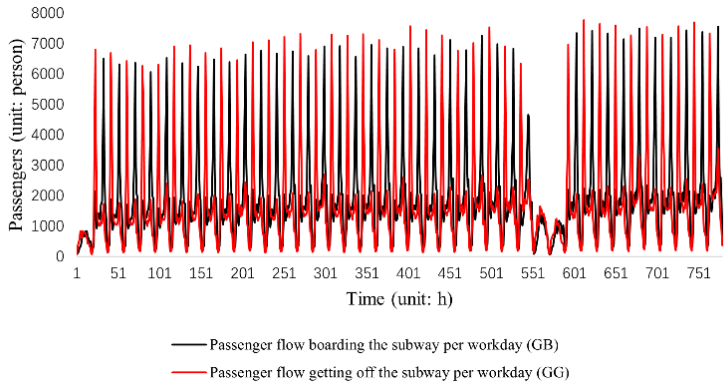

Figure 4. Diagram of Passenger flow per workday of Seoul Station from January to February 2021

Table 1. Example of Processed Data

Date	State	Before 06:00	06:00-07:00	07:00-08:00	…	After 23:00
2021-1-1	Boarding	86	111	157	…	101
2021-1-1	Get off	85	355	438	…	77

Table 2. The Trend of Processed Data

Index	Dataset categories			
	GB	GG	XB	XG
Mean value	1756.890	1632.937	855.047	792.679
Variance	2074840.048	2314184.627	216668.2	170879.8
Standard deviation	1440.431	1521.244	465.476	413.376

2.2.3 Research Methods

This study forecasts the flow of arriving and departing passengers at a time granularity of 1 hour at Seoul Station, where three lines intersect and transfer. Table 1 displays the processed sample data. The problem studied in this paper can be expressed as at a given time t, use the sequence of passenger flows at the previous time of the station to predict the corresponding sequence of passenger flows from the station

in the next continuous time period. The detailed processing steps of the data are divided into three parts: data preprocessing, model training, and model prediction.

2.2.4 Data preprocessing

The time series must be thoroughly preprocessed taking into account both the features of the time series and numerous predictive model attributes before training the model. In this paper, standardization is carried out before model training and prediction so that the data can adjust to the input requirements of different prediction models, specifically the RNN-LSTM model and ConvLSTM model, in order to facilitate neural network convergence and enhance prediction accuracy. A list of the specific data processing steps is as followed.

Step 1: Assume the original passenger flow time series is $P = \{p_1, p_2, \dots, p_n\}, |P| = n$ and p_i indicates the passenger flow of the station on the i-th day.

Step 2: The previous training data series P_{T_1} before a given time t is $P_{T_1} = \{p_{t-T-1}, p_{t-T}, \dots, p_t\}$, the prediction data series P_{T_2} is $P_{T_2} = \{p_{t+1}, p_{t+2}, \dots, p_{t+T_2}\}$.

Step 3: Since the prediction model will use a default activation function when generating the state and output at time t, for example, the tanh is the default activation function as operation 1. To ensure the consistency of the input and output value data of the prediction model in dimension, the input data must be standardized so that it is between [-1, 1]. Normalize the passenger flow sequence data to the [-1, 1] interval using MinMaxScaler in python software, the transformed dataset is $X = \{x_1, x_2, \dots, x_n\}$.

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{1}$$

$$x_t = \frac{p_t - \min(P)}{\max(P) - \min(P)} \tag{2}$$

Step 4: The data set X is reconstructed by sliding window segmentation. Assuming that the input time step of the prediction model is L_1, and the output prediction time step is L_2, then the length of the sliding window is $L_1 + L_2$, and with a unit slides each time, a total of $n - L_1 - L_2 + 1$ sequences of length $L_1 + L_2$ can be generated. The first L_1 data in each sequence constructs an input sequence X_j, and the last L_2 data constructs an output sequence Y_j.

$$X_j = \{X_j, X_{j+1}, \dots, X_{j+L_1-1}\}, x_j \in X \; and \; |X_j| = L_1 \tag{3}$$

$$Y_j = \{X_{j+L_1}, X_{j+L_1+1}, \dots, X_{j+L_1+L_2-1}\}, x_{j+L_1} \in X \; and \; |Y_j| = L_2 \tag{4}$$

The processed data is marked as $D = \{X^{input}, Y^{output}\}$, and the data set is divided into the training set $D^{train} = \{X^{train}, Y^{train}\}$ and test set $D^{test} = \{X^{test}, Y^{test}\}$ according to a certain proportion ε, since this article only analyzes the data of a single station, the amount of data is small. This paper utilizes 90% of the training set and 10% of the test set aiming to increase the number of training times and enhance the training effect.

2.2.5 Model training

After processing the data according to the above model input parameters, it is required to identify the model configuration parameters before executing model training. The following takes the LSTM model as an example. The input parameters of this model include activation function, learning rate, loss function, data input batch, etc. After the model parameters are determined, each time sequence input model with a specified input batch size is selected from D^{train}, and the LSTM model is continuously trained.

2.2.6 Model prediction and evaluation

The root mean square error (RMSE) is the evaluation metric used in this paper following model training. Following is the operation of RMSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i^{test} - p_i^{forecast})} \tag{5}$$

In operation 5, p_i^{test} represents the actual passenger flow data at time i, $p_i^{forecast}$ represents the forecasting passenger flow data at the time i.

3. Result and Discussion

3.1 RNN-LSTM Prediction Model

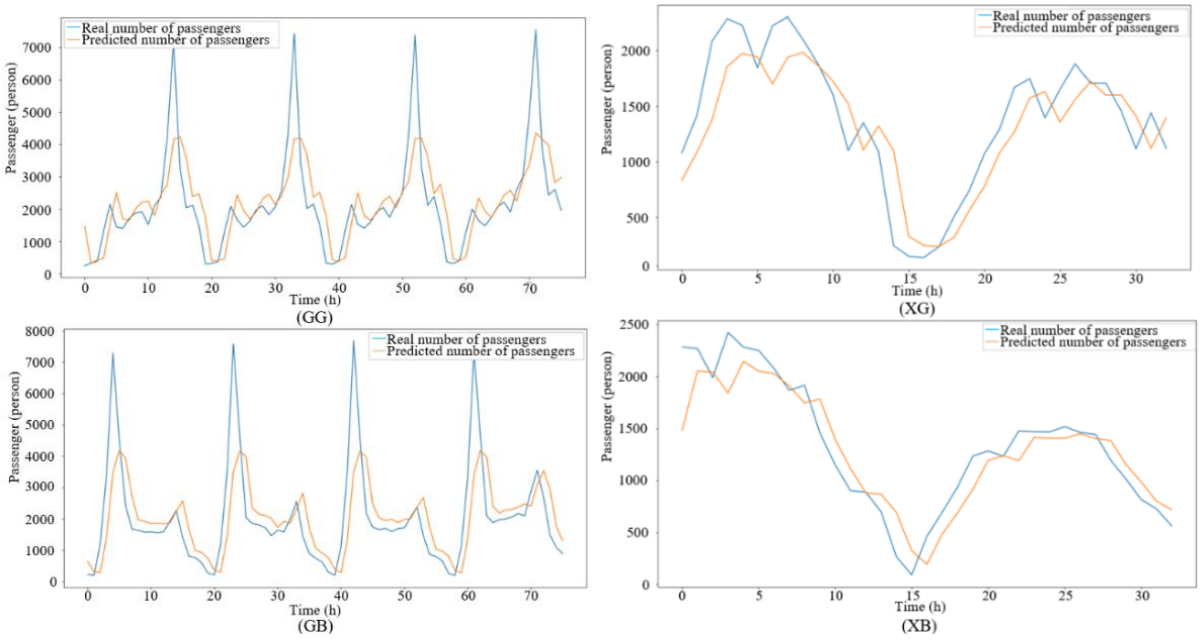

Figure 5. Comparisons of the predicted and actual daily number of passengers in four categories dataset

The results of the RNN-LSTM model passenger flow prediction for the four datasets above are shown in Figure 4. The figure's yellow line represents the predicted passenger flow, whereas the blue line depicts the actual passenger flow. The anticipated curve and the real curve essentially exhibit equivalent trends, and the RNN-LSTM model has strong predictability for short-term passenger flow prediction. This is demonstrated by comparing the actual passenger flow and the predicted passenger flow curve. However, there are still two problems worth noting: the first is that the distance between the forecast curve and the actual curve is relatively large, indicating that the model has a long forecast period when predicting time series passenger flow data with a large time granularity, for example, in this paper the selected time granularity is 1h data; the second is the problem of insufficient prediction of extreme values in the regional extremum of the predicted curve and actual curve. At the same time, the problem of insufficient prediction accuracy of the model is also reflected in the relatively large value of RMSE of model evaluation, as shown in Table 3. For the above problems, this paper speculates that the lack of prediction accuracy of the RNN-LSTM model is due to its own relatively simple model architecture, indicating that the model is not suitable for predicting time series data with large time granularity.

3.2 ConvLSTM Prediction Model

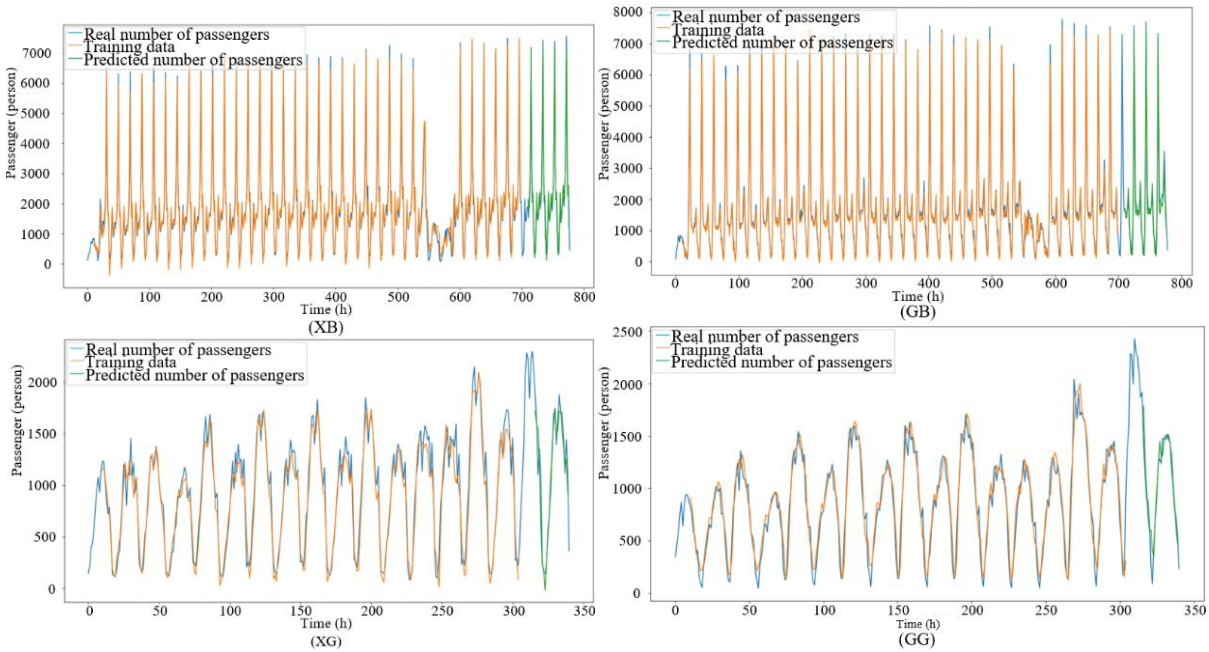

Figure 6. Comparisons of the predicted and actual daily number of passengers in four categories dataset

The results of the ConvLSTM model passenger flow prediction for the four datasets above are shown in Figure 5. The blue line in the figure shows the actual passenger flow, the yellow line shows the training data which covers 90% of all data, and the green line shows the estimated passenger flow which covers 10% of all data. As can be seen from the above four figures, the ConvLSTM model fully combines the advantages of the convolutional network and the long-short memory network. The trend of the forecast curve is consistent with the actual curve, and both of them fundamentally fit together with excellent prediction accuracy, regardless of whether the weekend passenger flow data are large-volume or relatively small. Compared with the RNN-LSTM model, the ConvLSTM model has a stronger predictive ability for time series data with longer time granularity, particularly the prediction of regional extreme values, which is more accurate and has a shorter prediction period. In Table 3, the smaller RMSE value of ConvLSTM also indicates its higher prediction accuracy.

Table 3. Model RMSE Index

Data source	Model	State	Proportion (Train data: test data)	RMSE
Workdays from Jan. to Feb. 2021	RNN-LSTM	Boarding	9:1	1050.03
		Get off	9:1	1227.31
	ConvLSTM	Boarding	9:1	233.15 (train score) 295.39 (test score)
		Get off	9:1	186.48 (train score) 270.76 (test score)
Weekends from Jan. to Feb. 2021	RNN-LSTM	Boarding	9:1	318.50
		Get off	9:1	252.46
	ConvLSTM	Boarding	9:1	129.71 (train score) 176.38 (test score)

	Get off	9:1	126.35 (train score) 144.05 (test score)

4. Conclusion

Due to the difference in the structure of the prediction model, this paper uses RNN-LSTM and ConvLSTM, which are currently commonly used, respectively, through the four subway passenger flow datasets --GB, GG, XB, and XG--of Seoul Station of Seoul Metro line 1 with 1 hour time granularities. The prediction results demonstrate that the ConvLSTM model, which combines the architectural benefits of convolutional models and the LSTM model, can more effectively solve the time series passenger flow prediction problem with longer time granularity when compared to the RNN-LSTM model with a simpler structure. To sum up, this paper believes that the ConvLSTM model has great application value in the short-term prediction of subway passenger flow. This research on the short-term passenger flow forecasting of Seoul Station--a typical business district metro station--will help to set a typical example for the passenger flow forecasting and control work of other similar metro stations to avoid the recurrence of the subway crowd surge accident.

References

[1] B S He, Y J Zhu, Chen L F, Wen K Y. A Spatial-temporal Graph Neural Network for Prediction of Short-term Passenger Flow at Urban railway Station [J]. Journal of the China railway society, 2022,44(09) :pp. 1-8.

[2] B M Zhang. Short-term forecasting of passenger flow in Shanghai-Nanjing inter-city railway [J]. Chinese Railways, 2014(9): pp. 29-33.

[3] P P Jiao, Li, R M Sun, T Hou, Z H Ibrahim, Amir. Three Revised Kalman Filtering Models for Short-Term Rail Transit Passenger Flow Prediction. Mathematical Problems in Engineering, 2016, pp. 1–10.

[4] Z H Zhu, Z S Weng. Railway passenger and freight volume forecasting based on chaos theory [J]. Journal of the China Railway Society, 2011, 33(6): 1-7.

[5] K Xu, X Y Bao, Wang Q C. Railway passenger volume forecasting based on GA-GM (1, N, α) power model[J]. Railway Standard Design, 2018, 62(1): 6-10.

[6] J Li, Q Y Peng, Wen C. Short term passenger flow prediction of high speed railway based on LSTM deep neural network[J]. Systems Engineering – Theory & Practice, 2021, 41(10): 2669-2682.

[7] R Fu, Z Zhang, L Li. Using LSTM and GRU neural network methods for traffic flow prediction [C]. 2016 31st youth academic annual conference of Chinese association of automation: Wuhan, China, 2016.

[8] J Li, Q Y Peng, C Wen. Short term passenger flow prediction of high speed railway based on LSTM deep neural network[J]. Systems Engineering – Theory & Practice, 2021, 41(10): pp. 2669-2682.

[9] Y Wang, Z F Wang, H Y Wang,, et al. Prediction of passenger flow based on CNN-LSTM hybrid model[C]//2019 12th international symposium on computational intelligence and design (ISCID). IEEE, 2019, 2: pp. 132-135..

[10] M Wang, H Lv, Y Zhai. Prediction of Short-term Passenger Flow of Urban Rail Transit based on Data Decomposition[C]//2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE, 2022: pp. 1088-1095..