Short-term Subway Passenger Flow Prediction based on LSTM and ARIMA Model

Yunhao Ma *

Department of Electronic Information and Computer Engineering, The Engineering & Technical College of Chengdu University of Technology, Leshan, 614000, China

* Corresponding Author Email: 1811030403@stu.hrbust.edu.cn

Abstract. With the acceleration of urbanization, urban traffic problems have become increasingly severe. To alleviate urban traffic congestion, the development of public transportation has been greatly emphasized. As a convenient and rapid means of urban travel, the metro has become the preferred choice for citizens, leading to significant passenger flow pressure. Therefore, it is crucial to forecast passenger flow in advance. This paper utilizes passenger flow data from Chengdu Metro Line 1 from May to September 2019 to conduct a study. LSTM and ARIMA models were constructed to simulate the actual passenger flow data. The prediction performance of both models shows that the trends simulated by these models fit the original data well, verifying their applicability in passenger flow forecasting. Furthermore, it was found that the LSTM model outperformed the ARIMA model among the constructed models in this paper. Accurate metro passenger flow forecasting can help relevant departments manage the metro system more effectively, thereby improving passenger travel efficiency and safety.

Keywords: LSTM; ARIMA; passenger flow prediction.

1. Introduction

With the development of modern cities and the growth of urban populations, the demand for transportation has gradually expanded. The urban road network has become increasingly extensive and complex, leading to traffic congestion and accidents, which significantly increase the difficulty and pressure of traffic management. The development and construction of underground rail transit have greatly alleviated the pressure on surface transportation [1]. However, as rail transit becomes more developed and convenient, it has become the primary choice for resulting in a significant increase in the pressure on rail transit systems [2]. This surge in demand presents considerable challenges for the operation and management of rail transit systems. Therefore, it is necessary to forecast passenger flow in advance.

The accurate prediction of passenger flow has always been one of the key work contents of urban rail transit management department, especially in the context of the rapid growth of rail transit capacity and changing passenger demand, passenger flow forecast, the necessity is even more obvious [3]. As a pivotal topic in urban rail transit, passenger flow prediction has seen the emergence of numerous directions and models. To date, there are three commonly used models for short-term passenger flow prediction: statistical models such as Autoregressive Integrated Moving Average (ARIMA) and Kalman Filter (KF), machine learning models such as Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN), and hybrid prediction models [4].

As a variant of the Recurrent Neural Network (RNN), the LSTM model incorporates a memory cell and gating mechanism, which enhances its efficiency in processing sequential data. Consequently, LSTM has been widely applied to time series data prediction [5-7]. Due to the extensive use of the LSTM model, various researchers have conducted comprehensive studies and improvements, often integrating LSTM with other algorithms to form hybrid prediction models. For instance, Zhang et al. employed wavelet denoising to mitigate the impact of noise on the LSTM model when predicting rail transit passenger flow [8]. Similarly, Qi et al. introduced ensemble empirical mode decomposition and Bayesian optimization algorithms to enhance the LSTM model, thereby constructing a hybrid prediction model [9]. It is evident that the LSTM model is predominantly used in most passenger flow predictions. However, researchers often implement specific improvements to the LSTM model to enhance its performance. Each model tends to excel in certain aspects of prediction, but to better accommodate the complex network structures of transportation forecasting, some scholars have proposed hybrid models. These hybrid models combine the strengths of different individual models to create a new model that overcomes the limitations of single-model predictions. This approach aims to improve the accuracy and reliability of passenger flow predictions in transportation systems [10].

This study uses daily metro passenger flow data from Chengdu between May and September 2019 as a reference. After processing and optimizing the data, LSTM model and an Arima model are established to predict passenger flow across different metro lines. By comparing the prediction results of various models, and try to identify the most accurate prediction model, thereby providing reliable forecasting information for metro traffic departments. This information is intended to aid in the scheduling and operation of metro services, ensuring that appropriate plans and responses can be implemented to handle high passenger volumes. Ultimately, this will enhance travel safety and efficiency for passengers [11].

2. Methods

2.1. Data Source

The data used in this study is sourced from the Chengdu Public Data Open Platform, specifically the daily metro ridership statistics of Chengdu from May to September 2019. The original dataset includes daily entry and transfer volumes for six metro lines in Chengdu, comprising a total of 911 entries. A sample of this data is shown in Table 1.

Table 1. Sample passenger flow data

Field name	Data type	Description	Example
date	object	date	2019/5/1
line	int	subway route	1
entry	int	number of arrivals	434257
transfer	int	transfer number	264399

2.2. Indicator Selection and Description

This study focuses on the entry volumes of Line 1 from May to September, extracted from the original dataset, for metro ridership prediction research. To ensure the accuracy, simplicity, and cleanliness of the data, a preprocessing step was conducted to remove anomalies and non-compliant data, such as excessively large or small values. Data from May to August was designated as the training set, and data from September was used as the test set.

2.3. LSTM Model

LSTM is a variant of RNN widely used for time series prediction. LSTM enhances RNN by incorporating three gating mechanisms: the forget gate, input gate, and output gate. These gates filter the input information and update the states between units, facilitating information flow across different layers. Figure 1 illustrates the structure of the LSTM model.

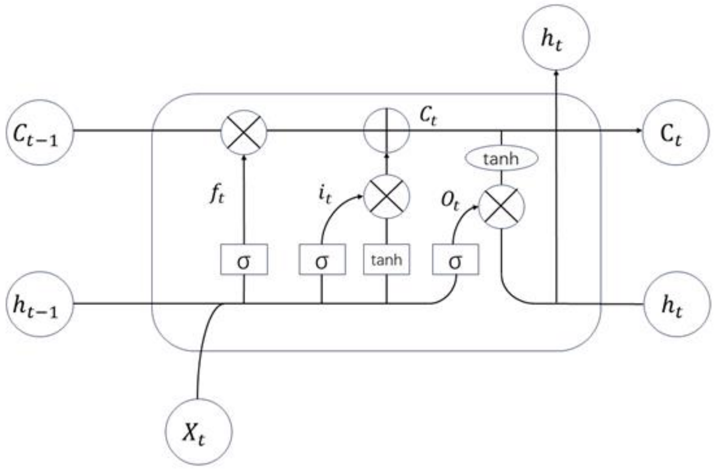

Fig. 1 LSTM model structure

LSTM Forget Gate Equation:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

Where W_f represents the weights between the input and the previous hidden state h_{t-1}, b_f is the bias, and f_t is the final transmitted information. LSTM Input Gate Equations:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh\,(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

Here, W_i and W_c are the weights between the input and the previous hidden state h_{t-1}, b_i and b_c are the biases, i_t is the input gate's output, and $\tilde{C}_t$ represents the candidate cell state. LSTM Output Gate Equations:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4}$$

$$h_t = o_t * \tanh\,(C_t) \tag{5}$$

In these equations, W_o denotes the weights between the input and the previous hidden state h_{t-1}, b_o is the bias, o_t is the output gate's output, and h_t is the hidden state at time t [12].

2.4. ARIMA Model

The ARIMA model consists of three main components: the Autoregressive model (AR), the Integrated process (I), and the Moving Average model (MA). The ARIMA model is particularly effective in handling various characteristics of time series data. It achieves this by analyzing and fitting historical data to forecast future trends and changes. Essentially, the ARIMA model leverages the autocorrelation and differencing of the data to extract the underlying patterns within the time series. These extracted patterns are then used to predict future values.

By capturing the dependencies and structure in the data, the ARIMA model can provide robust forecasts for time series with trends and seasonality. This makes it a valuable tool for understanding and predicting the behavior of complex time-dependent data.

3. Results and Discussion

3.1. LSTM Prediction Results

To facilitate data handling, enhance model training efficiency, and avoid potential negative impacts on prediction performance, the original dataset was normalized because the value of the original dataset is large. After preprocessing, the visualized data is shown in Figure 2.

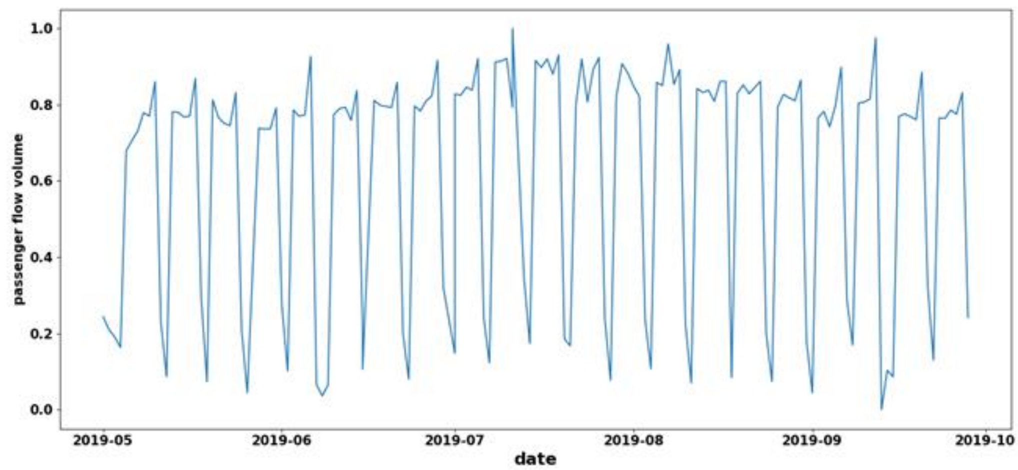

Fig. 2 Visualization of subway passenger flow data

The visualization clearly reveals that the passenger flow exhibits a distinct pattern, approximately cyclical on a weekly basis, with consistent peaks and troughs. The peaks predominantly occur on weekdays, while the troughs are observed during weekends, which aligns closely with real-world observations.

Based on the features identified in the data visualization, this paper proceeded to construct the LSTM model and split the data into training and test sets. Four-fifths of the data, corresponding to May through August, was used as the training set, while the remaining data from September was used as the test set.

The LSTM model was built with two layers, each containing 50 units, and a single output node. The model was trained over 100 epochs. The predicted passenger flow data compared to the actual data is shown in Figure 3.

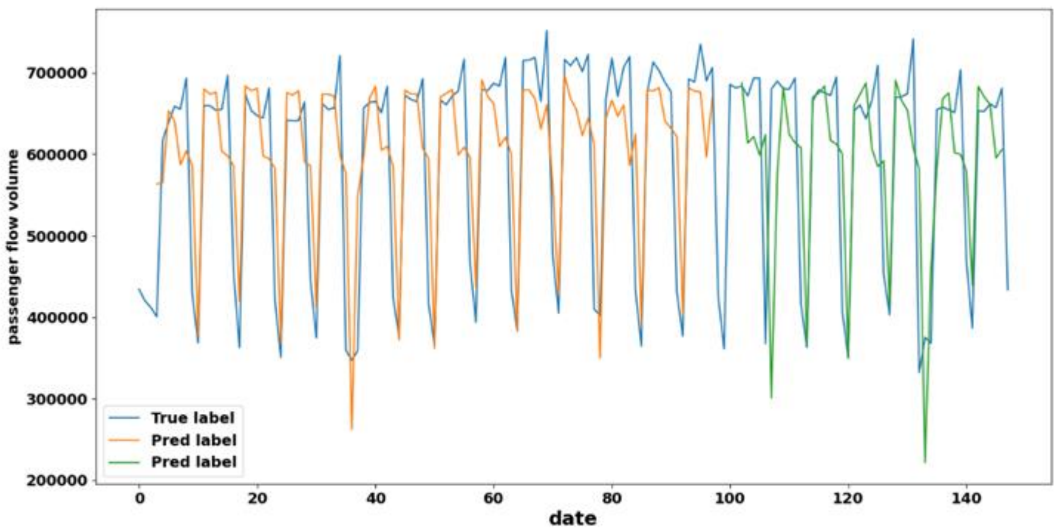

Fig. 3 Entry prediction results

The figure illustrates that the predicted data curve closely follows the actual data trends in terms of peaks and troughs, indicating that the LSTM model can effectively capture the patterns in passenger flow.

3.2. ARIMA Model Results

The Augmented Dickey-Fuller (ADF) test was used to assess the stationarity of the data. The null hypothesis of the ADF test is that there is a unit root, implying non-stationarity. Rejection of the null hypothesis indicates stationarity. As shown in Table 2, after the first difference, the sequence is stable and allowing to proceed with ARIMA model construction.

Table 2. Results of ADF test

difference order	t	p
0	-1.800	0.380
1	-7.009	0.000
2	-7.336	0.000

The initial step in constructing the ARIMA model involves observing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the time series. The values of parameters p and q can be roughly determined by examining the truncation and tailing characteristics. Truncation refers to the residual sequence's autocorrelation coefficients becoming small or near zero after a certain lag, while tailing refers to the coefficients maintaining a high value or decreasing slowly at higher lags. Based on Figure 4, the chosen model parameters are p=5 and q=7.

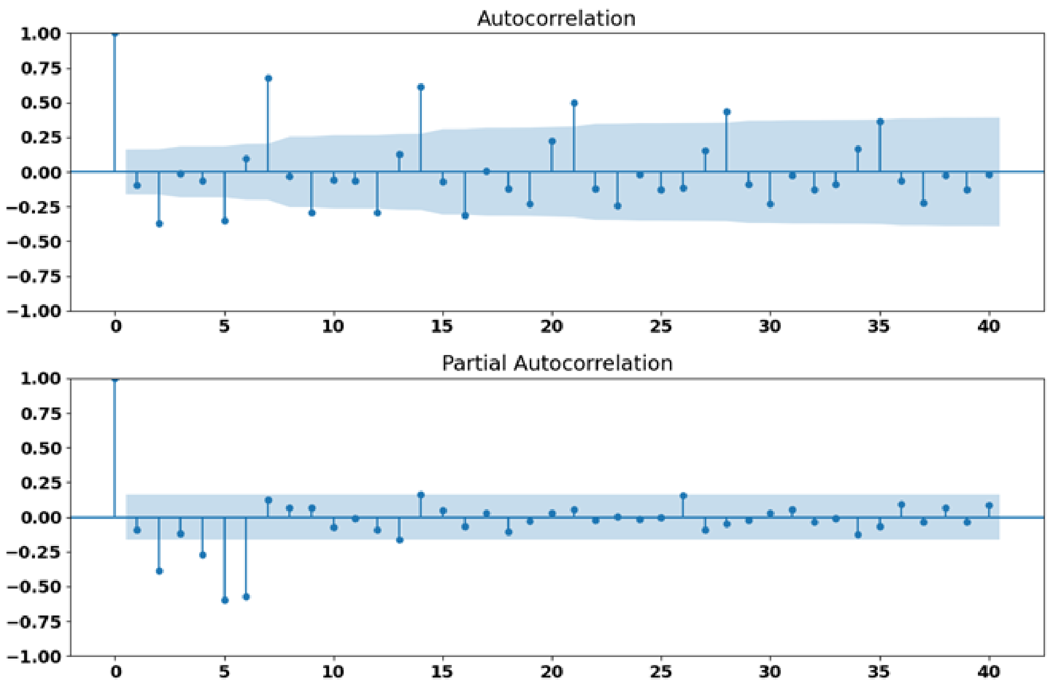

Fig. 4 Results of ACF and PACF

The comparison between the predicted values and the actual data is shown in Figure 5, demonstrating that the model is generally reliable.

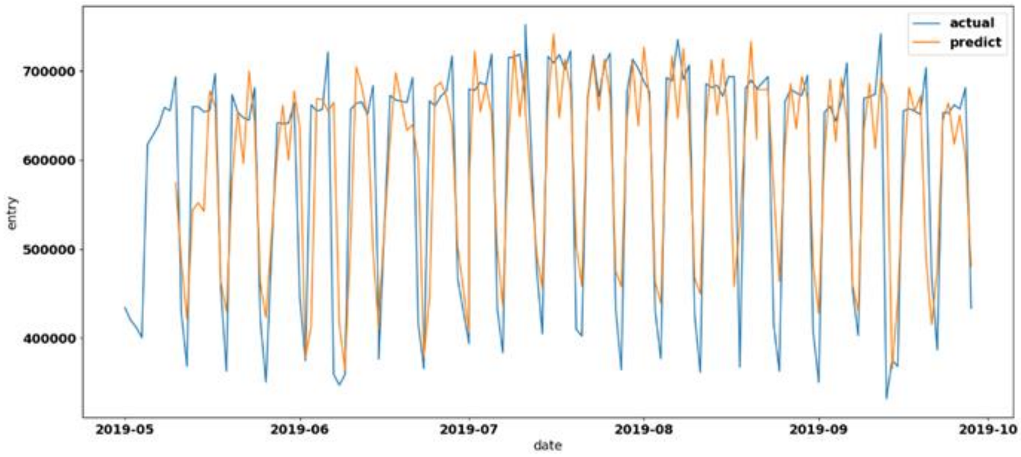

Fig. 5 Entry prediction results

In order to verify whether the parameters of ARIMA model have a better choice, this paper also adopts the combination of different parameters to obtain the minimum value of BIC and AIC to determine the best parameters (Table 3). The end result is shown in Figure 6.

Table 3. Results of AIC and BIC

	value	model
AIC	3793.128	(4, 4)
BIC	3823.100	(4, 4)

The prediction results from the ARMA (4, 4) model are shown in Figure 6.

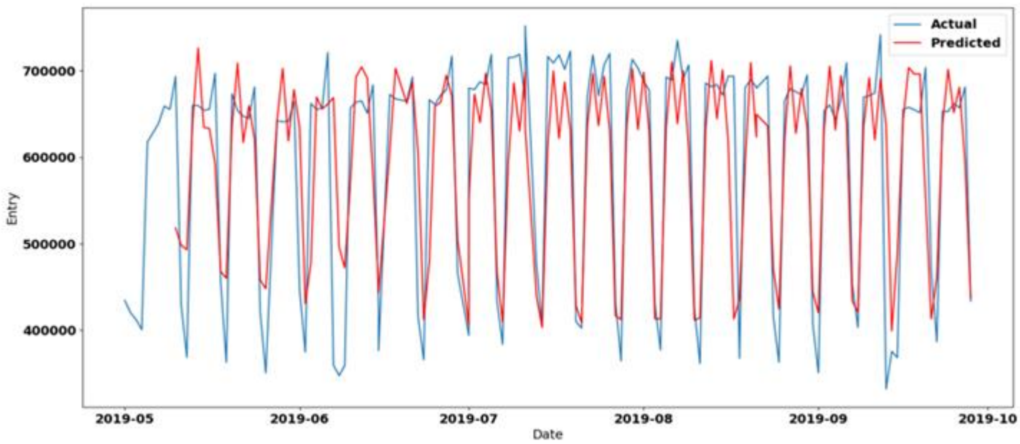

Fig. 6 Entry prediction results

By comparing the RMSE values obtained by the two models, the RMSE value of ARIMA (5, 1, 7) is 81626.798, and the RMSE value of ARMA (4, 4) is 78636.729. Finally, the ARMA (4, 4) model is selected in this paper.

3.3. Discussion

Both the LSTM and ARIMA models provided reliable predictions under the conditions of this study. By comparing the RMSE value 78636.729 of the ARIMA model with the RMSE value 70462.242 of the LSTM model, it can be concluded that the LSTM model is superior to the ARIMA model. However, it is evident that both models can effectively simulate data trends close to the actual data. Therefore, both LSTM and ARIMA models are suitable for univariate time series forecasting.

4. Conclusion

This study analyzed and forecasted metro passenger flow using entry passenger volume data from Chengdu Metro Line 1 from May to September 2019. LSTM and ARIMA models were constructed to evaluate their applicability in passenger flow forecasting. Based on the data used in this study, both models were able to fit the original data's trends and periodicity effectively. The prediction results from both models indicate that the LSTM model constructed in this study outperforms the ARIMA model in simulating the original data. It can thus be concluded that, for univariate data similar to that used in this study, the LSTM model can effectively simulate and forecast the original data.

However, the data used in this study is limited to a univariate time series, without considering other factors that might affect passenger flow. Future research should incorporate additional variables, such as weather conditions, holidays, and other factors, to enhance the accuracy and applicability of the forecasting models. This would provide practical assistance for metro management and operations, improve metro efficiency, and ensure passenger safety.

References

[1] Pan N. Urban rail transit passenger flow prediction based on ARIMA and LSTM. Scientific and Technological Innovation, 2022, (08): 165-168.

[2] He A, Wang B. Urban rail transit OD passenger flow prediction method based on improved LSTM algorithm. Comprehensive Transportation, 2021, 43(04): 67-72+87.

[3] Du X, Zhao X, Li L. Short-term passenger flow prediction of rail transit based on LSTM. Journal of Guizhou University, 2021, 38(05): 109-118.

[4] Xue Q, Zhang W, Ding M, et al. Passenger flow forecasting approaches for urban rail transit: A survey. International Journal of General Systems, 2023, 52(8): 919-947.

[5] Ren N. Short-term passenger flow prediction of urban rail transit using deep learning algorithms. System Simulation Technology, 2021, 17(04): 259-264.

[6] Cui H, Chen X, Yang, C, et al. Subway station passenger flow prediction based on deep long short-term memory network. Urban Rail Transit Research, 2019, 22(09): 41-45.

[7] Liu Y, Liu Z, Jia R. DeepPF: A deep learning based architecture for metro passenger flow prediction. Transportation Research Part C: Emerging Technologies, 2019, 101(APR): 18-34.

[8] Zhao Q, Feng X, Zhang L, et al. Research on short-term passenger flow prediction of LSTM rail transit based on wavelet denoising. Mathematics, 2023, 11(19): 4204.

[9] Qi X, Fu C. Improved LSTM method for short-term subway passenger flow prediction. Transportation Technology and Economy, 2024, 26(02): 58-64.

[10] Liu X. Passenger flow data prediction of rail transit stations based on deep learning. Beijing University of Posts and Telecommunications, 2024.

[11] Zhang Y. Short-term passenger flow prediction method for urban rail transit based on deep learning. Lanzhou Jiaotong University, 2022.

[12] Liu X. Passenger flow data prediction of rail transit stations based on deep learning. Beijing University of Posts and Telecommunications, 2024.