Enhancing Intelligent Mobility: LSTM Neural Networks for Short-Term Passenger Flow Prediction in Rail Transit

Jiajun Zhang[1,*] and Jingrong Zhang[2]
[1]College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, 300457, China
[2]School of Transportation Engineering, Dalian Jiaotong University, Dalian, 116028, China

Keywords: URS, Short-Term Passenger Flow Prediction, LSTM, Traditional Time-Series Prediction Models.

Abstract: In Urban Rail Systems (URS), traffic flow prediction has a long history. However, due to the inherent high non-linearity and randomness of transportation systems, it remains a challenging issue. Based on the passenger flow data from subway stations in Seoul, South Korea, this study aims to conduct short-term passenger flow predictions within the Seoul Metropolitan Subway system in South Korea using the Long Short-Term Memory (LSTM) network model, thereby verifying the effectiveness and accuracy of the LSTM model in urban subway passenger flow prediction. The model aids in facilitating early safety warnings and evacuations for passenger flow. According to the results, the LSTM model is more accurate for short-term passenger flow prediction in a high traffic station ("Geongdeok Station") and for a stable station ("Jamwon Station"). Compared to traditional time-series prediction models, LSTM shows superior forecasting capabilities. The findings and methodology in this study could serve as references and lessons for other researchers in similar fields.

1 INTRODUCTION

With the accelerated progress of urbanization, rail transit is become one of the indispensable means of transportation in most of the cities. and as the rapid development of urbanization and the continuous increasing of population, the passenger flow of URS is showing an accelerated trend of increase. Accurate short-term prediction of passenger flow at rail transit stations play an import role in the optimization of operation management and the improvement of transportation efficiency (Wu et al 2019). Zhao et al, proposed that the application of short-term traffic flow prediction in subway systems can plan subway departure intervals rationally based on the prediction results and serve as a reference for subway and light rail departure intervals (Zhao et al 2019).

Regarding existing traffic prediction models, these models can be categorised into two main types: parametric models and non-parametric models. Parametric models use particular assumptions to estimate parameters, providing simplicity, interpretability, and computing efficiency. However, they may struggle with complex non-linear relationships and accuracy if assumptions are not met. Non-parametric models, on the other hand, are more flexible and adaptable, capable of capturing complex data patterns. However, they often require more training data and have more complex structures, potentially leading to overfitting. The classic parametric models include the historical average model, regression analysis method, Bayesian methods, etc (Tang et al 2021 & Sun and Wei 2017). Common non-parametric models include the K-nearest neighbours approach, support vector machines, neural networks, and others (Cai et al 2016 & Liyanage et al 2022).

Deep learning has achieved significant advancements in various domains, including image identification, natural language processing, speech recognition, and recommendation systems. This has substantially accelerated the progress and use of artificial intelligence technologies. Armando Fandango et al. proposed the application of iterative strategies in constructing Recurrent Neural Network (RNN) models for short-term traffic flow prediction (Fandango and Wiegand 2018). Although RNN models are applied to traffic passenger flow prediction because of their ability to process sequential data, they have some drawbacks in practical applications, especially the problems of gradient vanishing or explosion, which affect the

Zhang, J. and Zhang, J.
Enhancing Intelligent Mobility: LSTM Neural Networks for Short-Term Passenger Flow Prediction in Rail Transit.
DOI: 10.5220/0012797600004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 410-415
ISBN: 978-989-758-690-3

stability of model training and the accuracy of prediction (Lv et al 2015). LSTM is a unique iteration of the Recurrent Neural Network (RNN) model, making it more effective than conventional RNN models. The LSTM model incorporates forget gates, input gates, and output gates to effectively capture the long-term dependencies in time series data, resulting in enhanced prediction accuracy. Furthermore, due to the LSTM model's superior ability to handle high-dimensional data and extract valuable information, numerous researchers employ deep learning techniques to analyse high-dimensional spatiotemporal data. Shi et al. conducted a comparative analysis of the efficacy of random forests, Back Propagation (BP) neural networks, and LSTM models in predicting rail transit flow. They concluded that the LSTM model exhibited superior fitting results and demonstrated more comprehensive predictive capabilities and higher average prediction accuracy compared to the other models (Shi et al 2020).

This paper mainly utilizes the LSTM model to predict urban subway passenger flow, reaffirming the accuracy and irreplaceability of the LSTM model in the field of short-term passenger flow prediction, and providing more accurate passenger flow information for subway operation managers, offering a scientific basis for decision-making.

2 METHOD AND DATE

2.1 Data Source and Preprocessing

The dataset employed in this study consists of passenger flow data from the Seoul Metropolitan Subway system, spanning from 2015 to 2018. This dataset is publicly accessible on the Kaggle website. It records the number of passengers entering and exiting each of the 275 stations of the Seoul subway from 5 AM to midnight daily, with a time resolution of one hour. The format of the original data is presented as shown in Table 1.

Data preprocessing is a crucial initial step in data analysis, aiming to ensure the quality and consistency of the data for effective analysis and modeling. The Jupyter Notebook development tool and the Python programming language were used in this paper to preprocess the data. The pandas library was used for data cleaning, while the glob library was used to look for data file directories and paths. By removing duplicate entries, confirming data types, looking for missing values, and doing statistical analysis, the dataset's accuracy and integrity were verified. There were no missing or duplicate records discovered, and there were no outliers in the number of passengers for any given period.

Because different features differ significantly from one another, it is easy for little data to be missed during training. Thus, for analysis to be effective, all data must be normalized.

$$\mathbf{x}_{\text{norm}} = \frac{\mathbf{x}-\mathbf{x}_{\text{min}}}{\mathbf{x}_{\text{max}}-\mathbf{x}_{\text{min}}} \quad (1)$$

Where, x_"norm" represents the normalized value. x is the sample value, x_"min" is the minimum value, and x_"max" is the maximum value. Normalization, by unifying the data scale, enhances the learning efficiency of the algorithm and reduces the risk of gradient vanishing or explosion, thereby improving the performance and stability of short-term prediction models. Normalization, through standardizing the data scale, not only increases the efficiency of algorithm learning but also minimizes the risk of gradient disappearance or explosion, consequently enhancing the effectiveness and steadiness of short-term forecast models.

2.2 Cluster Analysis

By evaluating the correlation between the silhouette coefficient of the dataset and the number of clusters, it can be deduced that the ideal number of clusters is either 2 or 4, as depicted in Fig. 1. From a research perspective, two clusters often lack significant research value; therefore, four clusters were selected for subsequent clustering.

The 3D visualization following clustering clearly indicates that the clustering effect on the original data is not pronounced. Therefore, the data was subjected to PCA (Principal Component Analysis) for dimensionality reduction before clustering, as shown in Fig. 2.

Table 1: Setting Word's margins.

USE_DT	station_code	station_name	division	05~06	……	23~24
2018/11/1 0:00	2530	Gongdeok	in	74	……	342
2018/3/26 0:00	309	Jichuk	Out	19	……	2

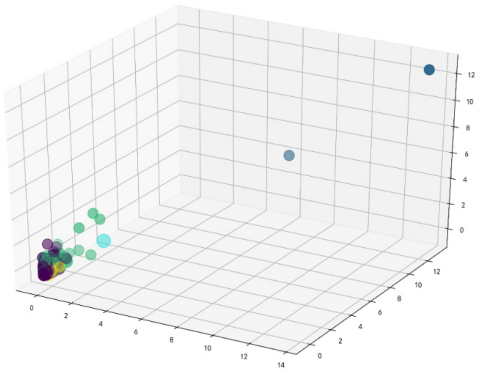

Figure 1: Initial Clustering Results of K-means (Picture credit: Original).

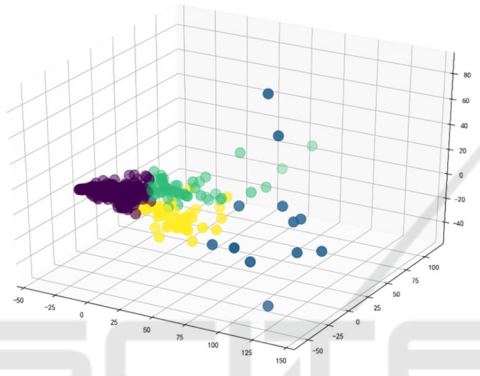

Figure 2: Clustering Results after Dimensionality Reduction Using K-means (Picture credit: Original).

Fig. 3 is a heatmap of the clustering results for 20 selected stations on the Seoul Metro line in South Korea. For clearer visualization, the passenger flow is categorized into four classes, ranging from high to low. It is observable that, spatially, the passenger flow radiates from areas of higher to lower density. There is a strong correlation in passenger flow among these stations.

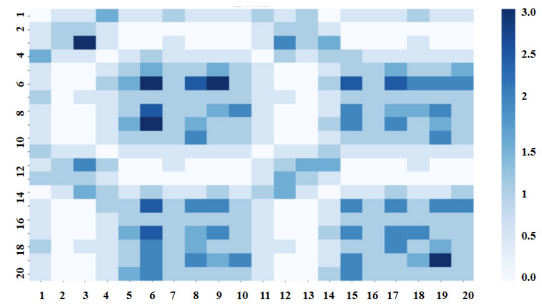

Figure 3: Spatial Heat Map (Picture credit: Original).

2.3 LSTM Model

The LSTM model is a sophisticated recurrent neural network architecture that stands out due to its internal structures, capable of maintaining a long-term flow of information when processing sequential data. The core component of the LSTM model is the cell state, often metaphorically referred to as a 'highway' for information transfer between network layers. It remains almost unchanged, effectively preserving long-term information continuity. To precisely control the flow of information, the LSTM model is equipped with three meticulously designed gate mechanisms: the forget gate determines which irrelevant information to discard from the cell state; the input gate controls the entry of new information; and the output gate decides what information to output based on the cell state. The intricate weight and bias parameters acquired through the network's training process govern these gates, enabling the LSTM model to excel and exhibit remarkable adaptability in various sequence prediction tasks, particularly those involving extended time intervals and delays. LSTM, an RNN variation, solves the RNN's long-term reliance issue. Fig. 4 shows a visualization of the procedure (Yousfi 2017).

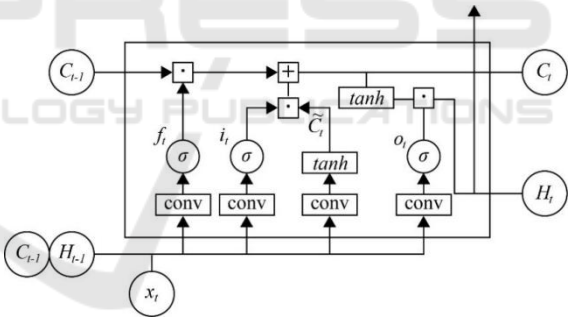

Figure 4: LSTM Flowchart (Picture credit: Original).

The formula for the forget gate in an LSTM is:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (2)$$

The forget gate is responsible for selecting and discarding specific information from the cell state.

The formulas for the input gate and the candidate values in an LSTM are:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (4)$$

The input gate determines the selection of fresh information to be stored in the cell state, whereas the candidate values encompass the potential information that can be incorporated into the cell state.

The formula for the cell state update in an LSTM is:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \quad (5)$$

The cell state is updated through the process of discarding previous information and incorporating new information.

The formulas for the output gate and the output value in an LSTM are:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t * \tanh(C_t) \quad (7)$$

The output gate determines the specific portion of the cell state that will be emitted, and the resulting output value is the refined representation of the cell state regulated by this gate.

The core of an LSTM network is its cell state C_t,which is updated through a gating mechanism. These gates control the flow of information in and out, including the forget gatef_t, the input gatei_t,and the output gateo_t.W_f,W_i,W_o,$W_c$represent the weight matrices for the forget gate, input gate, output gate, and cell candidate values respectively, each determining how information flows and updates within the model.b_f,b_i,b_o,b_c are the bias terms for the forget gate, input gate, output gate, and the cell state candidate vector, used to adjust the activation level of each gate and cell candidate vector. The LSTM equations employ the sigmoid function to control the selective transmission of information and the hyperbolic tangent function to modulate the value range of the cell state and output. The LSTM equations describe how the long short-term memory units dynamically regulate the long-term storage and forgetting of information by combining current input with past state information, using a complex gating mechanism, thereby solving the vanishing gradient problem found in traditional recurrent neural networks while maintaining the temporal dependencies of sequential data.

3 RESULTS AND DISCUSSION

3.1 Temporal Segmentation

The total daily passenger traffic of the Seoul subway in South Korea is depicted in Fig. 5. It is observed to exhibit a bimodal pattern, with peaks occurring between 8-10 AM and 6-7 PM. This aligns with the commuting times for most people and the school hours for students, consistent with real-world scenarios.

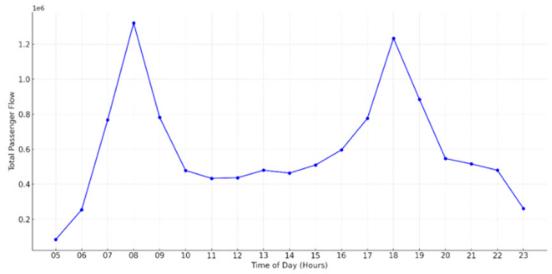

Figure 5: Seoul Subway daily station passenger flow (Picture credit: Original).

3.2 Temporal Segmentation

As shown in Fig. 6, the horizontal axis represents the daily passenger flow in December 2018, with each day serving as an observation point. The observed results indicate a significant periodic pattern in passenger flow, particularly evident in the troughs observed on December 2nd, 9th, and 16th, with a strict seven-day interval between these dates. This periodic pattern is likely influenced by the changes between weekends and weekdays.

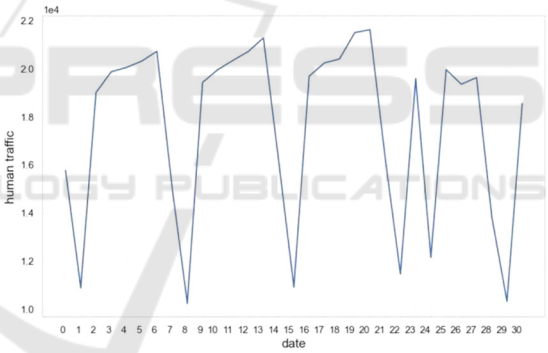

Figure 6: Seoul Subway one-month passenger flow (Picture credit: Original).

3.3 Spatial Characteristics

Fig. 7 demonstrates the significant differences in passenger traffic across different subway stations. This highlights the necessity of analyzing and comparing the daily passenger flow distribution at various subway stations.

Due to the significant differences in passenger flow distribution characteristics among various subway stations, this study has selected two representative stations for in-depth analysis. "Geongdeo Station" (transfer station) and "Jamwon Station" (non-transfer station).

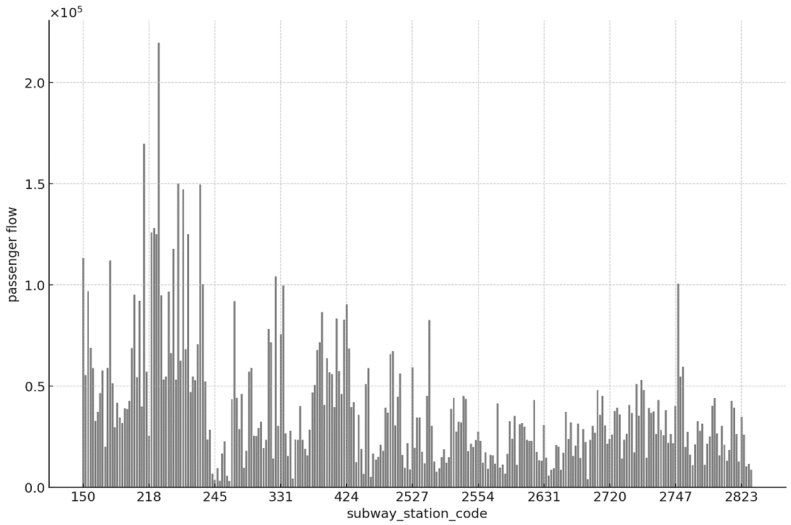

Figure 7: One day's passenger flow at each subway station (Picture credit: Original).

3.4 Results

Given the complexity of the Seoul subway system in South Korea, which comprises 23 lines, this paper focuses solely on "Geongdeo Station" and "Jamwon Station" for study. Gongdeok Station, as a transportation hub connecting Lines 5 and 6 of the Seoul Subway, has a high volume of passenger flow and complex traffic patterns due to its transfer characteristics. On the other hand, Jamwon Station, a standard station on Line 3, has a more stable flow of passengers, reflecting the daily commuter status of ordinary stations in the Seoul subway system.

The selection of these two stations as research subjects aims to compare and analyze the passenger flow characteristics of different types of subway stations. The data from Gongdeok Station helps to understand and predict the fluctuation of passenger flow at transfer stations during different periods. Meanwhile, Jamwon Station provides a benchmark for assessing the accuracy of the model in predicting regular passenger flow.

After a thorough analysis of the graphical data characteristics, this study selected a specific 8-day period as the subject for short-term passenger flow prediction based on LSTM. The parameter adjustments of the LSTM model used are detailed in Table 2.

Table 2 : Parameter Tuning.

learning_rate	Batch_Size	Epochs
0.01	32	100

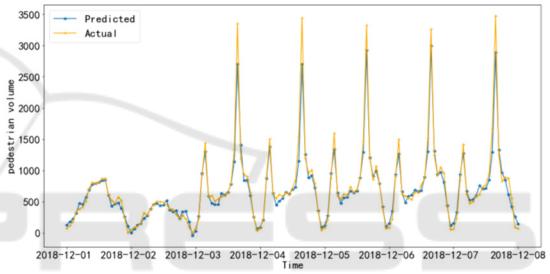

Figure 8: 2530 station Predicted data and true results of the test set (Picture credit: Original).

Fig. 8 presents the experimental results for passenger inflow at "Geongdeo Station" (station_code: 2530). The horizontal axis represents the date, while the vertical axis indicates passenger volume. The time interval spans from 5 a.m. to 24 p.m. each day. The blue star line denotes the predicted values, and the yellow plus line represents the actual values. It can be observed that most of the predictions closely align with the real values, though there are slight deviations in certain areas.

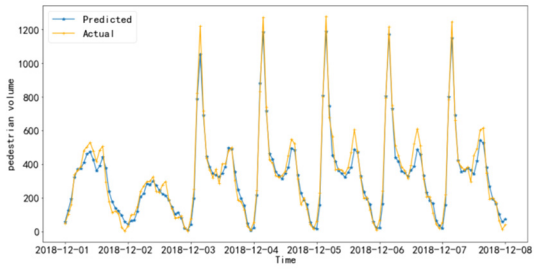

Figure 9: 328 Station Predicted data and true results of the test set (Picture credit: Original).

Fig. 9 illustrates the predicted inbound passenger flow results for "Jamwon Station" (station code: 328). Similar to the case of "Geongdeo Station," the overall predictions are accurate.

In this paper, ARIMA and BP models are used as comparative models. The study utilised two evaluation metrics, namely Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE), to conduct error analysis. As shown in Table 3, it is evident that, whether it is MAPE or RMSE, the performance of the LSTM model is superior to the other models.

Table 3: MAPE and RMSE of Three Test Models.

Model	MAPE/%	RMSE
LSTM	9.19	156.67
BP	15.86	228.74
ARIMA	18.43	279.87

4 CONCLUSION

This paper focuses on short-term passenger flow prediction for specific stations within the Seoul Metropolitan Subway system in South Korea using LSTM network model. The LSTM network is capable of learning time series with long-term dependencies, offering a flexible framework for employing various combinations of variables. The study not only demonstrates the effectiveness of the LSTM model in forecasting passenger flows at different types of subway stations but also highlights the advantages of deep learning in handling high-dimensional spatiotemporal data. These findings provide more accurate passenger flow information for subway operation management and scientific decision-making support for policymakers. While this research focuses on the impact of time on traffic flow, it does not account for other factors such as commercial zones or weather conditions. Future studies could consider incorporating these additional elements, including weather and the presence of commercial areas, to further enhance the model's predictive accuracy.

AUTHORS CONTRIBUTION

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

REFERENCES

Z. Q. Wu, T. Y. Huang, Y. W. Yan, M. S. Zhao, Urban Mass Transit. 22, 31 (2019)
H. Zhao, D. M. Zhai, C. H. Shi, Urban Mass Transit 32, 50 (2019)
J. Q. Tang, X. W. Zhong, J. Liu, T. R. Li, J. Chongqing Jiaotong Univ. (Nat. Sci.) 40, 31 (2021)
B. Sun, M. Wei, Highways Autom. Appl. 4, 20 (2017)
P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, J. Sun, Transp. Res. Part C Emerg. Technol. 62, 21 (2016)
S. Liyanage, R. Abduljabbar, H. Dia, P. W. Tsai, J. Urban Manag. 11, 365 (2022)
A. Fandango, R. P. Wiegand, Towards investigation of iterative strategy for data mining of short-term traffic flow with recurrent neural networks, in Proceedings of the 2nd International Conference on Information System and Data Mining (2018)
Y. Lv, Y. Duan, W. Kang, Z. Li, F. Y. Wang, IEEE Trans. Intell. Transp. Syst. 16, 865 (2015)
X. R. Shi, C. H. Wang, D. J. Liu, X. Zhang, B. Zhang, Electr. Technol. Softw. Eng. 182 (2020)
S. Yousfi, S. A. Berrani, C. Garcia, Pattern Recogn. 64, 245 (2017)