Prediction of Metro Short-term Entry Flow Based on Passenger Flow Characteristic Analysis

Shuzhi Zhao1,a, Chongyu Zhang2,b, Huasheng Liu3,c*
1 2 3College of Transportation, Jilin University, Changchun 130022, China
*aemail: zhaosz@jlu.edu.cn, bemail: zhangcy_jlu@163.com,
c*Corresponding author’s e-mail: liuhuasheng521@163.com

Abstract: Prediction of short-term incoming passenger flow at metro stations is of great significance to the stable operation of Metro networks. Taking AFC data of Chengdu Metro as data source, based on Stochastic Forest model, support vector machine regression model and neural network model in machine learning method, this paper makes short-term prediction of metro entry flow and comparative analysis of three models. Metro stations are divided into four types according to the passenger flow characteristics: residential type, commercial type, business type and terminal type. Three models are trained in workdays and weekends conditions for different types of stations. The average absolute percentage error (MAPE) and root mean square error (RMSE) were used to evaluate the accuracy and stability of the prediction results. The results show that BP neural network has the best comprehensive performance and random forest has better prediction accuracy for stations with strong periodicity.

1. Introduction
With the acceleration of the urban rail transit network process, grasping the real-time changes of passenger flow accurately has become a key element to guarantee the transportation order effectively and improve the quality of transportation services. High-precision short-term passenger flow forecasting for inbound stations can help managers respond to passenger flow fluctuations quickly, adjust transportation plans in time, alleviate the uneven passenger flow of urban rail transit lines, and improve the quality of transportation services. At the same time, it also provides necessary decision-making support for the rational layout of station facilities to ensure the safe and efficient operation of the rail transit system.

For short-term passenger flow forecasting, it can be divided into two categories: parametric methods (such as autoregressive integral moving average, Kalman filtering.) and non-parametric methods (such as support vector machines, neural networks, etc.). In the study of passenger flow prediction based on parametric method, Xiong Jie et al. [1] built a subway transfer passenger flow model based on the Kalman filter principle, and found that the passenger flow prediction during the morning peak of the working day is more accurate; Wang Yi et al. [2] used Marko Fulian improved the GM (1,1) gray prediction model and proved that it has high accuracy in the prediction of volatile passenger flow; Cai Changjun et al. [3] constructed an ARIMA inbound and outbound prediction model, and verified that the model is good Forecast accuracy. In the study of non-parametric passenger flow forecasting, Zou Wei et al. [4] established a genetic algorithm modified prediction model based on wavelet neural network, and obtained the prediction accuracy compared with genetic algorithm optimized BP neural network and single wavelet neural network Higher; Jiang X [5] established a non-parametric dynamic
time delay recursive wavelet neural network prediction model with the help of wavelet neural network, and proved that it has excellent performance in both short-term and long-term prediction; Dong Shengwei [6] based on BP neural network The network model uses genetic algorithms to improve short-term predictions of rail transit stations, cross-sections, and interchange passenger flows, which greatly improves the prediction accuracy of the model. Some scholars also compare parametric and non-parametric methods. For example, Castro-Neto M[7] established an OL-SVM model to predict traffic flow, and verified it by comparison with Gaussian maximum likelihood method and Holt exponential smoothing method.

In the current research on traffic flow prediction, most of the time granularity selected is 15 minutes or more. Due to the large sampling time span and the obvious numerical characteristics of the sample data, the noise factors in the real-time data are not considered enough, and the prediction method is difficult to apply to the real environment. In order to make the model more suitable for real scenarios, the time granularity of inbound passenger flow statistics selected in this paper is 5 minutes, and based on the analysis of the characteristics of the site passenger flow, the inbound passenger flow of different types of stations on working days and non-working days are separately carried out. Forecast, and finally get the best prediction model corresponding to different types of sites. In addition, the currently obtained data has problems such as increased data dimensionality, low data quality, and strong data randomness. Machine learning methods can improve their performance by continuously learning data, and obtain models with higher prediction accuracy. Therefore, it is of great practical significance to conduct model comparison and analysis based on machine learning methods, and to find models with strong robustness and high accuracy. In machine learning methods, models such as SVM and neural networks can be seen to have a strong fitting effect from previous studies, while random forests have excellent performance in both classification and regression problems, so this article selects these three models for short-term passenger flow forecasting research.

2. Analysis on Characteristics of Passenger Flow in Metro Station

A thorough understanding of the time distribution of inbound passenger flow is an important prerequisite for inbound passenger flow forecasting research. This article first preprocessed Chengdu’s April 2018 AFC data, filtered and eliminated AFC abnormal data, and also excluded data with a single journey time of more than three hours to ensure the validity and effectiveness of the incoming data. Authenticity. Finally, the original data of inbound passenger flow was counted at a granularity of 5 minutes.

This paper is based on the 5-minute time granularity of inbound data from each station of Chengdu Metro Line 2, and according to the time distribution characteristics of passenger flow of each station, the stations are divided into residential area, commercial area, business area, and hub type. Four categories. As shown in Figure 1 and Figure 2, different types of stations exhibit different characteristics of passenger flow in the same time series. For hub-type and commercial-area-type stations, the inbound traffic on weekdays or weekends is relatively large, and passenger flow fluctuations are also more obvious. Residential and business district stations exhibit a single peak in passenger flow during workdays, and their respective morning peak and evening peak characteristics are obvious. Inbound passenger flow at weekend stations has been in a low state and has small fluctuations.
In addition, the passenger flow distribution of the same station is also relevant and periodic on weekdays and weekends. As shown in Figures 3 and 4, there is a strong correlation between the passenger flow distribution on Tianhe Road on April 18 on weekdays and the passenger flow distribution of the previous week and the day before. On weekends, the passenger flow distribution on April 22 of Chunxi Road and the previous day The passenger flow distribution of the week and the previous day also has a strong correlation, but in comparison, after 19:00, the passenger flow distribution on April 15 and April 22 is more relevant, and the passenger flow trend on April 21 is similar. But the deviation is slightly larger. Based on the above characteristics of passenger flow, this article divides the forecast scenarios into two types, working days and non-working days, for different types of stations.

3. Model building
Combining the above passenger flow distribution characteristics, the model's training set and test set feature selection will fully consider the periodicity and relevance of the passenger flow distribution; for different prediction scenarios of different types of sites, model construction and training; finally select the average absolute percentage Error and root mean square error are used as the standard to measure the applicability of the model.

3.1. Establishment of data sets.
Because the inbound passenger flow has strong time correlation, the inbound flow in $q$ time period adjacent to the current period is considered in the model data set. As shown in formula 1, the reasonable
value of \( q \) can be obtained by calculating the autocorrelation coefficient of historical data. \( R_{i,x,q}^{i} \) is the self-correlation number of \( q \) periods of the station \( i \) lag on the \( x \) day; \( V_{x,n}^{i}, V_{x,n+q}^{i} \) is the station entry quantity \( n \cdot n + q \) on the site; \( \bar{V}_{x}^{i} \) is the mean value of the station \( i \)'s inbound volume on \( x \) day; \( N \) is the number of valid periods.

\[
R_{i,x,q}^{i} = \frac{\sum_{n=1}^{N-q} (V_{x,n}^{i} - \bar{V}_{x}^{i})(V_{x,n+q}^{i} - \bar{V}_{x}^{i})}{\sum_{n=1}^{N} (V_{x,n}^{i} - \bar{V}_{x}^{i})^2}
\]

\[
\bar{V}_{x}^{i} = \frac{1}{N} \sum_{n=1}^{N} V_{x,n}^{i}
\]

This paper takes 192 daily inbound passenger flow data of Chengdu Metro Line 2 as a sequence sample in April 2018, For 21 working days and 6 weekends in April, calculate their self-correlation. It is generally believed that when greater than 0.5, The correlation between adjacent periods in the sequence is significant. The results are shown in Table 1. If more than 80% of the total sample meets the condition of more than 0.5, It is considered that the adjacent previous period has a strong correlation, Therefore, it can be determined that the working day, weekend corresponding values are 4.

| Order \( q \) | Working days | Weekend |
|--------------|--------------|---------|
| 1            | 0.9607       | 0.9423  |
| 2            | 0.9276       | 0.8951  |
| 3            | 0.8876       | 0.8599  |
| 4            | 0.8396       | 0.8214  |
| 5            | 0.7882       | 0.7829  |
| 6            | 0.7279       | 0.7301  |
| 7            | 0.6574       | 0.6898  |

To sum up, if the passenger flow in the forecast period is \( V_{I_{2}}^{k} \), the mapping relationship between the characteristic variables and the dependent variables in the data set is shown in formula 3. In the weekend scenario, it can be seen from figure 2 that the deviation of passenger flow distribution on adjacent dates of the same site is slightly larger, so the passenger flow sequence of adjacent dates is eliminated, and the mapping relationship is shown in formula 4.

\[
F(V_{i-1,j}^{k}, V_{i,j-2}^{k}, V_{i,j-1}^{k}, V_{i,j}^{k-4}, V_{i,j}^{k-3}, V_{i,j}^{k-2}, V_{i,j}^{k-1}) \Rightarrow V_{i,j}^{k}
\]

\[
F(V_{i-1,j}^{k}, V_{i,j}^{k-4}, V_{i,j}^{k-3}, V_{i,j}^{k-2}, V_{i,j}^{k-1}) \Rightarrow V_{i,j}^{k}
\]

\( V_{I_{2}}^{k} \) for each site \( i \) week \( j \) day the \( k \) time period of incoming traffic. In addition, the training set data need to be normalized for SVM and BP neural networks.

3.2. Model building
Stochastic forests were first proposed by Leo Breiman in 2001 and are mainly applied to classification and regression problems. Its basic unit is the decision tree. By self-help method, a new training set is generated by randomly sampling samples and features from the original training set, and a decision tree is generated according to the new training set. Vote and output the results according to the training results of these decision trees. By using the idea of Bagging, random forest overcomes the disadvantage of relatively weak learning ability of decision tree, and because each tree selects some samples and attributes, In addition to good noise resistance and stability, the training model will not be over-fitted.
For support vector machine, the core idea of regression problem is: given training sample set \( \{x_i, y_i\} \) \( (i = 1, 2, 3, \ldots, n) \), for this sample set, \( x_i \in \mathbb{R}^n \) is n-dimensional input sample, \( x_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \), and \( y_i \in \mathbb{R}^n \) is output sample. By mapping \( \phi: \mathbb{R}^n \rightarrow H \), \( H \) as the feature space, the training samples are mapped from the initial low-dimensional sample space to the high-dimensional feature space and transformed into the linear regression problem in the feature space. Finally, the nonlinear regression of the low-dimensional space is realized. The regression function is as follows:

\[
f(x) = w^T \phi(x) + b
\]

In the regression algorithm of support vector machine, the error function is defined as the loss function. For type \( \varepsilon \) insensitive loss function \( \varepsilon \) is insensitive, that is, there is a maximum deviation of \( \varepsilon \) between tolerance \( y \) and \( f(x) \). If the absolute value of deviation does not exceed \( \varepsilon \), the loss can not be counted, otherwise the loss should be calculated. The loss function is:

\[
F(f(x), y, \varepsilon) = \begin{cases} 
0, & |y - f(x)| \leq \varepsilon \\
|y - f(x)| - \varepsilon, & |y - f(x)| > \varepsilon 
\end{cases}
\]

Therefore, for the regression problem, the objective function expression is:

\[
\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} F(f(x_i), y_i, \varepsilon)
\]

The neural network is widely used in passenger flow prediction, such as Back Propagation, BP, Radial Basis Function, convolution neural network and so on. Considering the structure and dimension of the data set, this paper selects the BP neural network to construct the model. For a given training set \( \{x_i, y_i\} \), the output function for each independent neuron is:

\[
y = f(\sum_{i=1}^{n} w_i x_i + b - \theta)
\]

\( f \) is the activation function of neurons. In this paper, the Sigmoid function is selected as the activation function. \( x_i \) input signal for neuron \( i \); \( w_i \) connection weight for neuron \( i \). BP objective functions of neural networks are:

\[
\min E = \frac{1}{2n} \sum_{j=1}^{n} \sum_{k=1}^{m} (y_{jk}^k - \bar{y}_{jk}^k)^2
\]

\( E \) is sample cumulative error ; \( y_{jk}^k \) is training output of \( j \) neurons in training case \( k \); \( \bar{y}_{jk}^k \) is output of \( j \) neurons in training case \( k \). Finally, the parameters are updated by the bias derivation of the weight \( \theta \) and the bias value \( b \).

4. Comparative analysis of results

Different models of machine learning methods have different hyperparameters. In this paper, the mean square error is minimized, and the hyperparameters of each model are searched by cross-validation method. The calibration results of model parameters are shown in Table 2.

| Model | Time          | Parameters         | Features |
|-------|---------------|-------------------|----------|
| RF    | Working days  | \( N = 500 \)    | Criterion=mse |
|       | Weekend       | \( N = 750 \)    | Criterion=mse |
| SVM   | Working days  | Kernel=rbf        | \( C = 10 \) |
|       | Weekend       | Kernel=rbf        | \( C = 1 \) |
| BP    | Working days  | Activation=Sigmoid | Layers=[20,10] |
|       | Weekend       | Activation=Sigmoid | Layers=[50,10] |

To better measure the accuracy and stability of the model, this paper selects the average absolute percentage error (MAPE) and root mean square error (RMSE) of the site as indicators for comparative analysis of different models. The formulas for calculating the two indicators are:
\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%
\]

(10)

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

(11)

\(y_i\) and \(\hat{y}_i\) represent the actual and predicted values of inbound traffic in the \(i\) period of the station. The closer the value of the MAPE is to 0, the better the accuracy of the prediction is, and the smaller the value of the RMSE is, the stronger the stability of the fitting result is. For different types of sites, the results are shown in the table below. Table 3 shows the MAPE calculation results of different types of sites, and it can be seen intuitively that BP the comparison of neural networks, the prediction error is the smallest and the accuracy of prediction results is the highest for different prediction scenarios. Random forest and support vector machine have a little deviation in prediction effect and perform well in other scenarios.

| Site type            | RF  | SVR | BP  |
|----------------------|-----|-----|-----|
|                      | 2018/4/25 | 2018/4/29 | 2018/4/25 | 2018/4/29 | 2018/4/25 | 2018/4/29 |
| Residential type     | 13.65 | 16.90 | 12.99 | 15.17 | 15.53 | 16.00 |
| Business District    | 13.94 | 21.00 | 13.23 | 20.01 | 14.66 | 20.34 |
| Commercial district type | 8.87 | 14.17 | 8.53 | 12.97 | 8.97 | 9.02 |
| Hub type             | 25.78 | 22.46 | 29.35 | 26.17 | 20.29 | 23.60 |

Table 4 shows the RMSE calculation results of different types of sites, it can be seen that the stability of BP neural networks is the best in the three models. Random forests only have poor prediction effect at the end of commercial site.

| Site type            | RF  | SVR | BP  |
|----------------------|-----|-----|-----|
|                      | 2018/4/25 | 2018/4/29 | 2018/4/25 | 2018/4/29 | 2018/4/25 | 2018/4/29 |
| Residential type     | 15.06 | 17.05 | 15.77 | 17.63 | 19.98 | 19.70 |
| Business District    | 16.75 | 14.83 | 17.95 | 13.99 | 21.96 | 14.22 |
| Commercial district type | 38.94 | 174.91 | 37.87 | 108.79 | 45.38 | 90.74 |
| Hub type             | 78.54 | 140.35 | 91.26 | 177.73 | 77.33 | 174.88 |

The stochastic forest model only has poor prediction accuracy and large fluctuation in hub and commercial area stations. For analyzing the reason, this paper calculates the importance of each feature by measuring the prediction error [8] after random sorting of each feature in the test set. The results are shown in figure 5.

Among them LW, B1, B2, D1, D2, D3, D4 each attribute characteristic in the corresponding formula 3 is the passenger flow sequence of the previous week, the previous day, the first two days, the hysteresis one period, the hysteresis two periods, the hysteresis three periods, the hysteresis four periods respectively. As can be seen intuitively, the effect of attribute LW, D1 on prediction results is the greatest, while commercial area stations and hub stations often fluctuate greatly, which leads to poor prediction accuracy and stability of random forests. In addition, due to the small time granularity selected in this paper, the prediction results are also affected.
5. Conclusion

Based on the selection of a smaller time granularity, combined with model methods such as random forest, support vector machine, BP neural network, etc., the passenger flow prediction of subway inbound stations and a comparative study between different models are carried out. Finally, the following conclusions are drawn:

(1) In terms of comprehensive accuracy and stability, the BP neural network has the best overall performance in predicting different types of sites among the three models;

(2) The random forest model has the best prediction effect for stations with strong passenger flow cycles, and is most suitable for the prediction of residential stations. The BP neural network has a better prediction effect for commercial district and hub-type stations, indicating that the BP neural network is more suitable for stations with large passenger flow fluctuations.

(3) Through the analysis of the importance of each influencing factor of the random forest, it can be concluded that among the seven influencing factors, the passenger flow sequence engraved at the same time in the last week of the subway station has a greater impact on the prediction result, which verifies that the subway passenger flow has a strong periodicity.

In addition, if you consider improving the three models and adjusting the time granularity, the performance results of the three models may have new changes. This is the direction of the next stage of research.

Acknowledgments

This paper is one of the phased results of the National Natural Science Foundation of China's general project "Research on the Topology Analysis and Layout Optimization Model of Comprehensive Passenger Transport Hubs from the Perspective of Urban Agglomeration" (71871103).

References

[1] Xiong Jie, Guan Wei, Sun Yuxing. (2013)Subway transfer passenger flow prediction based on Kalman filter[J]. Journal of Beijing Jiaotong University,, 37(03): 112-116+121
[2] Wang Yi, Xu Ruihua. (2010) Research on Short-term Passenger Flow Forecast of Urban Rail Transit Based on Period and Time Variation Characteristics[J]. Urban Rail Transit Research, 13(01): 46-49.

[3] Cai Changjun, Yao Enjian, Wang Meiyin, Zhang Yongsheng. (2014) Urban rail transit passenger flow prediction based on product ARIMA model[J]. Journal of Beijing Jiaotong University, 38(02): 135-140.

[4] Zou Wei, Lu Baichuan, Deng Jie, Ma Qinglu, Qiu Shichong, Zhang Qin. (2014) Research on passenger flow prediction based on genetic algorithm and wavelet neural network[J]. Journal of Wuhan University of Technology (Transportation Science and Engineering Edition), 38(05):1148-1151+1157.

[5] Jiang X, Adeli H. (2015) Dynamic Wavelet Neural Network Model for Traffic Flow Forecasting[J]. Journal of Transportation Engineering, 131(10):771-779.

[6] Dong Shengwei. (2013) Research on Rail Transit Short-term Passenger Flow Forecast Method Based on Improved BP Neural Network [D]. Beijing Jiaotong University,

[7] Castro-Neto M, Jeong YS, Jeong MK, et al. (2009)Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions[J]. Expert Systems with Applications, 36(3):6164-6173.

[8] Nathans, L L, Oswald, F L, Nimon, K. (2012) Interpreting multiple linear regression: A guidebook of variable importance[J]. Practical Assessment, Research & Evaluation, 12:1–19.