Analysis and Prediction of Subway Ridership

-- Take a Station in Hangzhou as an Example

Jingyi Liu *

Southern University of Science and Technology, Shenzhen, China

* Corresponding author email: 12012639@mail.sustech.edu.cn

Abstract. Many people go by subway in China. Huge passenger flow brings much trouble to the police and passengers, such as crowded carriages, long waiting times and low-efficiency transport. So it is important to know the traffic jam before it brings many problems. With the data from Tianchi Competition, this paper analyzes and predicts the subway ridership of a station in Hangzhou based on time series and linear regression. Taking line 3, station 5 as an example, there were two peaks which have many passengers in the station. Combined with the results, it proposed that traffic police need to pay more attention to the rush hours. People who do not need to commute can avoid these times. The government may also use this prediction results in the subway service management, as well as in the planning for the future development and subway lines projection.

Keywords: Linear Regression; Time Series; Passenger Flow Prediction.

1. Introduction

Subway is a kind of vehicle that many people appreciate for its convenience, environment friendly, and efficiency. There are 1.4 billion people in China. And increasing travel demand made the traffic very heavy and many people go to work or school by subway. So, the ridership can be very high at some specific times and traffic police need to evacuate the traffic in time to ensure the safety of the passengers. These have brought many problems, such as crowded carriages, long waiting times and low-efficiency transport. So if people can predict the ridership of a station, it can be the best help for everyone. The prediction can help people know when the peak hours will be in advance, so the traffic police can manage the order in time, passengers can avoid going by subway at these peak hours and the management can change the frequency of the railway at different times.

Analyzing the feature of the ridership and setting up a model to predict its changes can not only help police to keep order but also help people to avoid taking the subway during rush hours. Many researchers have proposed many different models to predict ridership, for example, time series [1], neural networks [2], LSTM Model [3], only to name a few. In previous research, the non-parametric regression model was used by Yujuan Sun and was found that could predict the subway passenger flow accurately for different intervals [4]. And most models predict the passenger flow in a particular period of time using a piece of known data that was collected in the previous time [5].

In this paper, linear regression is thought to be a great solution to this problem, with the data of Hangzhou Metro's ridership from Jan. 1st to Jan. 25th in 2019 offered by Tianchi. Station 5 of line B was chosen as an example in this work for the analysis and prediction of ridership. For evaluation, the previous 24 hours' data will be used to predict the future ridership in the next twenty-fifth hour [6][7]. And other stations and lines' data will be used to verify the model.

In the next section, this paper will introduce the data set, where the data collection process and the features will be introduced. In the third section, the method will be presented, including the basic definition, principle, and formulas of linear regression. The fourth section is about the experiments. The evaluation metrics, the experiments' results and the discussion will be included. The paper is concluded with the experiment results and discussion, which aims to give ideas and advice for the passengers in taking subways, as well as provide the possibility of serving as a reference point for the government when planning for future stations and subway lines.

2. Dataset

2.1 Data Collection

We got Hangzhou Metro's data from Tianchi Competition organized by Alibaba. The data set includes data from Jan.1st to Jan.25th in 2019 with 3 rail lines and 80 stations. There are 70 million pieces of data in total. Data fields include the time, the subway line ID, passenger ID, station ID, facility ID, different statuses, and the passengers' pay type.

In this model, we chose station 5 of line B as an example, which contains 1007753 data samples, and only time, line ID, and station ID are used in the model. We use station 69 of line A, station 6 of line B, and station 59 of line C to test the model and compare the results with the original data. Each of them includes 502976, 515388, and 707457 pieces of data.

2.2 Exploratory Data Analysis

After drawing line B, station 5 25 days ridership line chart (Figure 1), it is obvious that almost all the ridership reached the peaks at 8 a.m. and 5 p.m., and it has a clear law for each line. This is because people go to work early in the morning and go home in the afternoon. However, it was different for workdays and rest days because most people need to go to work or school by subway on workdays and would not by subway on rest days.

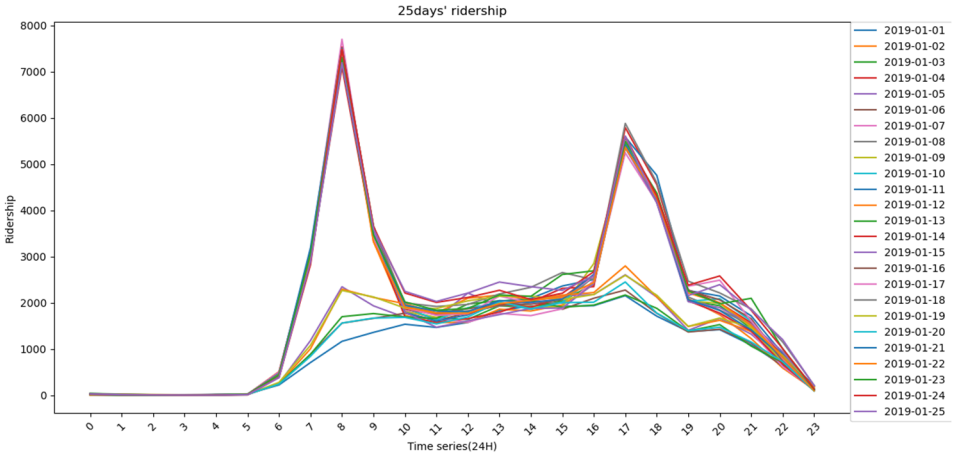

Fig 1. The 25 days ridership of line B, station 5

3. Method

3.1 Linear Regression

In a linear equation in two unknowns, we use y as the dependent variable and x as the independent variable and we have:

$$y = ax + b \tag{1}$$

If we know the specific a and b, we can draw a linear line in the coordinate system. Furthermore, if we try to use an x to predict its y, this is called linear regression. Linear regression is trying to use a line to match the data and find its regular [8].

We usually use the least square method to count the a and b in linear regression. The least square method is used to get the best function that matches the data set by minimizing the number of squares of the errors. The errors e_i between the real value $\hat{y}_i$ and the predicted value y_i [9]:

$$e_i = \hat{y}_i - y_i z \tag{2}$$

$$e_i = \hat{y}_i - (ax_i + b) \tag{3}$$

The sum of squares of the error Q is:

$$\begin{aligned} Q &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \\ &= \sum_{i=1}^{n} [\hat{y}_i - (ax_i + b)]^2 \end{aligned} \tag{4}$$

So the least square method is to minimize the error Q:

$$\begin{cases} \frac{\partial Q}{\partial a} = 2\sum_{i=1}^{n} (\hat{y}_i - b - ax_i) \times (-x_i) = 0 \\ \frac{\partial Q}{\partial b} = 2\sum_{i=1}^{n} (\hat{y}_i - b - ax_i) \times (-1) = 0 \end{cases} \tag{5}$$

And the final results are:

$$\begin{cases} a = \frac{n\sum_{i=1}^{n} x_i\hat{y}_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \hat{y}_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} \hat{y}_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\hat{y}_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \end{cases} \tag{6}$$

After we get a and b, we can use the equation to predict y with a specific x.

3.2 Time Series

Time series is to observe the transformation of the data in chronological order. And in the Time Series for Data Science described, "A time series is a collection of random variables, $\{Xi\}$, indexed on time" [10]. A series of known data are often used to predict future values. Usually, we use time series prediction to predict series that are observed periodically (e.g., per minute, per hour, per day, per week, etc.). Classic cases include precipitation, sales, flow, and so on.

4. Experiments

4.1 Evaluation Metrics

We evaluate the prediction performance using several metrics, including Mean square error (MSE), Mean absolute error (MAE), and R-square. MSE can show us how the predicted values y_i match the true values $\hat{y}_i$, and it is often used as a loss function for regression problems. Compared to MSE, MAE is less sensitive to outliers and more inclusive. R-squared is often used to measure the degree of fitting of linear regressions. If the R-square is closed to 1, it shows the model can fit the true data and the predicted value is more closed to the true value.

4.2 Results and Discussion

Table 1. Evaluation Results

Number	MSE	MAE	R2
1	312205.92	380.23	0.8265
2	370101.29	335.87	0.8420
3	278585.43	353.51	0.8630
4	226051.56	329.71	0.9100
5	170443.54	287.66	0.9313
6	322811.98	387.12	0.8740

We randomly chose 33% data as the test set and 67% data as the training set 6 times (Table 1). Figure 2 shows the results of each model and their real values. The line of prediction almost coincided with the real line. And the best fitting model is the fifth. Its mean square error is 170443.54, the mean absolute error is 287.66 and the R square is 0.9313. The R square is very close to 1 and the model is able to predict the ridership well.

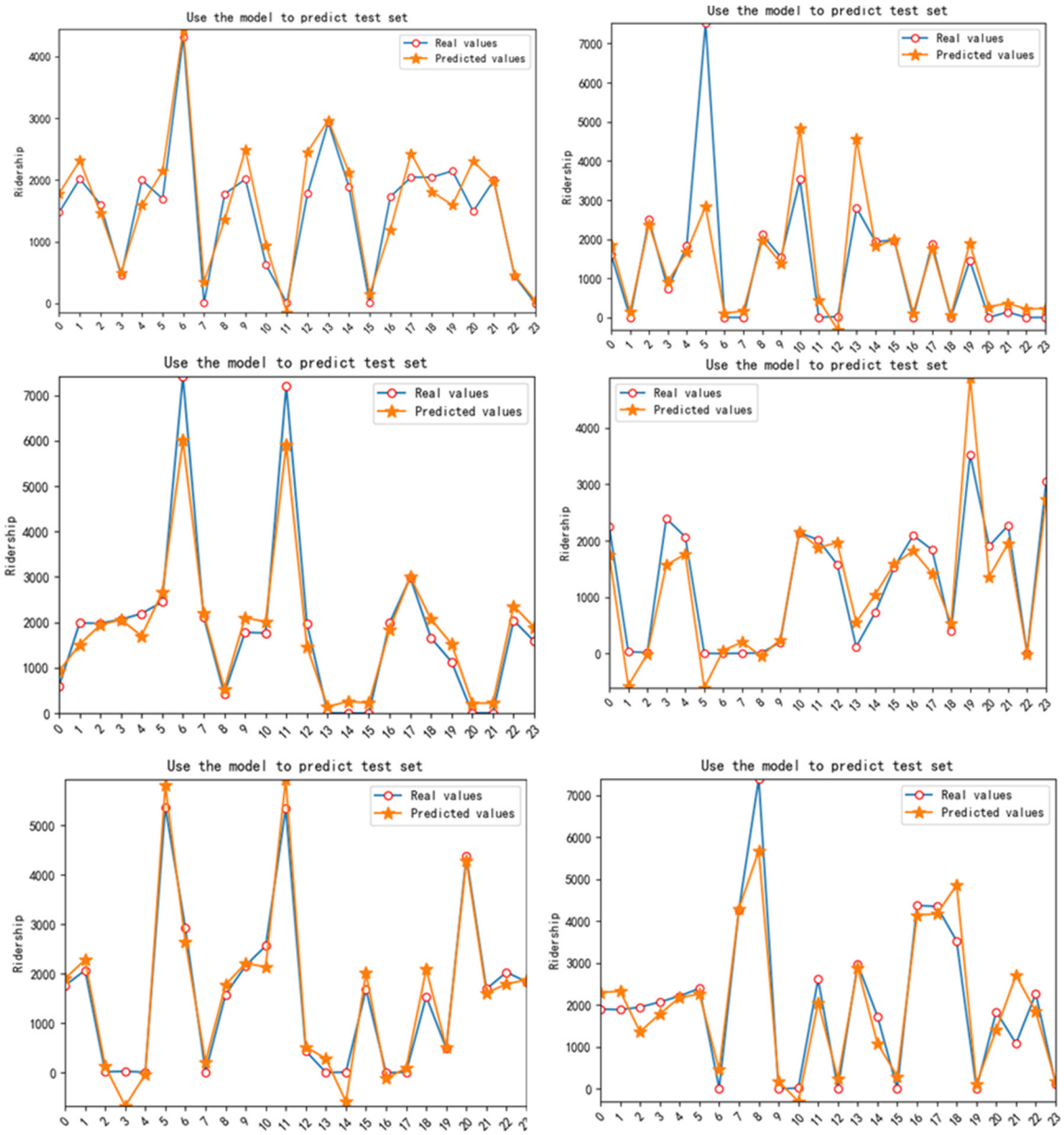

Fig 2. The predicted results for each line and the real values

In Figure 3, every chart on the right shows the actual passenger flow of the station, and every chart on the left shows the prediction of one day's 24 hours ridership using the fifth model and the last 24 hours' ridership. We used 23 days' data to predict the next day's ridership, and here are 23 lines showing the 23 days' ridership prediction. The prediction results are very close to the ground truth data. This also shows the usability and effectiveness of the proposed model.

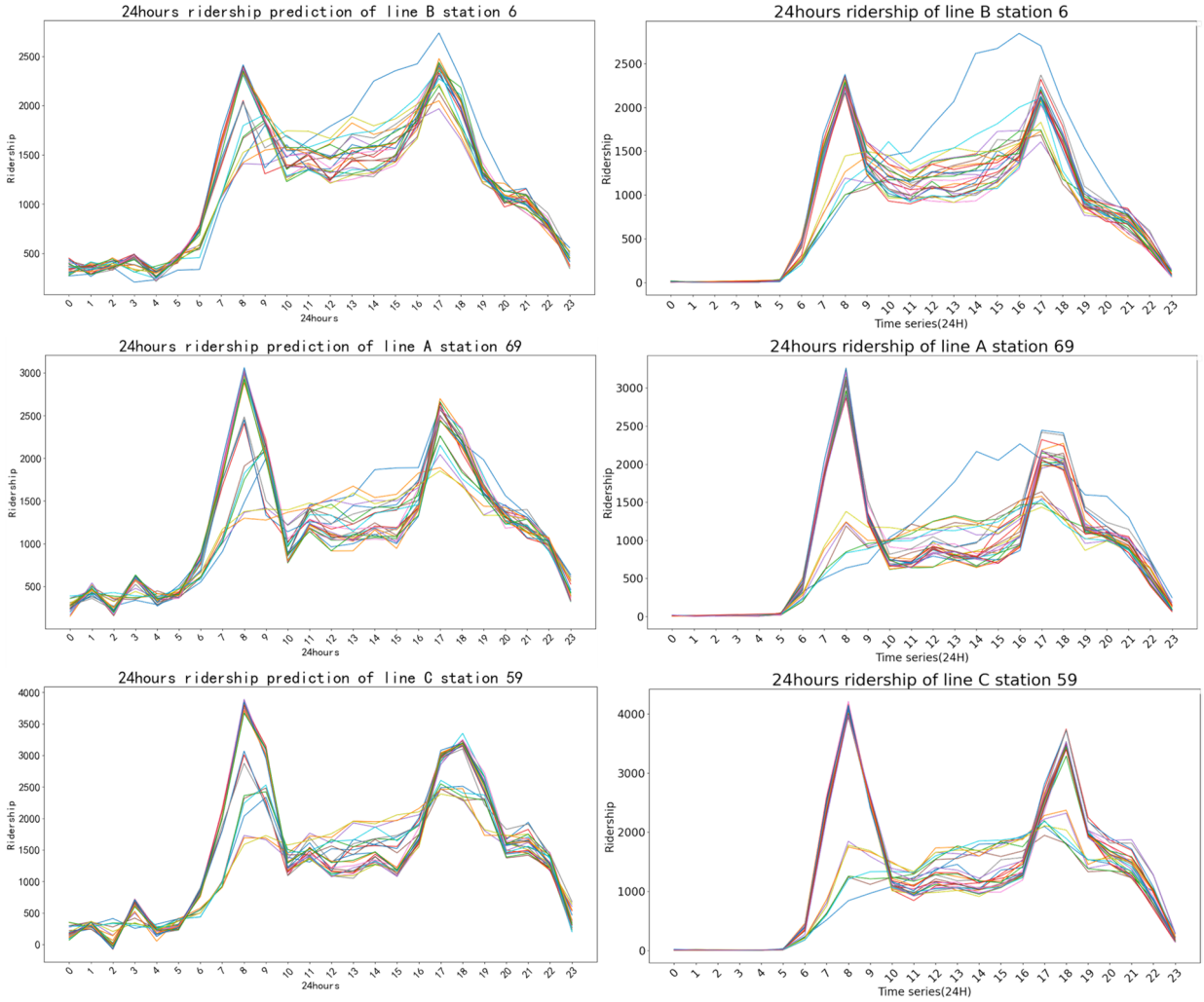

Fig 3. The predicted results for each line and the actual values

5. Conclusion

An accurate and effective prediction model can be beneficial to transportation management, traffic police and citizens. In this paper, I build a model using linear regression and time series to predict and analyze the subway passenger flow in the Hangzhou metro. Through the prediction, we can know that workdays around 8 a.m. and 5 p.m. are the rush hours that have many passengers in the station. And the peak in the morning has higher traffic than in the afternoon. So traffic police need to pay more attention to the two rush hours and people who do not need to commute can avoid these times. And maybe the management can increase the rail line frequency to make people move around the city conveniently and economically.

Many details are being overlooked in this model. It has too few variables that may cause some inaccurate predictions. Many features like weather and the COVID-19 pandemic can be taken into consideration and we can try to find whether different stations are relative to predict the ridership of the whole city in the future study.

References

[1] Pasini, K., Khouadjia, M., Samé, A., Trépanier, M., & Oukhellou, L. (2022). Contextual anomaly detection on time series: a case study of metro ridership analysis. Neural Computing & Applications, 34 (2), 1483–1507. https://doi.org/10.1007/s00521-021-06455-z.

[2] Lu, Y., Ding, H., Ji, S., Sze, N. N., & He, Z. (2021). Dual attentive graph neural network for metro passenger flow prediction. Neural Computing & Applications, 33(20), 13417–13431. https:// doi. org/ 10.1007/s00521-021-05966-z.

[3] Doğan, E. (2021). LSTM training set analysis and clustering model development for short-term traffic flow prediction. Neural Computing & Applications, 33(17), 11175–11188. https:// doi.org/ 10. 1007/ s00521-020-05564-5.

[4] Yujuan Sun, Guanghou Zhang, & Huanhuan Yin. (2014). Passenger Flow Prediction of Subway Transfer Stations Based on Nonparametric Regression Model. Discrete Dynamics in Nature & Society, 1–8. https:// doi. org/10.1155/2014/397154.

[5] Ling, X., Huang, Z., Wang, C., Zhang, F., & Wang, P. (2018). Predicting subway passenger flows under different traffic conditions. PLoS ONE, 13(8), 1–23. https://doi.org/10.1371/journal.pone.0202707.

[6] Lai, Y., & Dzombak, D. A. (2019). Use of Historical Data to Assess Regional Climate Change. Journal of Climate, 32(14), 4299–4320. https://doi.org/10.1175/JCLI-D-18-0630.1.

[7] Kumar, G. (2018). Time Series Analysis of Pm10 for Noida Sector 1 Industrial Area in Ncr Using Multiple Linear Regression. Bulletin of Pure & Applied Sciences-Mathematics, 37E(2), 273–277. https:// doi.org/ 10. 5958/2320-3226.2018.00028.0.

[8] Martin, P. (2022). Linear regression: An introduction to statistical models. SAGE Publications, Limited.

[9] Xu, J. (2021). Design of a Cultural Tourism Passenger Flow Prediction Model in the Yangtze River Delta Based on Regression Analysis. Scientific Programming, 1–9. https://doi.org/10.1155/2021/9913468.

[10] Woodward, W. A., Sadler, B. P., & Robertson, S. (2022). Time series for data science: Analysis and forecasting. CRC Press LLC.