The Usage Mode of Beijing Subway System

Runchen Ye[1, *, †] and Tianrui Huang[2, †]

[1]Beijing International Bilingual Academy, Beijing, China

[2]Beijing National Day School, Beijing, China

*Corresponding author: 2024jye@biba-student.org

Abstract. With the rapid development of engineering in China, the subway network in Beijing has become more and more dense. Studying the pattern of passengers using subway is not only helpful to the travel planning of ordinary citizens, but also of great significance in the future management and planning of the subway system. Previous studies on Beijing metro only focused on individual lines, but this paper analyzes the overall trend of passengers' use of Beijing subway system based on the data of passengers' swipe card using data visualization and linear regression model. This study found that most passengers spend less than 40 minutes on a single subway trip. Peak times for subway use are from 7 a.m. to 9 a.m. and from 5 p.m. to 7 p.m. Passenger's travel time is larger on lines operate in suburban areas with longer distance between each station, shorter on lines with short travel distance, and varied on lines runs in urban area. Whether the subway line the passenger takes passes through the central area; whether the passenger transfers; the total number of stations taken by passengers, these three variables are all positively correlated with the travel time of passengers, among which the transfer variable has the most significant relationship.

Keywords: Subway; passenger flows; data visualization; regression.

1. Introduction

As engineering and technology becoming more and more developed, increasing number infrastructure is being built all around the world. In recent years, public transportation system of big cities worldwide been improved qualitatively, including rail transit and bus lines.

A survey of a subway system is significant, it can help local citizens get more familiar with the subway, so that they can choose appropriate travel routes and avoid lines with heavy passenger flow or peak times of subway rides simultaneously. In addition, understanding the operate scenario of each part of the subway system also beneficial for systematic management. It enables staff to make reasonable decisions and allocate resources effectively. For example, determine which lines need to adjust the departure interval and which stations need to increase the staff according to passenger flows. Besides, realizing the passenger's travel demand in each region can also promote the future construction planning of the subway. To planning new routes in areas with high demand for rides or to alleviate the passenger flow of several existing lines.

With those benefits, a lot of research about modern public transport have been initiated. Meng, Huang and Yin measured the elasticity of residents' choice of subway travel and departure time by using subway card swiping data [1]. Wang summarized the ridership distribution pattern of each bus line in Beijing Guomao area based on the bus IC card data [2]. Lin, Pang and Yu using multiple regression model, examined the influencing factors of Shanghai tourism income base on relevant statistics [3]. He conducted a confirmatory study on the influencing factors of film box office in mainland China based on online movie data, correlation analysis and regression analysis [4]. Wang and Chen analyzed overall passenger flow index, long-term passenger flow time distribution and station passenger flow characteristics according to the passenger flow forecast report of Shenzhen Rail Transit Line 5 [5]. Yu analyzed the IC card swiping data of Shenzhen Metro, extracted the spatial and temporal distribution characteristics of subway passenger flow and user travel characteristics, and studied the changing rules of passenger flow [6]. Duan, Chen et al. analyzed the passenger flow characteristics of Beijing subway line 4 using modeling assignment based on passenger card swiping data [7]. Chen, Huang et al. summarized the passenger flow characteristics and growth pattern of

Beijing subway line 13 through survey data and related questionnaires [8]. Yin, Meng, and Zhang classified 195 subway stations in Beijing by using time-series clustering method based on subway card data and passenger flow characteristics [9]. Lv and Li et al. adopted backward stepwise regression analysis to model and analyze the influencing factors of passenger flow in metro stations in the downtown area of Guangzhou [10].

However, the previous studies about metros are limited to a single subway line, which can only reflect the situation in a local area, no research focus the patterns of passenger flow in the whole subway system. While it is important to conduct detailed investigations of individual lines, it is equally important to study the subway system as a whole. Through the overall survey, we can get the number of passengers in the subway system in different time periods. Through statistics, the function of each subway line in the system can be clearly reflected, which is very useful for managers to see the shortcomings of the subway system. Routes can also be compared so that passengers can make better choices. More importantly, investigating multiple routes, the transfer (from on subway line to another) of passenger can be evaluated so that more meaningful conclusions can be drew.

This paper is about Beijing subway system, which already has 27 subway lines in operation, covering a total length of 783 kilometers, ranking the second in China, with 463 stations and the largest passenger volume in the world. Based on the card swiping data of the passengers, four things are going to be revealed, the general travel time of passengers, the overall passenger flow at different period of times, the travel time of passenger on various lines and the influence of various factors on the duration of passengers' subway ride. They can reflect how most passengers use the subway system, how each line plays a role in the whole system and find the main factors that affect travel time. R language will be used for data visualization, building linear regression model and conduct data testing.

2. Methodology

2.1. The Data Sets

To discover the usage mode of Beijing Subway system, this research is mainly based on a dataset from git-hub which obtained when people swipe their cards as they enter and exit the subway system. This dataset includes the data of passenger's subway card number, the time, which subway line and station passengers enter and exit the subway system. The Travel Time of each passenger can be obtained by their exit time minus enter time. In order to make the study more reflective of the actual situation, lines data provided by Beijing Municipal Commission of Transport is also involved. It includes the number of stations each passenger takes, whether the line they take pass through the central area of Beijing and whether they transfer.

2.2. Data Filtration

Referring to the previous study, data screening is necessary to ensure that the card swiping data for the study came from passengers with commuting as their main purpose [1]. Too much irrelevant data will affect the viewability of data visualization, but will also generate difficulty to fit the model accurately, making the regression model estimator invalid. According to the real situation, the shortest time taken for a one-station subway ride is three minutes, and it takes no more than two hours to travel all the way cross the city by subway. After the calculation of Travel Time by get off time minus get on time of each passenger, if there is negative value, which means get off time is earlier than get on time, or it is greater than 24 hours, this may possibly be an error data. Therefore, once there is travel time that less than 3 minutes or greater than 2 hours, this row of data were removed from the dataset.

2.3. Variables

Since this research aims to explore how Beijing subway system is being used and involves multiple research questions, there are more than one independent variable and dependent variable. There are five independent variables, the certain time period within the day; subway line the passenger has taken; whether the subway line that the passenger takes passes through the central area; whether the passenger transfers; the total number of stations the passenger has taken. Two dependent variables, the travel time of each passenger; the total number of passengers in the subway system at a given time.

2.4. Methods of Inquiry

First, make density plot and histogram with travel time as the x-axis and density as the y-axis to demonstrate the trend of passengers' travel time, travel time of majority passenger and minority passenger. To show passenger flow within each time period. Divide 24 hours in a day into 12 intervals, two hours per segment. Base on the enter time of passengers into the subway system, filter out the number of people entering the subway system at each time period. Then, with time interval as x-axis, number of passengers entered the subway system as the y-axis, sketch a broken line graph [2].

After that, to display the travel time of passengers on each line, select passengers who have not transferred. Then, using box and whisker plot with travel time as x-axis and name of subway lines as y-axis. Finally, make an ANOVA test based on the mean travel time of passengers on each line, to detect the significance of their differences.

Finally, using R language to construct liner regression models [3, 4], with the travel time of each passenger as the dependent variable, conduct with variables including whether the subway line taken by the passenger passes through the central area, whether the passenger transfers, and the total number of stations taken by the passenger as independent variables to explore the influence of these factors on the travel time of the passenger. Finally, use three factors are simultaneously as independent variables to perform multiple regression with travel time to explore whether the three variables together can predict travel time well.

2.5. Hypothesis

Most residential areas in Beijing are located outside the Fourth Ring Road, and it takes about 30 minutes to reach the city center from the fourth Ring Road by subway. Therefore, the first hypothesis would be majority of passengers' travel time is about 30 minutes.

To forecast passenger flow within each time period, referring to the results previous research results [5, 6], there will be two subway travel peaks, one in morning and one in evening. According to the actual life, the second hypothesis estimate to be the passenger flow will have a peak between 8:00 and 10:00 and another peak between 18:00 and 20:00.

According to a previous study on passenger flow characteristics of Beijing subway Line 13 and Line 4, these two lines play different roles in Beijing subway network, so their passenger flow characteristics is different [7, 8, 9]. The third hypothesis is that due to the diversity of subway line in Beijing, it can be assumed that the range of passengers' travel time on every subway line is different.

To estimate the impact of different factors on travel time, based on existing data, there are three factors that can be studied: whether the subway passengers take passes through the central area, whether passengers transfer, and the total number of stations passengers take. Research about influencing factors of passenger flow in Guangzhou central subway station had conclude that the number of commercial buildings around subway stations have a certain impact on passenger flow of the station [10]. It follows that subway lines pass through central area where most of the commercial buildings located in tends to be more popular. Plus, passengers who has transferred or has taken more subway stations indicate they traveled to further places which likely to take longer. Thus, the last hypothesis would be all the three factors has positive relationship with passengers' travel time.

3. Results and Discussion

3.1. Travel Time

Figure 1 is the travel time versus density graph. The travel time ranges from 3 to 120 minutes, while the density maximally achieves around 0.23. The plot of travel time exhibits a unimodal shape, with a skewness to the right. The mean travel time is around thirty-seven minutes. The graph indicates that most people stay on the subway with more than 20 and less than 40 minutes, that's little bit lower compare to the first assumption of 30 minutes roughly.

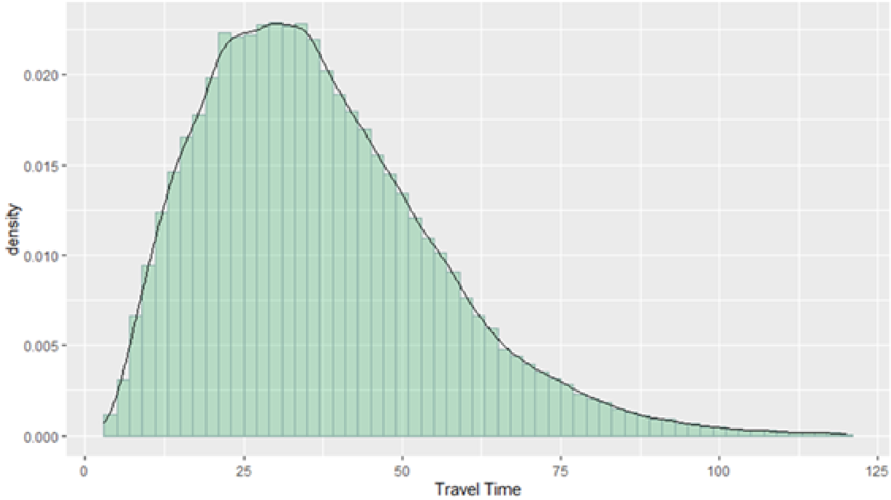

Fig. 1 Density plot of travel time

3.2. Passengers Flow at Different Time Periods

Figure 2 demonstrates the relationship between the time of the day and number of passengers on the subway corresponds with time. The graph is bimodal, which indicates that it has two peaks. The first peak occurs at time between 7 a.m. and 9 a.m. The second peak is roughly at time between 5 p.m. and 7 p.m., the number of passengers starts to increase at 6 a.m. and dramatically reach near six hundred thousand passengers. Until around 11 a.m. the number of passengers declines to about one hundred thousand and stays constant till 3 p.m. And then the number begin to increase again and reach the peak at 6.20 p.m. Finally, the number slowly decreases to zero at 12 a.m. The reason for two peaks to occur is that they represent with the commute time of most of the citizens. Most of the people are on their way to the office or back home at these moments. Compared to the second hypothesis, it is true that there are two peak time of subway ride, but the time of the peak is 7 to 9 a.m. and 5 p.m. to 7 p.m. slightly different from the predicted time of 8 to 10 a.m. and 6 to 8 p.m.

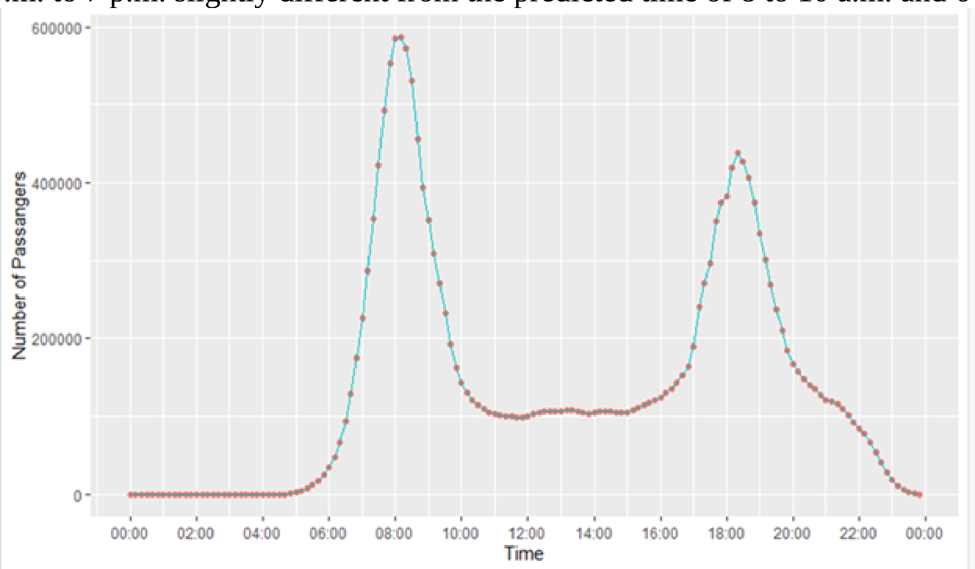

Fig. 2 Scatter plot of passengers vary by time

3.3. Travel Time of Passengers on Each Line

Figure 3 shows the range of travel times for each line. Each line is represented by a number on the x-axis. The range is show by the form of boxplot. The medians of the travel time mainly fall in the range of 25 to 50 minutes. Some lines with relatively short mean travel time like line 2 and line 91,98 are lines with short travel distance. Lines with much larger median travel time like line 92 to 97 are suburban lines with long distances between each stop. The rest of the lines are lines pass through central area, so they have a medium range of travel time, many outliers occur in these lines as well since people in urban area might has distinct travel demand. The distribution of travel time of each line varies which agrees with the third hypothesis. By ANOVA test we have p-value less than 0.05, which means that different subway lines do have variant mean travel time.

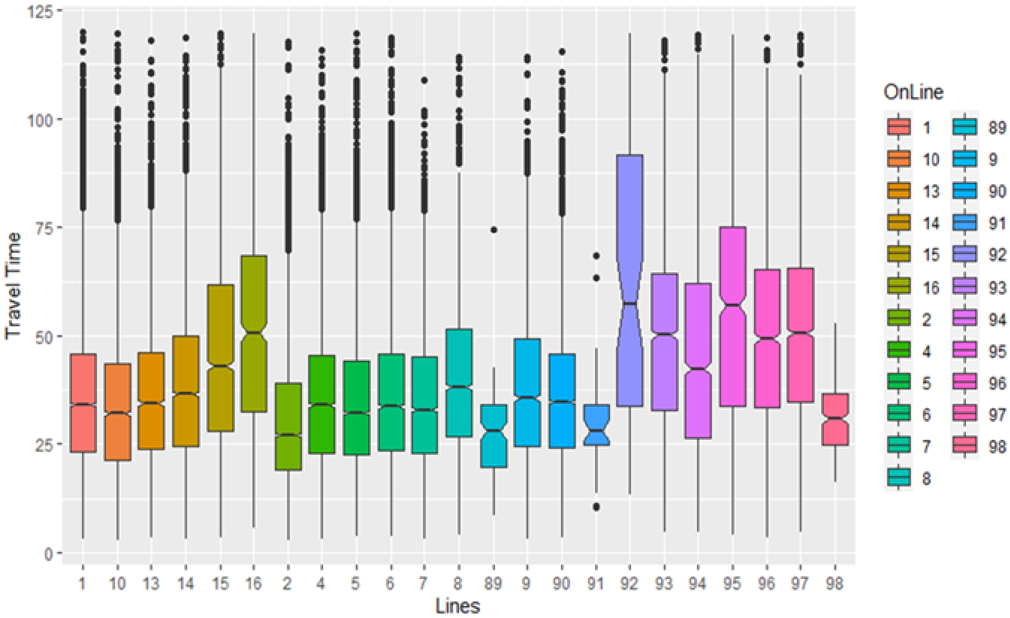

Fig. 3 Boxplot of passenger's travel time vary by lines

3.4. Impact of Different Factors on Travel Time

The first regression of whether the passenger transfers versus travel time shows that transfer variable has a positive influence on travel time. The p-value less than 0.001, which indicates two variables has relationship. Transfer may cause a positive effect on travel time with coefficient 17.3664 which is large. The R-squared of 0.1683 means that transfer variable has a more significant positive correlation with travel time, which agree with the hypothesis.

Table 1. Regression result of Central vs. Travel Time

Coefficients	Std.	Error	t value	P value
(Intercept)	23.893	0.105	226.4	<0.001
Transfer	17. 366	0.122	142.1	<0.001

According to the second regression of whether the line passenger takes pass through central area of the city with travel time, traveling across the central city may cause a negative effect with coefficient -10.5317. The R-squared equal to 0.04779, it follows that transfer variable does not have strong relationship with travel time, not exactly consistent with the previous hypothesis of a positive correlation.

Table 2. Regression result of Transfer vs. Travel Time

Coefficients	Std.	Error	t value	P value
(Intercept)	45.506	0.134	337.25	<0.001
Transfer	-10.531	0.148	-70.77	<0.001

The third regression estimated the relationship between the total number of station passenger took and his travel time shows that transfer variable also has a positive influence on travel time. The p-value is still less than 0.001, which demonstrate stations variable does has impact on travel time. The R-squared equals 0.02347, in other words, transfer variable does have a positive relationship but not significant, this does not strongly support the previous hypothesis.

Table 3. Regression Result of Stations vs. Travel Time

Coefficients	Std.	Error	t value	P value
(Intercept)	49.504	0.263	188.03	<0.001
Transfer	-0. 617	0.012	-49.26	<0.001

The last regression is a multiple regression predicting the impact of three variables on travel time simultaneously. These three variables have an impact on travel time because the P-value is less than 0.001. R-squared is equal to 0.2094, which indicates that the three variables together have a great impact on travel time.

Table 4. Multiple Regression of Three Variables vs. Travel time

Coefficients	Std.	Error	t value	P value
(Intercept)	32.224	0.262	122.705	<0.001
Transfer	17.030	0.119	142.704	<0.001
Central	-9.751	0.177	-54.85	<0.001
Stations	-0.003	0.014	-0.211	0.833

3.5. Cause and Limitation Analysis

For the estimation of passengers' travel time, the actual mean travel time is slightly higher than the expectation of 30 minutes. That's probably because the travel time of suburban lines isn't take into account. The travel time of suburban lines is usually larger since the suburban lines travel through long distances while have fewer stops.

In terms of the forecast of passenger flow peak, although the morning and evening peak period obtained from the actual data is one hour earlier than expected, it is assumed that the basis is accurate in the general direction. There are two peak periods for subway travel, namely in the morning and in the evening. The reason for this accurate prediction is that considering the main function of the subway is for commuters to commute to work in the morning and home in the evening, there must be two periods of peak time.

As for the prediction of passengers' travel time on different metro lines, as assumed, different lines have different range of travel time. This is because different lines in the subway network have different functions, and the time for passengers to take the lines with different functions is bound to be different.

In terms of the prediction of influencing factors of travel time, transfer variable can indeed affect travel time, which is because passengers who need to transfer tend to have farther destinations. However, the other two variables had no significant effect. The downtown line is popular because there are more commercial buildings in the city center, but that doesn't necessarily mean passengers will spend longer on it. Passenger demand in the urban area is variable, some people may take the subway just to get from their homes to the nearest shopping mall. Suburban routes, on the other hand, can even have longer travel times because the purpose of most people use them is to get home or away from home. As for why station variable is not significantly related to travel time, it may be due to the different station distances of subway lines. Since there are more developed zones, the station distance of the urban line might be smaller. In the suburbs, there may still be some areas that have not been exploited, so the distance between subway stations might be larger. Taking one stop on a suburban line may take as long as two stops on an urban line.

Errors might occur due to the simple analyzation of single variable, as well as the limited amount of data and inaccurate data. These might as well contribute to the possible errors occurred in our analyzation. We simply analyzed the data within one day. Therefore, the trend generated can only be a generalization of weekdays. Also, the data collected might contain some data that were inappropriately gathered, which might cause a deviation from the actual value and the value generated.

When studying the influencing factors of travel time, scatter plot would be better for showing the visual information. But considering this particular situation, more than 90k points on y-axis by a scatter plot and less than 30 numbers of value on x-axis makes the plot only multiple vertical lines. Even though we have the regression line, it cannot represent anything clearly in the plot.

4. Conclusion

Compare to the result and the hypothesis, the conclusions are the follows. The mean travel time of all the passengers is 37 minutes, which is slightly higher than estimation of 30 minutes. The estimation of passenger flow of having two peaks around 8 a.m. and 6 p.m. corresponds with the actual data gathered. The distribution of travel time of each line varies, which also agrees with the anticipation. The regression reveals the fact that whether the passenger transfers have the most significant impact on travel time, the other two variables, whether the line passes through central region and the number of stations passenger takes, theirs relationship with travel time is not severe. The result generated, including the 8 a.m. and 6 p.m. peaks occurred at Time versus Number of Passengers plot and the mean travel time of 37 minutes generated by travel time plot, can be used in estimating the number of passengers and improving the service and management of each subway station. Those can be helpful in the management and arrangement of Beijing subway service. The passengers flow diagram enables Beijing subway service company to put in more resources like number of subway trains and service stuff during the time when number of passengers is the largest. The distribution of travel time for each line allows Beijing subway service company to know the stations that are the most vital and also put in more resources in maintaining and improving services. In the future, more sets of data about different aspects on subway lines can be analyzed. For instance, the passenger flow of each station of the same line. Also, various methods can be used during analyzation but not only use R. In sum, the research still has limitations and can be improved through using various methods and consider more variables.

References

[1] Meng Bin, Huang Song, Yin Qin. Spatial and temporal Distribution characteristics of Metro travel departure time elasticity in Beijing. Journal of Geo-Information Science, 2019, 111-121.

[2] Wang Shuang. Analysis and Countermeasures of Bus Passenger Flow Characteristics in Guomao District of Beijing. Highway Transportation Technology, 2019, 331-334.

[3] Lin Xia, Pang You, Yu Haibin. A Study on the Influencing Factors of Tourism Income in Shanghai Based on Multiple Regression Analysis. Journal of Wuxi College of Commerce and Vocational Technology, 2017.

[4] He Shuangnan. An Empirical Study on the Influencing Factors of Film Box Office in Mainland China. Film Literature, 2017,6-10.

[5] Wang Qingyu, Chen Lei. Analysis of Passenger Flow Characteristics of Shenzhen Rail Transit Line 5. Transportation Technology and Economy, 2015, 77-81.

[6] Yu Zijuan. Analysis of Shenzhen Metro Passenger Flow Characteristics Based on IC Card Data. Comprehensive Transportation, 2019, 119-123+130.

[7] Duan Weijing, Chen Yanyan, Lai Jianhui.Analysis of Passenger Flow Characteristics of Beijing Metro Line 4. Urban Express Rail Transit, 2013, 54-57.

[8] Chen Tenglin, Huang Mingfang, Wu Xiongbin. Analysis of Passenger Flow Characteristics of Beijing Rail Transit Line 13 Station. Journal of Heilongjiang Institute of Technology, 2013, 40-44.

[9] Yin Qin, Meng Bin, Zhang Liying. Identification of subway station types in Beijing based on passenger flow characteristics. Progress in Geography,2016, 128-136.

[10] Lv Dijiang, Li Shaoying, Tan Zhangzhi, Wu Zhifeng, Gao Feng, Liu Xiaoping. Detailed modeling of passenger flow factors in Metro stations: A case study of Guangzhou City. Geography and Geo-Information Science, 2019, 64-71.