Research on Scheduling Based on Spatial and Temporal Distribution Characteristics of Piled Shared Bicycle

Jing Ge*, Yi Zhang
Shanghai Martime University, Shanghai, 201306, China

*Corresponding author e-mail: 202030610094@stu.shmtu.edu.cn

Abstract. With the development of bicycles, the problem of having no piles to park and no cars to use also appears. Because the driving mechanism of bicycle travel is not understood, the operation personnel in charge of redistribution have no way to start. In this paper, from the perspective of subway station and city bus transfer system, starting from the characteristics of road network, combining with the nature of land use, the spatial-temporal characteristics of quantitative analysis is carried out to study the spatial-temporal distribution law of travel behavior. Based on the above factors and the supporting conditions of the bicycle facilities, the common least square model was established. It was found that the number of stations in the traffic district was the most important variable, indicating that the basic supporting facilities had a great impact on the number of bicycle trips. Then, the number of bicycles and piles was deeply studied. Firstly, the stations are classified according to K-means, and defined as low-frequency high flow, high-frequency low flow, low-frequency low flow, and medium-frequency low flow according to the average of clustering results. Then, the redistribution suggestions are put forward respectively for these categories. For outflow clustering with large values of O and D, R can be predicted by BP neural network to provide reference for redistribution and distribution.

Key words: Shared bike; Scheduling; Clustering.

1. Introduction
As a product of the sharing economy, bicycles are becoming increasingly popular, but demand is uneven across the country, increasing the workload of dispatching. At present, there are a lot of literatures on the unbalanced problem of bicycle stations. Zupeng L [1] et al. solved the unreasonable number of bicycles allocated at each station and the untimely scheduling of bicycles between stations. Based on the improved ant colony algorithm, they solved the problem of how to optimize the scheduling between stations of the public bicycle system. Chen PC [2] et al. established a system based on five different structures of the complex neural network, which can predict the rental and return demand of each bicycle station in real time, and can be used to develop load balancing strategies between stations and effectively solve the bicycle shortage problem in some places. Liu J [3] et al. predicted the demand and balance of bike-sharing stations based on the prediction model of artificial neural network, and built an optimization model based on genetic algorithm to maximize the use of platform optimization and reduce the number of unbalanced stations. Jian Jiang [4] et al. proposed a deep learning model of destination
prediction network based on spatial data to predict the most likely destination for each user and provide convenient and flexible services for users. Mette S [5] et al. conducted a case study on the campus of Gaziantep University, and optimized the site by maximal covering mathematical models, the mathematical models of P-center and P-median, and undesirable facility location model, thus providing different solutions for campus planners to optimize the bike-sharing site. Xu H [6] et al. solved the lag problem of traditional bike sharing scheduling methods and proposed a dynamic scheduling model based on short-term demand prediction. Through scheduling test on one year's data of Chicago public bike sharing system, the performance of the model was analyzed. Most of these studies focus on the optimization analysis of the site itself, while few scholars have studied the optimization of bike-sharing sites based on the travel mechanism and cycling characteristics of users. This paper starts with quantitative research, determines significant variables, and controls the spatial allocation of bicycle resources in place.

2. Data collection and preprocessing
Because of the openness and accessibility of data, this article uses Washington, D.C. as an example.

2.1. Bicycle data
1) Trip history data for 5 working days from September 9, 2019 to September 15, 2019 in Washington, D.C. (https://www.capitalbikeshare.com/system-data)
2) Bicycle data obtained by SQL per minute crawler during 5 working days from January 12, 2021 to January 16, 2021. (https://secure.capitalbikeshare.com/map) to process the data and the crawler, with navicat database management tools, connect MySQL, data processing, time interval for 30 minutes, the average every 30 consecutive one minute, Obtaining the half-hour per-station average of 148,505 rows of data for 596 stations.
3) The positions of 596 bikes obtained by SQL crawler include: ID, site ID, site name, longitude and latitude. (https://secure.capitalbikeshare.com/map)

2.2. Characteristic data
1) Land use: the characteristic variables related to the built environment within 500 m of the site were selected and analyzed: residential land area, park green land area, office land area, commercial land area, university land area and population density
2) Number of rail transit hubs: bus stations, subway stations and tram stations
3) Passenger flow of subway stations: statistical data of passenger flow of subway stations on weekdays and during morning and evening peak hour

3. Model construction

| variable                | Moran’s I | z       | p       | model         |
|-------------------------|-----------|---------|---------|---------------|
| residential land        | 0.388733  | 13.778926 | 0       | agglomeration |
| park green space        | 0.231175  | 8.462102  | 0       | agglomeration |
| office space            | 0.271410  | 9.680286  | 0       | agglomeration |
| commercial land         | 0.191162  | 6.901996  | 0       | agglomeration |
| colleges and universities land | 0.322825 | 12.0397748 | 0 | agglomeration |
| population density      | 0.543141  | 19.091853 | 0       | agglomeration |
| bus station             | 0.322895  | 11.360207 | 0       | agglomeration |
| subway station          | -0.008917 | -0.224015 | 0.822746 | divergent     |
| tram stop               | 0.337190  | 12.843748 | 0       | agglomeration |
| number of intersections  | 0.442622  | 15.992122 | 0       | agglomeration |
| number of other sites    | 0.168808  | 5.986712  | 0       | agglomeration |

Before the regression analysis, the correlation analysis of variables was carried out and it was found that no threshold value exceeded 0.7 and no variables needed to be removed. The spatial autocorrelation test
is carried out on all regional variables to check whether there is any influence on each surface space. Since it is the calculation of surface elements, all elements should be standardized to reduce the deviation caused by different number of adjacent elements. As shown in Table 1, according to the Moran index, the subway station is not significant, and according to the Z-value, except for the subway station, all of them are also clustered, resulting in spatial heterogeneity.

Removal of subway station this variable do ordinary least squares method, the dependent variable is weekday traffic area in several sites, trip generation combined results such as table 2, check the multicollinearity, multicollinearity is due to the precise or highly relevant relationship between the variables, cause the distortion of the linear regression model, the variance in the parameter estimate increase, $1/(1-R^2)$ is the Variance Inflation Factor (VIF). The larger the Variance Inflation Factor is, the stronger the collinearity is. Check that VIF is not greater than 7.5 and there is no multicollinearity.

| Table 2. OLS inspection results |
|-------------------------------|
| **variable**                  | **coefficient** | **T** | **VIF** |
| residential land             | 42.54*          | 2.35  | 1.29    |
| park green space             | -31.21          | -1.41 | 1.13    |
| office space                 | 55.85*          | 3.20  | 1.27    |
| commercial land              | 15.44           | 0.66  | 1.15    |
| colleges and universities land| 16.07           | 0.57  | 1.19    |
| population density           | 21.65           | 1.99  | 1.22    |
| bus station                  | -48.99*         | -3.10 | 1.13    |
| tram stop                    | 4.71            | 0.15  | 1.06    |
| number of intersections      | 130.75*         | 4.43  | 1.70    |
| metro passenger flow         | 3.05            | 0.16  | 1.02    |
| number of other sites        | 354.63*         | 21.16 | 1.19    |

R$^2$ is 0.65, and the influence of other stations is the most significant. In a traffic district, the more other stations, the more available bicycles and empty piles. In other words, the more potential bicycles, the more they are used. Piles of the bicycle, poor flexibility, parking pile planning is not reasonable, when you want to use the car, the station is empty; When I returned the car, I couldn't find the parking post. It can be seen that the imbalance between supply and demand in peak periods ignores the respective characteristics of the stations, leading to the deviation of the relationship between traffic and characteristics at stations with different attributes. The following is an in-depth study on this point.

4. Cluster analysis

Net flow rate according to the average daily amount of O and D and P three characteristic values for different site classification first, according to different classification allocated to improve the efficiency of the operation personnel, the daily average 0, D reflect site bicycle use frequency, net flow rate $P = (D_i - O_i)/OD_i$, plus or minus said site bike mainly inflows/outflows, more close to 1 into more obvious, The closer you get to -1, the more obvious the outflow is.

As an unsupervised classification, K-means clustering can effectively explore the internal structural features of the data set and classify things according to their similarity, so that the travel features of the same cluster have the same identity. Therefore, K-means is selected in this paper to classify sites. Through the hierarchical clustering method of tree graph, the number of sets can be found dynamically in the hierarchical clustering. Therefore, the 16 sites that can achieve self-balance (P=0) are removed to conduct hierarchical clustering for inflow and outflow respectively.

For the net inflow, the clustering number was determined to be 4, and then K-means clustering was performed. The number of clustering was 23, 17, 40 and 58, respectively, and the contour coefficient was 60.8%. And according to the working day average O and D values and the net outflow rate, it can be summarized as Cluster 1 (low frequency and high inflow), Cluster 2 (high frequency and low inflow), Cluster 3 (low frequency and low inflow) and Cluster 4 (medium frequency and low inflow).
Fig. 1 Hierarchical clustering results of net inflow and net outflow

Table 3. Net inflow clustering results

| Clustering category                  | Working day average O value | Working day average D value | Net outflow rate |
|-------------------------------------|----------------------------|----------------------------|-----------------|
| Cluster 1 (low frequency and high inflow) | 20.94347826                | 34.83623188                | 0.231557596     |
| Cluster 2 (high frequency low inflow) | 91.50588235                | 101.7294118                | 0.056119504     |
| Cluster 3 (low frequency and low inflow) | 16.95875                   | 18.98833333                | 0.05933507      |
| Cluster 4 (medium frequency low inflow) | 46.76842105                | 51.75087719                | 0.048784112     |

The site of cluster 1 is not suitable for centralized scheduling due to its low frequency, so scheduling fees can be charged for the bikes placed in the cluster site to guide people to put the bikes in other sites. Clustering 2 and 4 require timely cleaning of bikes, which can be dispatched to other stations at night to avoid not being able to transfer bikes in time the next day, resulting in no piles available.

The number of the four clusters with net outflow is 45, 23, 17, 54, and the contour coefficient is 78.6%. According to the working day average O and D values and the net outflow rate, the clusters are generally classified as Cluster 1 (low outflow at medium frequency), Cluster 2 (high outflow at low frequency), Cluster 3 (low outflow at high frequency), and Cluster 4 (low outflow at low frequency).

Table 4. Net outflow clustering results

| Clustering category                  | Working day average O value | Working day average D value | Net outflow rate |
|-------------------------------------|----------------------------|----------------------------|-----------------|
| Cluster 1 (medium frequency low outflow) | 49.28                      | 41.58857143                | 0.0833687       |
| Cluster 2 (low frequency and high outflow) | 87.875                     | 79.6125                    | 0.044659378     |
| Cluster 3 (high frequency and low outflow) | 13.32541667                | 6.97375                    | 0.27231638      |
| Cluster 4 (low frequency and low outflow) | 23.10520833                | 19.47291667                | 0.082769855     |

Cluster 2: Due to the high net outflow rate, there will be a serious shortage of vehicles, so vehicles should be replenished in time to avoid no available vehicles. Incoming Cluster 3 and Outgoing Cluster 4 both belong to low frequency and low net flow rate, and the travel volume is not too large compared with other clustering sites, which can achieve its own balance, so it is not considered for the time being.

Outflow clusters with high values of 1, 3, O, and D can be centralized scheduling and timely supplement. Unlike scheduling at night, dynamic scheduling in the daytime is carried out in the peak
period with high demand for bicycles, which requires the support of prediction. Two stations were randomly selected for observation, Columbia & Ontario Rd NW and Massachusetts Ave & Duupont Circle NW. In order to ensure that the number of bikes and the number of empty piles at each station were regularly restored to the predefined target value M, introduce a response at the same time the two factors of variables: every half hour site cycling rate \( R \), \( R = \frac{\text{site time available bike number}}{\text{always pile number}} \), if it is a rising trend, represents a car also in, if it is a downward trend, represents a bicycle to lend, the closer the 0 bike is less, the more the more close to 1 bike, can reflect the site car capacity.

![Fig. 2 Prediction results of BP neural network](image)

According to the BP neural network prediction as shown in the figure, there is only a small difference between the predicted value and the real value. It is found that Columbia & Ontario RD NW is a morning destination type, which refers to arriving in the morning and leaving in the evening, with some leaving and returning in the middle. Please pay attention to supplement before 7 o’clock. Massachusetts Ave & Duupont Circle NW Night Destination, which generally starts down in the morning, reaches maximum lending at 18:30 and gradually increases, and their demand is significantly greater than supply. Please be sure to replenish your bike before 18:30. For the scheduling work with peak usage, the approximate number of dispatching cycles can be determined by dispatching number \( X = (\text{predicted } R + \text{target value } M) \times \text{the total number of piles at station } I N \).

5. Summarizes

Overreliance on cars will cause urban congestion, and the popularity of bicycles can reduce the pressure on the traffic network. By mining spatial-temporal data and analyzing the kernel density of peak period occurrence and attraction, variables that may have an impact on the travel volume are determined. It is found that the number of other stations in the traffic district is the most important characteristic by the establishment of ordinary least square method, and this characteristic is studied in depth.

Firstly, K-means is used to classify sites according to the average O and D values of working days and net outflow rate, and then redistribute them, which will greatly improve efficiency and better provide management and operation support for enterprises and governments. On the whole, it is defined as low frequency and high flow, high frequency and low flow, low frequency and low medium frequency through the average of clustering results. And put forward the opinion of redistribution to these several kinds respectively. Due to the high \medium frequency flow, O, D value can be larger centralized traffic control, and puts forward concrete practice about peak hours of scheduling, involving bicycles and pile figures of these two variables, thus introducing every half hour site cycling rate \( R \) variable, \( R \) is predicted by BP neural network can not only see the change trend, can also according to the characters of travel schedule in time and the number of bicycle, Provides effective information for scheduling and allocation.

References

[1] Zupeng L, Keping L, Xiaohong Z. Optimal dispatch between stations for public bicycle based on ant colony algorithm [J]. Journal of Transport Information and Safety, 2012, 4: 71-74.

[2] Chen P C, Hsieh H Y, Su K W, et al. Predicting station level demand in a bike-sharing system using recurrent neural networks [J]. IET Intelligent Transport Systems, 2020, 14(6): 554-561.
[3] Liu J, Li Q, Qu M, et al. Station site optimization in bike sharing systems[C]//2015 IEEE International Conference on Data Mining. IEEE, 2015: 883-888.

[4] Jian Jiang, Fei Lin, Jin Fan, Hang Lv, Jia Wu, Jianxin Li. A Destination Prediction Network Based on Spatiotemporal Data for Bike-Sharing [J]. Complexity, 2019, 2019.

[5] Mete S, Cil Z A, Özceylan E. Location and coverage analysis of bike-sharing stations in university campus [J]. Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy, 2018, 9(2): 80-95.

[6] Xu H, Duan F, Pu P. Dynamic bicycle scheduling problem based on short-term demand prediction [J]. Applied Intelligence, 2019, 49(5): 1968-1981.