ISSN: 2664-9640

# Empirical Research on Data Mining Based on Public Bicycle System

Jianglong Shen [1,2,3,4,5], Na Lei [1,2,3,4,5], Yufei Xiong [1,2,3,4,5], Xueying Wu [1,2,3,4,5]

[1]Shaanxi Provincial Land Engineering Construction Group Co.,Ltd., Xi'an 710075, China;

[2]Key Laboratory of Degraded and Unused Land Consolidation Engineering, the Ministry of Natural Resources, Xi'an 710021, China;

[3]Institute of Land Engineering and Technology, Shaanxi Provincial Land Engineering Construction Group Co.,Ltd., Xi'an 710021, China;

[4]Shaanxi Provincial Land Consolidation Engineering Technology Research Center, Xi'an 710075, China;

[5]Land Engineering Technology Innovation Center, Ministry of Natural Resources

## Abstract

**Massive data generated from Bicycle Sharing Systems were used to analyze the spatial and temporal characteristics of user behavior and plan stations setting. Because the existing station clustering indicators depend on static data of the station, they are insensitive to the actual station failures and affect the clustering effect. Therefore, we proposed a new clustering index: Normalized Net Bicycle Outflow, which used users ride datasets to dynamically describe the use status of the station, eliminated errors caused by faulty bikes and parking piles, and can reflect the actual use of the station. In this paper, we summarized the research progress in recent years from two aspects: stations clustering analysis and bicycle rental demand forecast. In addition, the real data of CitiBike system are used for stations clustering and prediction. The results show that the new clustering index can accurately capture the time-varying traffic flow of the station, distinguish the different states of the station on weekdays and weekends and retain the tidal effect of the station on weekdays, and Elman neural network can predict the bicycle demand of the station well.**

## Keywords

**BicycleSharing Systems; Data Mining; Cluster Analysis; Normalization Bicycles Net Outflow; Elman neural network.**

## 1. Introduction

In recent years, with the rapid development of social economy and the continuous improvement of people's material living standards, the demand for travel has continued to increase. Traffic congestion and air pollution in the city have plagued people's daily life. The public bicycle system is green, convenient, and inexpensive. Such characteristics have been welcomed by many cities. The global popularity of public bicycle systems has contributed to energy conservation and emission reduction, enhancement of citizens' fitness, enrichment of urban transportation types, optimization of road resource construction, enhancement of public transportation operation efficiency, and enhancement of urban image. Many contributions. The massive data generated by the use of public bicycles provides data support for the study of human riding behavior, travel preferences, and optimization of site layout. Research based on public bicycle data mining continues to develop

ISSN: 2664-9640

Predecessors' research on public bicycle data mining mainly focused on site clustering and bicycle demand forecasting. In 2009, Froehlich et al. [1] first used the number of site piles and the number of bicycles to build a Normalization Available Bicycles(NAB) as a clustering indicator. However, it is foreseeable that the site will face many complex situations in actual situations, such as being affected by construction, or malicious damage to parking spots, failure of public bicycles, etc. Problems will lead to errors in the calculation of NAB indicators. This article hopes to avoid errors caused by these problems by using real cycling data to construct clustering indicators, and use actual data to simulate the characteristics of the new indicators.

## 2. Cluster analysis of BSS sites

The cluster analysis of public bicycles can discover the usage patterns of public bicycles, reveal the basic temporal and spatial dynamics of the city, reflect the urban culture and spatial layout, and provide references for the design and operation of the public bicycle system. The research work of the cluster analysis of BSS sites can be divided into two Type: One type is to expand around the time series data of the rental point, through the construction of clustering indicators, select the clustering model to realize the cluster analysis of the site. The other type is based on the geographical distribution of the site, the location is similar, the use mode The same sites are classified into one category. Through the above two types of methods, analyze the usage patterns of the sites and understand the temporal and spatial properties of the sites.

### 2.1. Clustering based on site timing characteristics

The existing clustering methods based on site temporal characteristics are divided into the following three categories: hierarchical clustering, partitioned clustering, and model-based clustering. Hierarchical clustering first calculates the distance between samples, and merges the closest points into The same class, and then calculate the distance between the class and the class, the closest class is merged into one big class. It keeps on merging until it merges into one class. Its advantage is that there is no need to set the number of clusters in advance, but the disadvantage is that Outliers have great influence and high computational complexity. Partitioned clustering divides the initial N points into pre-set K clusters, so that the result is satisfied: the points within the class are close enough, and the points within the class are close enough. The points are far enough. The K-means algorithm is a typical partitioned clustering method. Its advantages are simple and easy to implement, and low time complexity; the disadvantage is that it is sensitive to noise and outliers. The EM algorithm is based on model clustering Method. It is achieved by alternating the two steps of calculating expectations and maximizing. EM algorithm is an effective tool for data with missing information.

2.1.1 Hierarchical clustering

In 2009, Froehlich et al. [1] obtained 26 million observations from the Bicing system in Barcelona, Spain from 2008/08/27 to 2008/12/01, constructed NAB as a clustering index, and selected hierarchical clustering to achieve cluster analysis of sites The clustering results show that nearby sites show similar usage patterns, and different types of sites depend on geographic relationships. In 2012, Lathia et al. [2] captured London Barclays BSS 2010/10/17-2010/12/03 A total of 11,440,440 observations, a total of 12,255,057 observations from 2011/01/02 to 2011/02/22 were used as the control group to compare the differences in NAB indicators between the two data sets on weekends and during the week, and the hierarchical clustering algorithm was used to classify the sites. For 6 categories, the impact of the London bicycle system on leisure users during the switching access strategy was studied. In 2017, Lin Yanping [3] used the New York CityBike and Washington SmartBikeDC systems 2014/04/01-2014/05/30, a total of 2318139 entries The data is used as a sample, and the hierarchical clustering algorithm is used to cluster the BSS sites in New York and Washington, and 23 and 27 site clusters are obtained respectively.

ISSN: 2664-9640

In 2018, Wu et al. [4] obtained data sets from Longgang and Luohu periods in Shenzhen for comparative analysis. The first period contains data sets from 2016/06/09-2016/10/19; the second period contains 2016/10/20-2016/12/21 data set. It is proposed to use the NAB value of the day minus the NAB average of all days to construct a normalized NAB analysis site using attributes. It will have the same NAB before clustering The value of the site is deleted, and the hierarchical clustering algorithm with dynamic time warp distance as the feature is used to cluster the average daily NNAB time series vector of the site, and the working days and non-working days in Luohu and Longgang are respectively NNAB conducted a cluster analysis, and the results showed that: on weekdays, Luohu sites can be divided into four categories based on usage. Clusters 1 and 2 have a lot of bicycles returning to these sites at around 8 o'clock in the morning and are relatively stable until 5 o'clock in the afternoon. Around midnight, a large number of bicycles left these sites at 6 o'clock. Therefore, clusters 1 and 2 are defined as morning destinations and night origin sites. The situation of cluster 4 is just the opposite of clusters 1 and 2, reflecting that cluster 4 belongs to morning origin , Night destination site. Combined with POI data analysis, it is found that there are three main central business districts near the cluster 1 and 2 sites, and the cluster 4 sites are mainly located in the residential building area. The validity of the analysis results is verified. The clustering results Spatial distribution can find sites with similar usage activity patterns, and there are geographic connections.

2.1.2 Partitioned clustering

In 2009, Borgnat et al.[5] obtained all trajectory information of Paris BSS 2005/05/25-2007/12/12, used K-means clustering algorithm to cluster the bicycle flow between stations, and divided all travel trajectories into There are four types of Sunday noon, 6 pm, 9 am, and noon time. The corresponding stations at noon on Sunday are parks and other types of places. The trajectory at 6 pm and 9 am is the reversal of work, and the corresponding station is the railway station. , Campus and other places, the trajectory at noon is for lunch break cycling, and the corresponding sites are communities and other places. In 2017, Feng et al. [6] used NAB as a clustering index and introduced different cluster number evaluation indicators to compare K-Means clustering and hierarchical clustering to divide the effect of different clusters. At the same time, based on the K-means clustering algorithm, the sites of the Paris Velib system are divided into four types: work sites, residential sites, sites that are in short supply, and sites that exceed demand. In 2018, Shen Xingfa et al. [7] used the Tyson polygon algorithm to divide the urban area based on the location information of the station, used the potential Dirichlet distribution model to mine the functional characteristics of the station, and used the K-means algorithm to cluster the station, and finally combined Point of interest data, analyze the temporal and spatial attributes and usage functions of the site.

2.1.3 Model clustering

In 2011, Vogel et al. [8] extracted the public bicycle rental and return amount per hour at each site on weekends and weekdays, except for the total number of rental and return vehicles on the day of the corresponding site, and each site got the normalized 48 attribute values as features. , Use the EM classification algorithm to divide all sites into 5 categories, and analyze the possible attributes and locations of different categories of sites according to the bicycle usage patterns of different categories. In 2014, Etienne et al. [9] took the Vélib system in Paris as an example , Obtained 2.5 million trajectory data of 1185 stations in the system for a month. It is proposed to construct the time series data of the stations using the index generated by the number of arrivals of public bicycles and the number of departures. The Poisson mixture model and the EM algorithm are used to divide 1185 stations into 8 types , Clusters can be well connected with parks, railway stations and other types of facilities, as well as social variables such as population, work, and service density. In 2018, Zhao Hong et al. [10] conducted statistical analysis based on the operating data of the Lanzhou public bicycle system, and adopted The

ISSN: 2664-9640

improved C4.5 classification algorithm clusters the sites according to the traffic information between the sites. Liu et al. [11] obtained a 5-day data set of Ningbo BSS site, which contains 617 sites 5 days and a total of 1000 statistical information, based on optical The density clustering model implements site clustering.

The above work is based on a large amount of BSS site time data, using data mining and clustering analysis techniques to identify cycling behavior, and extract the behavior pattern of each site and the relationship between the sites, which is helpful for in-depth study of the use characteristics and operation rules of the site, and prediction The site usage mode provides a basis for the optimal layout of the site, the setting of pile positions, and the scheduling decision.

### 2.2. Clustering based on the geographical distribution of sites

In 2016, Chen et al. [12] adopted a geographically restricted label propagation algorithm to group adjacent sites with similar usage patterns into one category. First, the sites were regarded as weighted network nodes. The location relationship produces a connecting edge. Based on the corresponding time, traffic, weather and other factors, the number of rental and return cars under the same factors is selected to construct the feature, and the weight of each edge is calculated to reflect the similarity of the site usage pattern. First of all. Each node is given a unique label, and its community label is continuously updated iteratively. After the end, each community obtained is regarded as a category of usage mode.

In 2018, Zhang et al. [13] took Zhongshan BSS as the research object to examine the patterns of bicycle sharing trips, travel chains and transition activities, and classified each site by site type according to the main activities around the site and the type of land use. These types are Residential (residential communities/buildings), commerce (such as shopping malls, markets, office buildings, banks, hotels), institutions (government buildings, schools/colleges, research institutions, hospitals, etc.), entertainment (parks, playgrounds, etc.) and transportation (Train station, intercity bus station, public bus terminal/hub). This definition method can clearly understand the type of land use near the station and the use mode of the station. Combined with the obtained workdays 1218244 travel trajectories, 334101 Travel chain, 462773 transition activities, using ArcGIS spatial statistical analysis tools to explore the travel mode and potential travel purpose of bicycle sharing. Yingshan et al.[14] obtained a total of 82,336 data from 2016/01/01-2016/08/31 According to the location and path conversion information of the station, firstly, all stations are initially clustered according to the latitude and longitude information of the station. On the basis of the first geographical location clustering, each station generates its own conversion matrix, and then combines it with K-means The algorithm forms the final clustering results of the sites to study the temporal and spatial distribution of the sites.

## 3. Research on BSS Car Rental Demand Forecast

Public bicycle borrowing and returning demand forecasting mainly predicts the number of rental and returning vehicles at all stations within a period of time. Public bicycle borrowing and returning demand forecasting occupies an important position in the research of public bicycles. It can predict the status of each station in the public bicycle network As time changes, it can provide managers with scientific and reasonable scheduling of bicycles in advance, and provide a basis for users to make reasonable travel plans. At present, the research on predicting the needs of public bicycle users mainly focuses on two ideas: one type is based on Time series model demand forecast; one type is demand forecast based on neural network model.

Time series models include autoregressive model(AR), moving average model(MA), autoregressive moving average model(ARMA), and autoregressive product moving average model (ARIMA). The autoregressive moving average model is suitable for stationary series, autoregressive The product moving average model is suitable for non-stationary sequences.

ISSN: 2664-9640

The neural network model has the characteristics of self-adaptation and non-linearity. It can quickly model non-linear data, and continuously adjust its network structure and connection weights through repeated learning of the training set. Prediction of unknown data. Neural networks currently widely used in time series trend forecasting include recursive delay networks, BP neural networks, Elman networks, etc. Due to the highly nonlinear and non-stationary characteristics of bicycle demand, it is related to time series The model can achieve more accurate predictions than the neural network model.

## 3.1. Demand forecast based on time series model

Since the data generated during the rental process of public bicycles constitutes a series of time series, a study based on time series analysis to explain the dynamic structure and laws of BSS and predict the demand for borrowing and returning cars is produced. In 2009, Borgnat et al. [15] obtained Based on the cycling data of the Paris BSS 2005/05/25-2007/12/12, the first-order autoregressive AR(1) model was used to predict the number of bicycle rentals per hour. In 2010, Kaltenbrunner et al. [16] aimed at the Barcelona BSS, The autoregressive moving average ARMA (10, 10) model is used to predict its hourly site demand. In 2011, Vogel et al. [17] used the Vienna BSS to divide the hourly rental and return time series model into relatively fixed rentals. The number of cars returned and the number of cars returned in a given hour fluctuate. The temperature, wind speed and other factors are selected, and the multiple linear regression model is used to predict the bicycle demand at the site.

In 2016, Lin Yanping [18] obtained the morning peak data of the Hangzhou BSS site for 20 working days, and divided the daily morning peak period into 8 time nodes at a time interval of 5 minutes, and obtained a total of 160 data points. The autoregressive quadrature moving average ARIMA(7, 1, 1) model predicts the demand for bicycles every 5 minutes in the morning peak period of the next day. The results show that the demand for public bicycles has obvious periodicity, which is similar to the baseline method. In contrast, the model can perform short-term time series forecasts better.

## 3.2. Demand forecast based on neural network model

In 2009, Froehlich et al. [1] used the data of Barcelona BSS 2008/11/02-2008/11/23 as the training set, and the data of 2008/11/24-2008/11/28 as the test set, using Naive Bayes Three prediction models: network model, historical average method, and historical trend method are used to predict the bicycle stock of the site. The results show that the prediction effect of Bayesian network is better than the other three models. In 2013, Xu et al. [19] obtained Hangzhou BSS 2011 /07/01-2011/12/31 half a year for all car rental and return information in the system, 20% of the records are randomly selected as the modeling data. A hybrid combination of K-means clustering and support vector machine (SVM) is used Model to predict the public bicycle traffic at the site.

In 2017, Xie Xiaoping [20] took Lanzhou BSS as the research object, obtained data on the morning peak demand of public bicycles for 6 consecutive working days from 2015/06/01 to 2015/07/10 for a single site, and constructed an improved Elman network forecast The model uses the data of the first 25 days as the training sample of the network, the demand for 3 consecutive days as the input vector, the demand for the 4th day as the target vector to obtain 22 sets of training samples, and the data of the last 5 days as the test sample of the network , Through simulation experiments, the prediction results of the model are compared with the actual demand to prove the effectiveness of the prediction method.

In 2018, Cao Xuening [21] used the BP neural network model to predict the short-term vehicle demand of public bicycle stations based on the idle pile data of a city for two consecutive working days for ten days, and used the mean square error to measure the prediction results. Lozano [22] took the Salamanca BSS as the research object, in 2019, Xu et al. [23] obtained the

ISSN: 2664-9640

Chicago BSS as the research object, and used the random forest model to predict the bicycle demand of the system. Du Mingyang et al. [24] obtained One month's travel data of Broadway&E14 S site in New York BSS, using an improved wavelet neural network model to predict the travel demand of the site during weekdays and weekends respectively.

# 4. Empirical analysis

## 4.1. Site cluster analysis

The NAB indicator is also known as the standardized number of bicycles available. It was proposed by Froehich [1] in 2009. It is a clustering indicator used to achieve clustering of public bicycle stations. Many scholars [2,4,6] are adopting the NAB indicator Cluster analysis is implemented for public bicycle stations in different cities. The calculation of the NAB indicator depends on the number of stakes and the number of bicycles available at the station. The value range is 0 to 1. When the NAB is close to 0, it means that the station is empty at the moment. When the NAB is close to 1, it means that the station is fully loaded at the moment, and the station has no free piles for parking. The NAB indicator has the following advantages: it can classify the stations according to the activity level of the station It can record the number of bicycles available at the station at different times; it can identify the full and empty status of the station, and provide a basis for bicycle dispatch. When the station is empty, it can dispatch bicycles in time to avoid the situation of no cars available. When When the site is fully loaded, the bicycles can be recalled from the site in time to prevent parking without a pile. However, in real life, the site may have many unexpected failures. For example: the site's piles are damaged and cannot be returned normally. Vehicles and damaged bicycles still occupy parking piles, and construction around the site leads to problems such as the failure to rent/return bicycles. As the NAB indicator excessively relies on the static data of the station's fixed piles and the number of available piles, when the bicycles are damaged or piles When it cannot be used normally, the damaged bicycles and damaged pile positions are still applied to the calculation of the NAB index. As a result, the NAB index cannot accurately reflect the actual use of the site in real time, and the clustering of the site is not accurate and effective. In addition, NAB The indicator can identify the change in the number of bicycles available at the site over time, but it cannot identify the change in the traffic flow of bicycles over time.

Using user cycling data, this paper proposes a new clustering index: Normalization Bicycles Net Outflow (NBNO) to describe the status of the site at any time. Namely:

$$BNO_{i,t} = B_{i,t}^{out} - B_{i,t}^{in} \tag{1}$$

Among them, represents the number of bicycles leaving the station at time $t$ at station $i$, and represents the number of bicycles entering the station at time t at station i. BNOi,t represents the net outflow of public bicycles at station $i$ at time $t$. For comparison, we normalize $BNO_{i,t}$ Processing is designed to normalize the result to between -1 and 1 for easy calculation and comparison.

$$NBNO_{i,t} = (1-(-1))\frac{(BNO_{i,t} - BNO_i^{\max})}{(BNO_i^{\max} - BNO_i^{\min})} + (-1) \tag{2}$$

Among them, $NBNO_{i,t}$ represents the normalized net outflow of bicycles at station $i$ at time $t$. The value ranges from -1 to 1. When $NBNO_{i,t}$ is close to -1, it means that the number of bicycles at station $i$ at time $t$ is far greater than the number of bicycles leaving the station. When the NBNO is close to 1, it means that the number of bicycles leaving the station at time $t$ is far greater than the number of bicycles entering the station, and the station $i$ at this time is represented as an outgoing station. When the NBNO is close to 0, It shows that the number of bicycles in and out of station $i$ at time $t$ reaches a balance, and station $i$ at this time is a balanced station.

ISSN: 2664-9640

The calculation of the NBNO indicator does not rely on static data such as the number of piles and empty piles at the site, but the historical data generated by the actual riding of the residents, which dynamically describes the use mode of the site. This avoids the failure caused by the actual situation. The NBNO indicator has the following three advantages: 1) Accurately identify the traffic flow in and out of the station at any time. Unlike the NAB indicator that identifies the number of empty piles at a station, the NBNO indicator can record the number of bicycles in and out of the station at each time, and further identify the in and out of stations The high peaks of and the corresponding moments can grasp the dynamic attributes of the site from a microscopic perspective. 2) The clustering results retain the site tidal effect; 3) Accurately identify weekdays and weekends. It can more effectively and intuitively describe the site usage. As follows Figures (a) and (b) represent the usage status of the site on weekdays and weekends, respectively. It can be found that there is a clear difference between the images on weekdays and weekends, and Figure (a) shows the tidal effect of the site.

Take the data from New York Citybike system on Sunday, 2017/10/01 and Monday, 2017/10/02 as samples, to conduct an empirical study on site clustering. Analyze site usage patterns corresponding to weekdays and weekends. The data is captured from the Citybike official website. The acquired data is in a non-standard format, so python3.7 is used to preprocess the data and extract features. Borgnat et al. [25] chose 15min, 30min, 1h, 2h, and 1day to divide the time window. The research found that the smaller the time window, The greater the fluctuation of the data; the larger the time window, the smoother the data but the short-term characteristics will be lost. When 1h is selected to divide the time window, it can better balance the volatility of the data and retain the short-term characteristics of the data, so this article chooses 1h to Divide the time window. The preprocessed data contains a total of 756 outbound time series data and 785 inbound time series data, which record the number of bicycle departures and arrivals per hour at each station.

Normally operating BSS sites must have two phenomena of bicycle entry and exit. There is no situation of only entering but not leaving, or only not leaving. Therefore, the two sets of time series data after preprocessing include only the stations where bicycles enter The sites that only contain the outflow of bicycles are eliminated, and the sites that have both in and out of bicycles are retained. After processing, a total of 1510 sequences containing NBNO features are generated. Each sequence is of equal length and contains 24 features, reflecting each The characteristics of the use of the site on the day. The K-means algorithm based on Euclidean distance [26] is used to cluster the NBNO sequence. Based on the results of previous studies [1,8,13], the site clustering is concentrated in about 5 clusters. Therefore, assign an initial value of 5 to the number of clusters k and observe the clustering results.

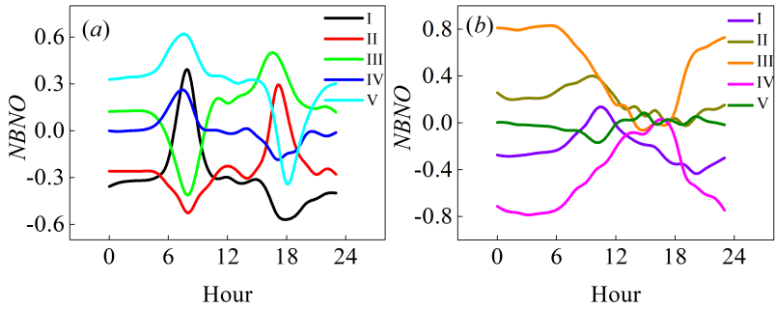


Fig. 1 Station clustering results(a)Station clusters on working day (b)Station clusters on weekend

Figure 1 depicts the clustering results of public bicycle stations of the CitiBike system in New York. (*a*) Represents the clusters of stations on weekdays. It can be found that the clusters of the five types of stations contain both peaks and valleys, at 7 am and 18 pm It is generated near the time, corresponding to the tidal effect produced by commuting to and from work on

ISSN: 2664-9640

weekdays. Class I and Class V sites correspond to residential sites. At around 7:00 in the morning, a large number of bicycles flow out due to people going to work. Around 18:00 in the afternoon , A large number of public bicycles flow back to the site with the peak off work. Such sites are generally deployed in residential areas and near residential buildings. Type II and III sites correspond to working sites, and they are the same as Type I and IV The usage pattern of such sites is just the opposite. Around 8 o'clock in the morning, because people go to work, there is a large influx of bicycles. Around 18 o'clock in the afternoon, there is a peak of vehicle outflow sites. Such sites are generally set up in government units, schools, and hospitals. Near other places. Type IV stations correspond to transportation and entertainment stations. Station activities generally occur during the day without obvious peaks. And during most of the day, the inflow and outflow balance can be maintained. This type of station is generally set up in subway stations, buses Stations and other transportation hubs as well as parks, playgrounds and other places. (*b*) Represents clusters of stations on weekends. It can be found that there is a significant difference from weekdays. Due to weekends, the stations no longer have tidal effects, and most of the stations are used frequently Decrease. Among them, Category I, Category II, and Category V sites are represented as balanced sites; Category III sites are represented as outgoing sites; Category IV sites are used in the opposite way to Category III sites, which are represented as inbound sites .

## 4.2. Forecast of bicycle demand at the site

Public bicycle demand forecasting occupies an important position in the research of public bicycles. It can provide dispatchers with accurate information by predicting the changes in bicycle demand at various stations in the BSS system over time, and shorten the time consumed by dispatch. Reduce scheduling costs.

The bicycle demand of the site is a set of time series, which is susceptible to the interference of various external factors such as temperature, wind speed, working days, holidays, etc., and has strong volatility and nonlinear characteristics. Due to the demand forecast of the neural network model It has strong adaptive and nonlinear characteristics, and can adapt to the time-varying, nonlinear and uncertain characteristics of public bicycle site demand changes. Therefore, this paper uses the Elman neural network model to predict the short-term bicycle demand at the site.

The Elman neural network is a typical dynamic recurrent neural network, which consists of an input layer, a hidden layer (middle layer), a receiving layer and an output layer [27]. Since the output of the hidden layer is delayed and stored by the receiving layer, it is self-connected to The input of the hidden layer ensures its sensitivity to historical data. The internal feedback network is added to enhance the ability of the network to process dynamic information, thereby achieving the purpose of dynamic modeling. The nonlinear state space expression of the Elman network is [28]:

$$y(t) = g(w^3 x(t)) \; ; \; x(t) = f(w^1 x_c(t) + w^2 (u(t-1))) \; ; \; x_c(t) = x(t-1) \tag{3}$$

Among them, $t$ represents the current moment of the neural network, $t$-1 represents the previous moment; $y$ is the m-dimensional output node vector; $x$ is the n-dimensional intermediate layer node unit vector; u is the r-dimensional input vector; xc is the n-dimensional feedback state vector ; $w^3$ is the weight of the connection between the middle layer and the output layer; $w^2$ is the weight of the connection between the input layer and the middle layer; $w^1$ is the weight of the connection between the receiving layer and the middle layer; $g$(*) represents the transfer function of the output neuron, which is the middle layer The linear combination of the output; $f$(*) represents the transfer function of the middle layer neuron.

Take the #312 site as an example, obtain the bicycle demand data of the site for a total of 31 days from 2017/10/1 to 2017/10/31 as a sample, and use the Elman neural network to

ISSN: 2664-9640

construct the site public bicycle demand prediction model. According to the site bicycle demand Historical data, the input and output nodes of the feedback neural network are selected to reflect the inherent law of bicycle demand, so as to achieve the purpose of predicting the bicycle demand at the site in the future. Use the data of the previous 30 days as the training sample of the network, the first day The demand data of the second day is used as the input vector, and the demand data of the second day is used as the target vector. In this way, 29 sets of training samples are obtained. The data of the last day is used as the test sample to verify that the network can reasonably predict the situation at each moment of the day Bicycle demand.

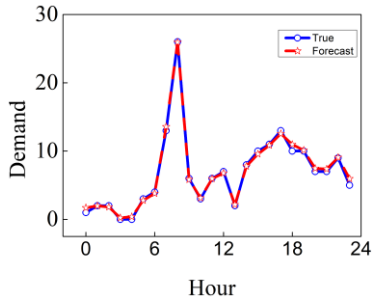


Fig. 2 Demand forecast results

Figure 2 is the prediction result of the Elman neural network. The red curve represents the predicted bicycle demand. The blue curve represents the real bicycle demand. According to the simulation results, it can be found that the two lines can fit together well, and the predicted data The trend is consistent with the actual situation. In order to more accurately describe the model's prediction of the bicycle demand at the site, this paper uses Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to evaluate the prediction effect of the model . RMSE is the square root of the ratio of the sum of squares of the deviation between the observed value and the true value to the number of observations m. It is used to measure the deviation of the observed value from the true value. MAE is the average of the absolute error, which can better reflect the true situation of the predicted value error This article uses two evaluation indicators to display the evaluation results, making the prediction results more rigorous. The calculated RMSE=0.5669, MAE=0.3158, indicating that the Elman neural network predicts The average error of the demand for public bicycles at the site is no more than 1 at 24 moments, which can achieve a good prediction effect. The empirical analysis proves that the neural network is effective in predicting the demand for public bicycles.

## 5. Conclusion

This paper uses user cycling data to construct a new clustering index to avoid errors caused by relying on site static data. In addition, empirical analysis is carried out by obtaining real data from CitiBike. The results show that the NBNO index retains the foundation of the NAB index's advantages The above shows the following three advantages: 1) Accurately identify the traffic flow in and out at all times, and grasp the dynamic attributes of the station from the micro level; 2) The clustering results show the tidal effect of the station, and it is easy to identify the station pattern; 3) Accurately identify weekdays and weekends , Intuitively display human behavior patterns and site usage. The NBNO indicator uses historical data generated by residents' cycling to avoid errors caused by pile damage and vehicle damage in the NAB indicator calculation process, and can be more realistic and effective Achieve site clustering, reflecting the actual use of the site. The clustering results in this article also further confirm that the use pattern of bicycle clusters is highly correlated with the land use patterns of different site settings. The site use pattern is inflow in the morning and outflow at night Sites are generally set up near office buildings, government units, and financial institutions, and are called work sites. Sites are used in the morning and inflow at night. They are generally located in residential areas and near

residential buildings, called residential sites. The station maintains a balanced amount of bicycle inflow and outflow, and only small fluctuations occur during commuting hours. The stations are generally set up near transportation hubs such as subway entrances, bus stations, and railway stations, which are called transportation-type stations. By clustering the stations, It can analyze the spatio-temporal characteristics of each type of site, which reduces the workload and improves the accuracy of the prediction. It can formulate reasonable scheduling strategies and site capacity adjustments based on the clustering results, effectively alleviating the unreasonable site distribution and capacity settings. Cars can be borrowed, no piles can be parked, etc., to achieve efficient site management. In addition, this article also verifies the effectiveness of the Elman neural network model in site prediction through one-month bicycle demand data at a single site.

Regarding the problem of BSS site clustering and demand forecasting, many scholars are still studying. Through the empirical analysis of this article, it is found that the method of clustering and forecasting sites based on user cycling data is effective. Research on BSS , And some scholars have developed from the perspective of complex networks. In 2013, Borgnat et al. [29] took Lyon's Velov system as the research object, and proposed that sites are regarded as nodes, bicycle sharing exchanges are connected edges, and a complex network of public bicycles is constructed to study user travel. The characteristics of the statistical characteristics in the temporal and spatial distribution. The community discovery algorithm based on the maximization of modularity is used to cluster the sites with the same usage pattern. In 2018, Lu Ling et al. [30] used a city's BSS for 5 months of cycling The data builds a complex network of public bicycles. The network contains 434 nodes and 19,472 edges. The key nodes are identified by analyzing the topological characteristics of the network, and then optimization suggestions are provided for the key nodes.

BSS data mining research helps us better understand the human behavior hidden behind the data, and explore the dynamics of human behavior behind it; at the same time, it can provide public bicycle operators with daily management, site location planning, and site location planning based on the results of data analysis. Provide reference for site capacity settings, vehicle scheduling plans, etc.; provide reference plans for users to rent cars and optimize user travel experience. Currently, research based on public bicycle data mining is still a hot issue. With the advent of artificial intelligence era, machine learning and deep learning And other data mining technology innovations will further promote the rapid development of research in this field. Future research may be more integrated with external environmental characteristics, such as weather, seasons, special events, etc., to provide BSS with more accurate load balancing and site capacity prediction, explanation The temporal and spatial dynamics of the city.

## Acknowledgments

We appreciate editor and reviewers for their positive and constructive comments and suggestions. Funding: This work was supported by Shaanxi Province Science and Technology Development Plan (Science and Technology Rising Star) [grant number 2020KJXX-051]; and partly supported by Shaanxi Provincial Natural Science Basic Research Program [grant numbers 2021JQ-961]; and partly supported by Funding Projects for Fundamental Scientific Research Operation Fees of Central Universities[grant number 300102351502].

## References

[1]Froehlich J, Neumann J, Oliver N.Sensing and Predicting the Pulse of the City Through Shared Bicycling[C]. Washington：IJCAI, 2009: 1420-1426.

[2]Lathia N, Ahmed S, Capra L.Measuring the impact of opening the London shared bicycle scheme to casual users[J].Transportation Research Part C, 2012, 22：88-102.

[3]Lin Yanping, Dou Wanfeng.Research on demand prediction of urban bicycle sharing based on network model[J]. Application Research of Computers, 2017, 34(09): 2692-2695.

[4]WuJiansheng, WangLuyi, LiWeifeng. Usage Patterns and Impact Factors of Public Bicycle Systems: Comparison between City Center and Suburban District in Shenzhen[J]. Journal of Urban Planning and Development, 2018, 144(3): 04018027.

[5]Borgnat P, Fleury E. Robardet C, Scherrer A. Spatial analysis of dynamic moventments of Velo'v Lyon's shared bicycle program[C]. ECCS'09, Warwick, HAL, 2009.

[6]FengYunlong, Roberta C A, Marc Z.Analysis of bike sharing system by clustering: the Velib case[J].IFAC- PapersOnLine, 2017, 50(1): 12422-12427.

[7]Shen Xingfa, Wang Landi. Rental Points Clustering and Function Identification of Public Bicycle System[J]. Computer Engineering, 2018, 44(01): 44-50.

[8]Vogel P, Mattfeld D.Strategic and Operational Planning of Bike-Sharing Systems by Data Mining-A Case Study[C]. Berlin, Springer-Verlag, 2011:127-141.

[9]Etienne C, Latifa O. Model-Based Count Series Clustering for Bike Sharing System Usage Mining: A Case Study with the Vélib' System of Paris[J]. ACM Transactions on Intelligent Systems and Technology-Special Section on Urban Computing, 2014, 5(3): 39.

[10]Zhao Hong, Shi Ruigang. Assessment and investigation of parking site setting and lock-pile disposition public bicycle system[J]. Journal of Lanzhou University of Technology, 2018, 44(05): 107-112.

[11]Liu Liangxu, Hu Zhenghua, Zhou Chaolan, Xu Guohu. Research on the clustering algorithm of the bicycle stations based on OPTICS[J]. Concurrency and computation-practice & experience, 2019, 31(10): e4876.

[12]Chen Longbiao, Yang Daqin, Wang Leye, Ma Xiaojuan, Li Shijian. Dynamic cluster - based over - demand prediction in bike sharing systems[C]. Heidelberg, ACM, 2016: 841-852.

[13]Zhang Ying, Brussel M J G, Thomas T, Maarseveen M F A M. Mining bike-sharing travel behavior data: An investigation into trip chains and transition activities[J]. Computers, Enviroment and Urban Systems, 2018, 69: 39-50.

[14]Chong Yingshan, Han Xiaoming. Prediction of shared bicycle site demand based on random forest and spatiotemporal clustering[J]. Science Technology and Engineering, 2018,18(32): 89—94.

[15]Borgnat P, Abry P, Flandrin P, Rouquier J-B. Studying Lyon's V´elo'V: A Statistical Cyclic Model[C]. ECCS'09, Warwick, HAL, 2009.

[16]Kaltenbrunner A, Meza R, Grivolla J, Codina J, Banchs R. Urban Cycles and Mobility Patterns: Exploring and Predicting Trends in a Bicycle-based Public Transport System[J]. Pervasive and Mobile Computing, 2010, 6(4): 455-466.

[17]Vogel P, Greiser T, Mattfeld D C. Understanding bike-sharing systems using data mining: exploring activity patterns[J]. Procedia-Social and Behavioral Sciences, 2011, 20: 514-523.

[18]Lin Yanping，Dou Wanfeng. Research on Short-Term Prediction Method of Demand Number in Urban Public Bicycle Based on the ARIMA Model[J]. Journal of Nanjing Normal University(Engineering and Technology Edition), 2016, 16(03): 36-40.

[19]Xu Haitao, Ying Jing, Wu Hao, Lin Fei.Public Bicycle Traffic Flow Prediction based on a Hybrid Model[J]. Applied Mathematics & Information Sciences, 2013, 7(2): 667-674.

[20]Xie Xiaoping, Qiu Jiandong, Tang Minan. Demand prediction of public bicycle rental station based on Elman neural network[J]. Computer Engineering and Applications, 2017, 53(16): 221-224.

[21]Cao Xuening. Research on Short-term forecasting Method for Public Bicycles [A]. Academic Committee of Urban Transportation Planning, China Urban Planning Society. Innovation Driven and Intelligent Development - Papers Collection of the 2018 Annual Conference on Urban Transportation Planning in China [C]. Academic Committee of Urban Transportation Planning, China Urban Planning and Design Institute, China Urban Planning and Design Institute, 2018: 12.

[22]Lozano A, Paz J F D, Gonzalez V, Iglesia D H D L, Bajo J. Multi-Agent System for Demand Prediction and Trip Visualization in Bike Sharing Systems[J]. Applied Sciences, 2018, 8(1): 67.

[23]Xu Haitao, Duan Feng, Pu Pan. Dynamic bicycle scheduling problem based on short-term demand prediction[J]. Applied Intelligence, 2019, 49: 1968-1981.

ISSN: 2664-9640

[24]Du Mingyang, Cheng Lin, Li Xuefeng. Prediction of Public Bike Trip Demand Based on APSO-WNN[J]. Journal of Highway and Transportation Research and Development, 2019, 36(06): 94-102.

[25]Borgnat P, Robardet C, Rouquier J -B, Abry P, Fleury E, Flandrin P. Shared bicycles in a city: a signal processing and data analysis perspective[J]. Advances in Complex Systems 2011, 14(03): 415-438.

[26]Preeti A, Deepali D, Shipra V. Analysis of k-means and k-medoids algorithm for big data[J]. Procedia Computer Science, 2016, 78: 507-512.

[27]Zhou Jianguo, Li Wei, Yu Xuechao, Xu Xiaolei, Yuan Xiaolei, Wang Jiashuai.Elman-Based Forecaster Integrated by Adaboost Algorithm in 15 min and 24 h ahead Power Output Prediction Using PM 2.5 Values, PV Module Temperature, Hours of Sunshine, and Meteorological Data[J]. Polish Journalof Environmental Studies, 2019, 28(03): 1999-2008.

[28]Zhu Qinghui, Li Guangru, Gou Xiangyu, Li Haili. Ship traffic flow prediction based on Elman neural network optimized by cyclic structure[J]. Chinese High Technology Letters, 2019, 29(03): 295-301.

[29]Borgnat P, Robardet C, Abry P, Flandrin P, Rouquier J-B, Tremblay N. A dynamical network view of Lyon's Velo'v shared bicycle system[J]. Dynamics on and of Complex Networks, 2013, 2: 267-284.

[30]Lu Ling, Peng Yali, Zeng Xinyi, Yang Yuxin, Huang Minghe. An accessibility index potential evaluation model for the complex network of public bicycles[J]. Computer Engineering & Science, 2018, 40(01): 175-183.