Research on the Prediction of Sharing Bicycles Based on ARIMA Model

Zijian Lin[1] and Yajie Zhang[2*]

[1,2]Department of Information and Technology, Wenzhou Polytechnic, Wenzhou 325035, China

**Corresponding Author:* [2]zhangyajie98@outlook.com

***Abstract*—** Bicycle-sharing systems are a new type of transportation service that provides bicycles for shared use; they allow users to rent a bicycle at one station, ride it, and return it to another station in the same city. In this paper the author, the trip data of Shanghai Mobike in August 2016 are taken as the main raw data, and the trip characteristics of shared bicycle system are deeply studied by using data mining method, and the ARIMA model is used to predict bicycle trip. Finally, RMSE is used to judge the prediction accuracy. The results show that the model can effectively predict the residents' trip, and the prediction accuracy is high, which can provide a certain reference for residents' trip.

***Keywords*—** Sharing Bicycles; ARIMA; Prediction; Time Series; Travel Characters.

I. INTRODUCTION

The Public Bicycles System (PBS), which originated in Europe, began to enter China in 2007, and began to pilot in large cities such as Beijing, Hangzhou and Wuhan, and gradually expanded to other provincial capital cities and small and medium-sized cities. The operation of the public bicycles system formally integrates bicycles into the field of public transport, seamlessly butt the slow traffic and public transport, and crack the "last kilometre" problem at the end of the traffic, and achieve a better city with low carbon travel. In 2012, the Ministry of housing and Construction issued the guidelines for the planning and design of urban walking and bicycle traffic systems. By 2016, the total amount of public bicycles in China has reached about 500000, exceeding the total of other countries in the world.

With the rapid development of the Internet, the application of mobile Internet technology has been paid more and more attention. In the past two years, the sharing economy or sharing economy has been developing in full swing. The typical representatives of foreign sharing economy include Uber and Airbnb. The domestic sharing economy is also developing rapidly. In the field of trip, DiDi as the representative of such as easy road vehicles, China special car, tick-sharing and other Internet enterprises, to the vast number of users to bring great convenience to trip. However, in the field of short distance trip above the enterprise still has no business coverage. In order to better solve the "last kilometer" pain point of short distance trip, a large number of shared bicycle enterprises have been set up. At present, share bike field originator and leader OFO and Mobike as the representative. The biggest innovation in sharing bicycles is that they abandon the traditional fixed parking stacks and scan the QR code with their smartphones to unlock them. The emergence of shared bikes not only solves the "last kilometer" trip problem of users, saves the time cost of waiting for cars and the cost of service, but also makes sharing bicycles more convenient and more flexible to take and return than public bicycles with piles. High cost performance.

Xi'an began to invest in sharing bicycles in 2016, up to now 52 thousand, more than 500 thousand public cards, and the average daily trial number reached 230 thousand. In the opinion polls, 95.3% of the citizens expressed satisfaction with the system. Xi'an sharing bicycles are the main supplement to the subway and bus, to solve the last 2 kilometers trip problem of the station and home, and take the subway export, the bus hub and the large area as the main points.

II. THE CURRENT SITUATION OF SHARING BICYCLES IN SHANGHAI

By the end of 2017, the total number of users in China's shared bicycle market had reached nearly 50 million. According to the 2017 China shared Cycling Market Research report released earlier by BIGDATA Consulting, a third-party data research firm. According to data from TRUSDATA on April 6, 2017, the number of active users in the domestic shared bicycle market is about 4.322 million per month. By that measure, Mobile's market share has reached an alarming 72.5%.

The scale of sharing bicycles is becoming larger and larger, which brings convenience to users and also brings some problems at the same time. For example, there are not enough vehicles in the vicinity of densely populated areas to meet the demand of residents; due to the tide of traffic, there will be a two-way flow imbalance in the morning and evening rush hours; and there is a large demand for rental in local locations in a single direction. On the other hand, the demand for parking is large, the demand for vehicles depends on

artificial experience, the lack of platform data support, intelligent management level to be promoted and so on.

At present, the demand scheduling of shared bicycle users is mainly carried out by the combination of manual inspection and site video monitoring. This approach has great disadvantages:

① the cost of manual inspection is high and the efficiency is low. The scale of shared bicycle is huge and wide, so it is difficult to solve the problem that other sites have no car to borrow for a long time.

②with the increasing number of nodes, video surveillance can not understand the situation of each node in real time;

In order to alleviate the problem of loan and return, improve customer satisfaction, reduce the operating cost of the system, and improve the service level of the shared bicycle system as a whole, it is necessary to reasonably and accurately forecast the demand for consumption. In this study, based on the ARIMA model, the data of Mobike trip in Shanghai and the location of the site are used to forecast the demand.

At present, there are also some problems in the operation of the sharing bicycles system in Shanghai. For example, the number of sites in the densely populated area is not enough to meet the needs of the residents to borrow and return the car. Because of the tide phenomenon, there will be a two-way flow imbalance in the peak period of the morning and evening, and the demand for a few stations in a single direction is large, In the other direction, there is a large demand for parking, and some sites appear empty, users can not borrow cars, some sites are full of full status users can not return the car; the site of the demand for vehicles depends on artificial experience, lack of platform data support, intelligent management level to be improved.

III. SOURCE OF DATA

The original data in this paper come from two parts. One is take from Mobike's one-month trip data in Shanghai from August 1 to 31 in 2016, which get a total of 102362 trip records. Each trip record contains information such as order number, bicycle number, user ID, start / end time, latitude and longitude points, etc., as shown in figure below:

Order NO.	Bike NO.	UID	Start Time	Start Lat	Start Long	End Time	End Lat	End Long	Trace Point
78387	158357	10080	2016/8/20:06:57	121.348	31.389	2016/8/4 6:02:45	121.348	31.389	#121.354,31.391#121.355,31.391#...
891333	92776	6605	2016/8/29:19:09	121.508	31.279	2016/8/29 19:31	121.489	31.271	#121.489,31.270#121.489,31.271#...
76435	347335	8135	2016/8/30:12:49	121.467	31.32	2016/8/4 13:03	121.447	31.318	121.447,31.317#121.447,31.318#...

Figure 1: Data Table for Sharing bicycles

This paper analyses the operations data of 20 days about sharing bicycles of the above 180 sites, and analyzes the characteristics of the system lending and return vehicle, such as the time characteristics, the frequency characteristics and the turnover rate, which will help to deepen the understanding of the sharing bicycles system and understand the running rules of the system, and make clear the main influencing factors for the system travel. It is helpful for the operation and management department to take effective measures to improve the

service level of the sharing bicycles system and the satisfaction of the users to the system. It is of great significance for the sustainable development of the sharing bicycles system to control the management cost of the system properly.

The other part of the data from the online questionnaire survey. The investigation was carried out in Xuhui District, Lujiazui District, Putuo District, Hongkou District and other typical areas in Shanghai. A total of nine days were issued on 14, 15 and 22 August and from 23 to 27 August, including five working days and four non-working days. Questionnaires were sent out in person and collected in person. The subjects of the survey are residents of the city, divided into the use of shared bicycles and non-use of the two groups of people, targeted design of two types of questionnaires. A total of 1347 valid questionnaires were collected, including 1205 for shared bicycles and 142 for non-users.

The questionnaire for the purpose of sharing bicycle trip is: which of the following purposes do you use shared bike for? (multiple options) ①Shopping ②Dining and entertainment③ Exercise ④Commute to and from work. ⑤Go on a trip.⑥Others. From the results of the questionnaire, commuting and shopping are the two main purposes of trip.

Table 1: Questionnaire

Which purpose do you use public bicycles for?(Multi Choices)					
Shopping	Enthronement	Exercise	Commute	Excursion	other

IV. MODEL AND PROCEDURE

A. ARIMA Model

The ARIMA model is called autoregressive integral moving average model, which is a famous time series prediction method proposed by Box and Jenkins in the early 1970s[1][2]. So it's also called the Box-Jenkins model and the Boxes-Jenkins method. The ARIMA (p, d, q) called differential autoregressive moving average model (AR) is the number of differences made when autoregressive p is the autoregressive term and MA is the moving average term and d is the time series when the time series becomes stationary [5]. The so-called ARIMA model is the model that transforms the non-stationary time series into the stationary time series, and then regresses the dependent variable only to its lag value and the present value and the lag value of the random error term. The ARIMA model includes moving average process MA, autoregressive process AR, autoregressive moving average process ARMA and ARIMA process according to whether the original sequence is stationary or not.

The characteristic of ARIMA model is that the change of other correlated random variables is not directly considered, but the data sequence formed by the prediction object over time is treated as a random sequence. And the random sequence can be generated by the autoregressive moving average process, that is, the time series can be explained by its own past value or lag value and random disturbance term. If the time series is stable, that is, its behavior does not change significantly over time, it can predict future values through the past and present values of the time series.

ARIMA model can effectively deal with autocorrelation non-stationary data, so this paper uses ARIMA model to predict the demand of bicycle in the short term, which can help users to know how many vehicles will be borrowed or returned in the future. Then according to this information to decide whether to rent or return to other sites; second, to enable the management to accurately grasp the site bicycle rental situation, so as to effectively allocate the site of the actual vehicle.

B. Modeling Process

The basic process of forecasting the trip demand of shared bicycle stations by establishing ARIMA (p, d, q) model. The flows were explained in detail as follows:

①Drawing time sequence diagram to judge whether there is a clear trend or period, that is, to identify the stationary. ② For the non-stationary time series data, the D-order difference is carried out, and the stationary sequence is obtained.

③ After the original non-stationary time series is smoothed, the hypothesis test is used to determine whether the difference time series is a white noise sequence. ④After stationary treatment, if the partial autocorrelation function is truncated, and the autocorrelation function is trailing, the AR model is established, if the partial autocorrelation function is trailing, and the autocorrelation function is truncated, the MA model is established. If both partial autocorrelation function and autocorrelation function are trailing, the sequence is suitable for ARMA model. ⑤ After the order of the model is determined, the parameters of the ARMA model are estimated, and the least square method is commonly used to estimate the parameters. ⑥ Establishing appropriate model. ⑦According to the principle of AIC, SBC, the optimal model is chosen. ⑧ Build the model to forecast the demand of shared bicycle trip station. A specific modeling flowchart, as follows:

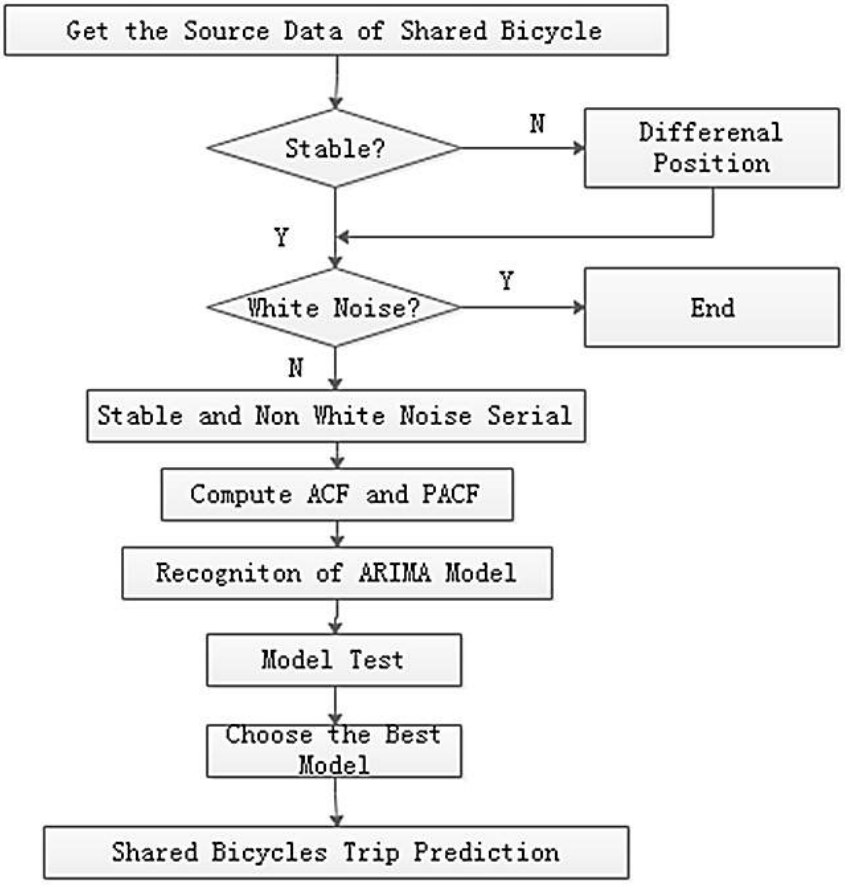

Figure 2: Flow Chart of the Model

C. Data Cleaning

Among the 102362 trip records of the original data, the invalid trip records should be screened out first, and the data should be screened out. Because of these invalid data will affect the sharing of bicycle loan-to-return space-time features and the characteristics of turnover and other characteristics of the analysis.

Shared bicycle trip time and distance is too short, can not be considered as an effective trip. According to the definition of trip, the one-way distance should be more than 400 meters, and the minimum trip time calculated by the average speed of 9.6 kilometers per hour is 2.5 minutes. The usage time of the shared bicycle is calculated as follows:

$$t = t_a + t_b + t_c \quad \ldots\ldots\ldots\ldots (1)$$

Formula (1), t is the use time of the shared bicycle, t_b is the actual time of sharing bicycle, t_a and t_c is the time spent in leasing and returning, which is generally 15 seconds. Therefore, if the sharing of bicycle trip data is valid, then the vehicle time should not be less than 3 minutes.

The other two invalid data are canceling the car loan data and abnormal car return data. For example, the cancellation of the car loan data is generally due to a change in the tripper's temporary plan, after the cancellation of the trip, or after the car found to be a malfunction car and a change. The data is typically available for less than three minutes and is available at the same site. Common features of abnormal car return data, such as the location and time of the car being borrowed, but not the place and time of return, or the

time of the car being used for more than 10 hours. After these invalid data are eliminated, the remaining data are used as the characteristic analysis of the number of times of sharing bicycle loan and return, the space-time characteristic analysis, the characteristic of turnover rate, and then the modeling is carried out.

D. Stationary Test

The original time series is tested after the first order difference. If we have not heard of the test, we need to carry out the quadratic difference transformation. The sequence diagram after the first order difference has no obvious trend, it can be preliminarily considered as a stationary time series.

The ARIMA (p,d,q) model, the AR is an autoregressive term or an autoregressive term, and MA is a moving average term Q is a moving average term, and d is the number of differences made when the time series becomes stationary.

Let y_t be a single integer sequence of order d, that is, yt~I (d), is denoted as

$$w_t = \Delta^d y_t = (1 - L^d) y_t \quad \text{.....(2)}$$

Is a stationary sequence, Ie. I (0), so the model is:

$$Y_{m+1}(t) = \alpha_{11} Y_{11}(t-5) + \cdots + \alpha_{mp} Y_{mp}(t - p*5) + \beta_{11}\mu_{11}(t-5) + \cdots + \beta_{mp}\mu_{mp}(t - p*5) + N(t) \quad \text{...(3)}$$

The $Y_{m+1}(t)$ represents the number of leased vehicles at the $t - p*5$ period in advance of m days. The model is composed of the observed values of 1 to p step sizes filled by m and the random term linearly.

E. White Noise Testing

After processing the original non-stationary time series, it is necessary to judge whether the difference time series is white noise or not. White noise is a pure stochastic process, it is strictly stationary, its original assumption is that the delay period is less than or equal to m period sequence values are independent of each other. The experimental results show that the p value is less than 0.05, which rejects the original hypothesis, so it is not a white noise sequence.

F. Recognition and Order Determination of Models

In many parameter estimation problems use likelihood function as the objective function. When there are enough training data, the accuracy of the model can be improved continuously, but at the cost of increasing the complexity of the model. At the same time, it brings a very common problem in machine learning-over-fitting. Therefore, the model selection problem seeks the best balance between the complexity of the model and the ability of the model to describe the data set.

Many information criteria are proposed to avoid over-fitting problem by adding penalty terms of model complexity. There are two commonly used model selection methods-Akaike Information Criterion(AIC) and Bayesian Information guideline (BIC). AIC is a standard to measure the good fit of statistical model, which was proposed by Japanese statistician Hirosuke Chichi in 1974. It is based on the concept of entropy and provides a standard to weigh the complexity of the model and the goodness of fitting data.

In general, AIC is defined as:

$$AIC = 2k - 2\ln(L) \quad \text{....(4)}$$

Where k is the number of model parameters, L is a likelihood function. When selecting the best model from a set of models available, the smallest AIC model is usually selected.

When there is a big difference between the two models, the difference is mainly reflected in the likelihood function item. When the likelihood function difference is not significant, the first term of the upper formula, that is, the complexity of the model, plays a role, so the model with a small number of parameters is a better choice.

In general, when the complexity of the model increases (k increases), the likelihood function L also increases, which makes the AIC become smaller, but when k is too large, the growth rate of likelihood function slows down, leading to the increase of AIC, and the model is too complex to cause the phenomenon of over-fitting. The goal is to select the model with the smallest AIC, not only to improve the model fitting degree (maximum likelihood), but also to introduce a penalty term, so that the model parameters are as few as possible, which will help to reduce the possibility of over-fitting.

BIC, similar to AIC, is used for model selection, which was proposed by Schwarz in 1978. When training the model, increasing the number of parameters, that is, increasing the complexity of the model, will increase the likelihood function, but it will also lead to the phenomenon of over-fitting. In order to solve the problem, both AIC and BIC introduce penalty terms related to the number of parameters of the model. The penalty term of BIC is larger than that of AIC. When the number of samples is too large, it can effectively prevent the model complexity caused by too high model precision.

$$BIC = k\ln(n) - 2\ln(L) \quad \text{..... (5)}$$

Where K is the number of model parameters, n is the number of samples and L is the likelihood function. $k\ln(n)$ is penalty item can effectively avoid the

phenomenon of dimensionality disaster when the dimension is too large and the training sample data is relatively small.

V. EXPERIMENTS AND ANALYSIS OF RESULTS

First of all, the data set is divided into two parts: the data of 8.1-8.19 in 2016 as model development data and the data of 8.20-8.31 as model validation data. Model assessment indicators:

In this paper, the RMSE (root mean square error) is used to measure the accuracy of the prediction. It is the square root of the square of the deviation between the predicted value and the actual value and the ratio of the prediction number n. The calculation method is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(\hat{y} - y_t)^2} \qquad (6)$$

Where m denotes the number of predicted data, which is the original data, which is the data predicted by the model. The smaller the numerical value is, the smaller the error of the model is, and the higher the fitting degree is.

VI. CONCLUSION

Based on the ARIMA model, this paper mainly analyzes and forecasts the characteristics of Mobike's residents' travel in Shanghai. Compared with other forecasting methods, it excludes some complicated influencing factors and carries out more objective and realistic sustainable prediction and improved the accuracy of the prediction. BIC criterion is used to identify and determine the order of the model. In the experiment, the trip is predicted and verified. By comparing the root mean square error and the chart, we can see that the prediction effect is remarkable.

It shows that the model can provide a reliable reference for residents to travel. The travel distance of shared bicycle is mainly 1-4km, which is convenient and flexible, and has the advantage that bus and car can't be replaced in short distance travel. We should give full play to the advantages of sharing bicycle, and combine the shared cycle system with conventional traffic, BRT and so on, to assist public transport to complete the whole process of public transport.

VII. SUGGESTIONS FOR DEVELOPMENT

In this paper, taking Shanghai as an example, taking the data of sharing bicycles brushing card as the main data, this paper discusses the characteristics of sharing bicycles travel along the city of Haiti City, carries out multi-dimensional characteristics analysis and operation evaluation, and gives the operation and scheduling suggestions.

Through the reform of the management process, such as the comfort of the sharing bicycles, the management of the card, and so on, it will provide the people with the humanized service and truly meet the user's needs, thus further improving the turnover rate of the sharing bicycles.

Sharing bicycle enters the market with the concept of green and low carbon environmental protection, according to the investigation. In recent years, the development of sharing bicycle enterprises has reached a bottleneck period and also increased the society. It is difficult to be able to manage the public. After explosive growth, sharing bicycles should be found in time. The problem is to focus on more refined management and green development, thus becoming ripening period.

ACKNOWLEDGMENT

This work was supported by the Xinmiao Talents Program of Zhejiang Province (2019R463009).

REFERENCES

[1] Castillo-Manzano J I, Castro-Nuño M, López-Valpuesta L. Analyzing the transition from a sharing bicycles system to bicycle ownership: A complex relationship[J]. Transportation Research Part D Transport & Environment, 2015, 38:15-26.

[2] Contreras J, Espinola R, Nogales F J, et al. ARIMA models to predict next-day electricity prices[J]. IEEE Transactions on Power Systems, 2003, 18(3):1014-1020.

[3] Calster T V, Baesens B, Lemahieu W. ProfARIMA: a profit-driven order identification algorithm for ARIMA models in sales forecasting[J]. Applied Soft Computing, 2017, 60.

[4] Zhang T G. Analysis and Forecast of Temporary Price of International Grain based on ARIMA Model- Taken Soybean as Example[J]. Prices Monthly, 2016.

[5] Chen Q, Sun T. A model for the layout of bike stations in public bike-sharing systems[J]. Journal of Advanced Transportation, 2016, 49(8):884-900.

[6] Benarbia T 1. Modelling and control of self-service sharing bicycles systems by using Petri nets[J]. International Journal of Modelling Identification & Control, 2018, 17(3):173-194.

[7] Benarbia T 1. Modelling and control of self-service sharing bicycles systems by using Petri nets[J]. International Journal of Modelling Identification & Control, 2018, 17(3):173-194.

[8] Shaheen S A, Martin E W, Cohen A P, et al. Public bikesharing in North America During a period of rapid expansion: Understanding business models,

industry trends & user impacts: MTI Report 12-29[J]. Mineta Transportation Institute, 2017.

[9] Li S ,Zhuang C,Tan Z,et al. Inferring the trip purposes and uncovering spatio-temporal activity patterns from dockless shared bike dataset in Shenzhen, China[J].Journal of Transport Geography, 2021, 91(1):102974.

[10] Maas S , Nikolaou P , Attard M , et al. Classifying bicycle sharing system use in Southern European island cities: cycling for transport or leisure[J]. Transportation Research Procedia, 2021, 52(1):565-572.

[11] Gao L,Ji Y,Yan X,et al. Incentive measures to avoid the illegal parking of dockless shared bikes: the relationships among incentive forms, intensity and policy compliance. 2021.

[12] Feizi A , Mastali M , Van Houten R , et al. Effects of bicycle passing distance law on drivers' behavior[J]. Transportation Research Part A: Policy and Practice, 2021, 145.

[13] Wei Z , Zhen F , Mo H , et al. Travel Behaviours of Sharing Bicycles in the Central Urban Area Based on Geographically Weighted Regression: The Case of Guangzhou, China[J]. Chinese Geographical Science, 2021.

[14] Shao P , Wang Q , Zhao C ,et al. Research on the factors influencing shared bicycle green use behavior and intention. Journal of Arid Land Resources and Environment, 2020.

[15] Sm A ,Ma A , Mac B . Assessing spatial and social dimensions of shared bicycle use in a Southern European island context: The case of Las Palmas de Gran Canaria[J]. Transportation Research Part A: Policy and Practice, 2020, 140:81-97.

[16] Yang L , F Zhang, Kwan M P , et al. Space-time demand cube for spatial-temporal coverage optimization model of shared bicycle system: A study using big bike GPS data[J]. Journal of Transport Geography, 2020, 88.

[17] Hu X ,Cheng J,Zhong W.Analysis and Exploration of Open Source Data in Traffic Network Based on Shared Bicycle Arrangement Model[J]. 电脑学刊, 2020, 31(2):227-240.

[18] Ji H,Xu H,Zhou C,et al.Design and implementation of analysis and visualization of shared bicycle information. 2020.