Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Short-Term Travel Volume Prediction and Delivery Volume Prediction of Shared Bicycles Built on Different Sites

Xintong Huang*

Department of Information Engineering, University of Finance & Economics, Nanjing, 210000, China

*Corresponding author: 2120222183@stu.nufe.edu.cn

Abstract. With the progressive development of the “sharing economy”, sharing bicycles are becoming increasingly popular in cities, and citizens’ last-mile travel problems are greatly alleviated. In addition, bicycles are also a good solution for short-distance travel. However, sharing bicycle operators do not have a good approach to dealing with supply and demand relationships, and bicycles are distributed in a position where supply is greater than demand, making it convenient for considerable discounts. The study aimed to establish a predictive model for predicting the short-term travel of bicycles to be shared in a particular area of the city. Shared bicycle data from 10 to 19 May 2017 were collected from the online platform of the Moby Algorithm Challenge, selecting data from land ranging from 152Mx152M, with the number of bicycles used reaching 2484561. ADF testing is designed to ensure that the parameter input in the model maximizes the accuracy of the model. The ARIMA model is then used to match the data and obtain a highly matching parameter model. As a result, the fixed value of the model is relatively close to the distribution state of the true value. The model predicts data for the next 12 days and obtains relatively accurate predictions.

Keywords: Shared bike; ARIMA; usage prediction.

1. Introduction

As a new travel mode of the "Internet + sharing economy", shared bicycles have become an important connection mode between rail transit and conventional public transportation under their low-carbon and environmental protection advantages, and solve the "last mile" travel problem. It has played an important role in alleviating urban traffic congestion, environmental pollution, and other problems. However, the inaccurate supply and demand scheduling of bicycles and the unreasonable built-up environment of the neighborhood are affecting the rational use and riding experience of shared bicycles, resulting in problems such as indiscriminate parking of shared bicycles and occupation of public resources.

To deal with the issue, many techniques are deployed for modeling the evolution of traffic circulation. In previous studies, machine learning such as random forest and gradient-boosted trees and other nonlinear models have been used to examine the impact of various influencing factors on shared bicycle riding. The prediction results cannot be effectively interpreted given the characteristics of machine learning "black box" operation. Wei et al. considered the influence of built-up environment interaction to predict the demand for bicycles and introduced bicycle density, based on the interaction of the number of bus stations and other traffic attributes, a prediction model of shared bicycle travel demand based on the Gradient Boosting Decision Tree (GBDT) model was proposed, and the model was explained with the help of the SHapley Additive explanation (SHAP) interpretation method, to improve the accuracy of the model [1]. Meanwhile, in time-series forecasting, deep learning is commonly employed. Demand forecasting for bicycle sharing is a spatiotemporal data forecasting challenge that incorporates both geographical and temporal variables. Kang et al. suggested a convolutional neural network prediction model after thoroughly taking into account the transportation network's spatial complexity, nonlinearity, and uncertainty [2]. This model disregards the time aspects while making good use of the traffic data's geographical information. As a consequence, Zhang et al. developed a prediction model based on convolution and residual networks, which

improves the accuracy of the prediction results by carefully taking into account time and spatial information [3].

Fu et al. employed long short-term memory (LSTM) and its variant network gated recurrent unit (GRU) to forecast short-term traffic flow for the temporal features [4]. Additionally, Yu et al. developed a traffic flow LSTM neural network forecast model and used LSTM and autoencoder to capture the temporal dependency of traffic prediction under harsh circumstances [5]. In addition, LSTM can identify the structure and pattern of data, can mine the nonlinearity and complexity contained in data, and is widely used in prediction research based on time series [6, 7].

However, when studying the scenario with latency-stringent, the traditional LSTM model has significant defects, Zhao et al. proposed the Random Connectivity Long Short-Term Memory (RCLSTM) model based on the traditional LSTM model. It is formed via stochastic connectivity between neurons. By revealing a certain amount of sparsity in this way, the RCLSTM model lowers computing costs and becomes more beneficial in applications with strict latency requirements. It is essential that the prediction accuracy of RCLSTM stays equal to that of the conventional LSTM, irrespective of changes made to the number of training samples or the length of input sequences [8]. In addition, there are breakthroughs in recurrent neural networks. Ma et al, from three perspectives of characteristics transmission method, spatial dimension, and characteristic dimension, discussed the direction of improvement of the structure of the enveloped neural network, gave some representative activation function contrast, gradient decrease algorithm, and its improved and adaptive optimization algorithm work principle and characteristics, clarified the application of enclosure neural networks in the field of intelligent transportation. They supported the enrolled neuron network algorithm with vector, differential integrated moving average return model, Karman filter, error wave reversal to the transmission of neurons, and long-time memory networks from the three main aspects of the application as well [9]. He et al. combined a linearity-based ARIMA model and a non-linearity-based Radial basis function (RBF) neural network and establish an ARIMA-RBF prediction model by the analysis of such time sequence characteristics as the weekly change and non-stability of passenger flow and the mechanism of ARIMA and RBF models [10]. Similar techniques have been used in the previously mentioned study to examine the demand for shared bicycles from the viewpoints of time and space.

This paper will predict the travel volume and delivery volume in the context of demand with ARIMA, compare experimental phenomena obtained by two different methods, then focus on the analysis and study of short-term travel volume and delivery volume of different construction lands, and put forward countermeasures and suggestions, to provide a basis for further in-depth research.

2. Methods

2.1. Data Source

The data in this article is sourced from the Mobike Cup Algorithm Challenge held by Mobike Company in 2017. Filter out the time period to be predicted and remove invalid values.

Table 1. Name and explanation of variables

Full name	Instruction	Data Type	Data Example
order ID	Order Number	Object	1893973
user ID	User Number	Object	451147
vehicle ID	Vehicle Number	Object	210617
vehicle category	Bicycle Type	Int	2
start time of cycling	Start Time	Object	2017-05-22 22:16:00

The dataset includes a total of 2484561 shared bicycle riding data from May 10th, 11th, 12th, 15th, 16th, 18th, and 19th (all on weekdays) in Beijing. The data contains 7 data fields: order ID, user ID, vehicle ID, vehicle category, start time of cycling, and starting and ending points of cycling. The starting and ending points are 7-digit Geohash codes, representing a space of approximately 152 m x 152 m, as shown in Table 1.

2.2. Method Introduction

In this paper, the author uses the ARIMA Model. Box-Jenkins introduced the ARIMA modeling idea, also known as the B-J method. The ARIMA modeling idea mainly includes the following key processes:

Step 1: The original time sequence is tested if it is stationary, and if the original series is stationary, the sequence can be transformed by d-grade differential transformation or other transformation (if the natural algebra changes are more common for the series) to turn the original sequence into a smooth sequence. An analysis of the original sequence or sequence after stabilization is performed, mainly by reference to the sequence's self-relevant and self-dependent functions, to observe whether the sequences imply seasonal changes. These analyses help to determine the shape of the ARIMA model. Step 2: The standard method for verifying sequence smoothness is unit root testing, commonly used to include Augmented Dickey-Fuller (ADF) test, Dickey-Fuller Test with GLS Detredding (DFGLS) test, Phillips–Perron test, Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test and No-deterministic polynomial time question (NP) test. Step 3: Estimate the parameters of the model, but also judge the model's compatibility and rationality. Step 4: Significant testing of the model residuals, mainly to test whether the residual sequence of the estimated results of a model meets the randomness requirement, i.e. if the residual sequence is a white noise sequence. Step: Use the final model to predict the original sequence and then evaluate the good and bad of the model.

3. Results and Discussion

3.1. Preliminary Work

The ARMA model is only suitable for smooth time sequences. For an irregular time sequence, a d-stage differentiation is required to make it a smooth sequence. To get the value of d, differencing is utilized first to establish the appropriate order of differencing. After that, the values of p and q are ascertained by figuring out the correct sequence for the moving regression and autoregression processes (Fig 1 and 2).

$$X_t = c + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}. \quad (1)$$

The first step in this model is to display the sequence and use the ADF test to measure the sequence's degree of stationarity to get the value of d. The non-stationary time series are then transformed into a stationary one based on the d value.

$$X_t = X_{t-1} + \omega_t, X_t(0, t\sigma^2). \quad (2)$$

$$\Delta X_t = X_t - X_{t-1} = \omega_t, U_t N(0, \sigma^2) \quad (3)$$

The p value can be determined using the partial autocorrelation function (PACF) plot's highest lag point. The value of q can be estimated using the autocorrelation function (ACF) plot's highest lag point.

$$\mathrm{PACF(k)} = \frac{E(Z_t - EZ_t)(Z_{t-k} - EZ_{t-k})}{\sqrt{E(Z_t - EZ_t)^2}\sqrt{E(Z_{t-k} - EZ_{t-k})^2}} = \frac{\mathrm{cov}[(Z_t - \overline{Z_t}),(Z_{t-k} - \overline{Z_{t-k}})]}{\sqrt{\mathrm{var}(Z_t - \overline{Z_t})}\sqrt{\mathrm{var}(Z_{t-k} - \overline{Z_{t-k}})}} \quad (4)$$

$$ACF(k) = \sum_{t=k+1}^{n} \frac{(Z_t - \overline{Z_t}),(Z_{t-k} - \overline{Z})}{\sum_{t=1}^{n} (Z_{t-k} - \overline{Z})^2} \quad (5)$$

After getting the three values, the ARIMA model can be fitted to predict the upcoming results. By comparing the forecast with the test data, the forecast model's accuracy can be assessed.

The t-value for these time series data is -9.132, and the p-value is 0.000 in the ADF test, as shown in Table 2. The critical values for 1%, 5%, and 10% are, respectively, -3.4948, -2.891, and -2.583. When the difference order is 0, the p value has already been 0 and the sequence is stationary. To make it easier to research, it suggests that a first-order differentiation on the sequence and a fresh ADF test are required.

Table 2. ADF test

Differencing Order	t	p	Critical Value		
			1%	5%	10%
0	-4.723	0.000	-3.512	-2.897	-2.586

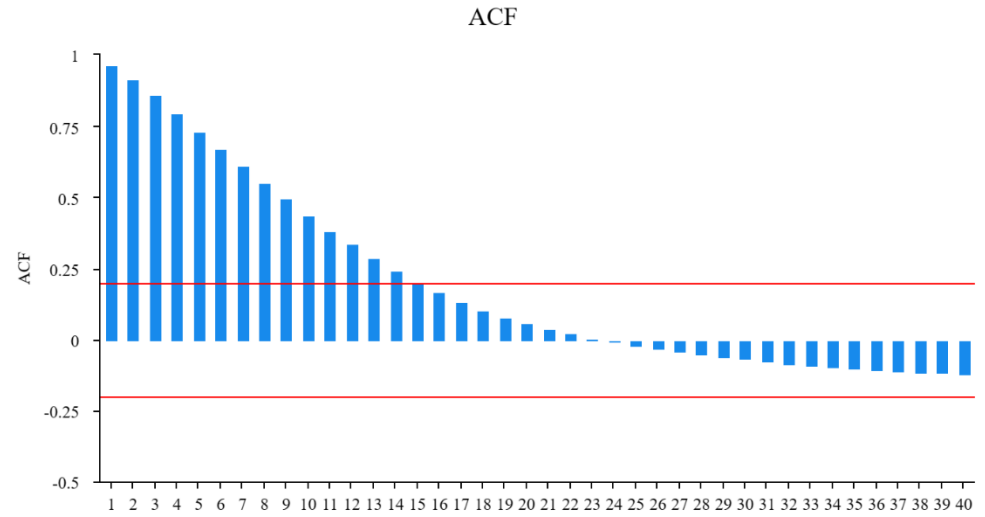

Fig. 1 ACF plot

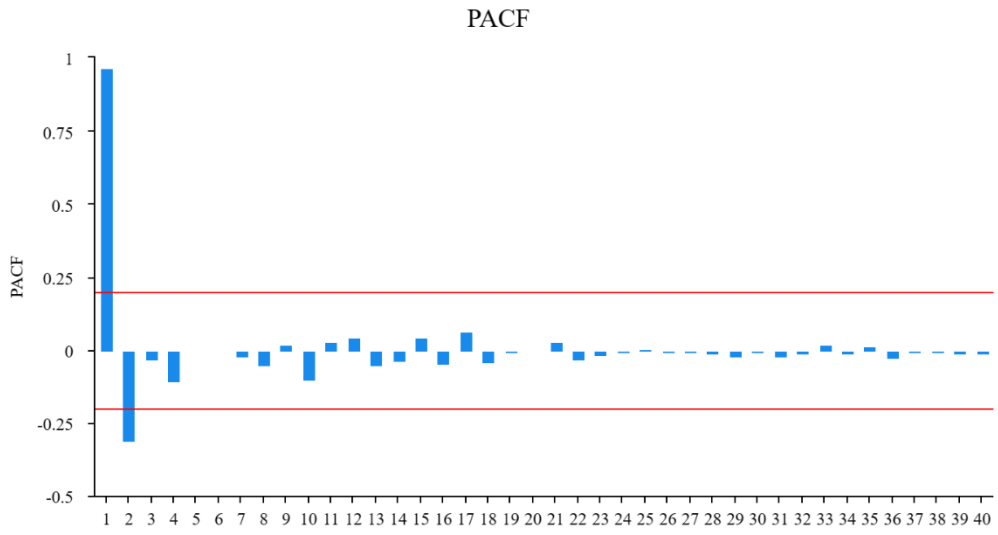

Fig. 2 PACF plot

3.2. Model Results

The Table 3 shows the model Q statistics information (specific for the Ljung-Box Q test statistics), including statistical measurements and p values: First, the ARIMA model requires the model residual to be white noise, i.e. the residual does not exist self-relevance, can be tested by the Q statistical test for white Noise (original assumption: residual is white noises). Second, for example, Q6 is used to test whether the corresponding coefficient of the 6 layers before the residues satisfies white noisy, usually its corresponding p value is greater than 0.1 indicates satisfying the White Noise test (on the contrary, it is not stated as a white noisiness), in the usual case, it can be analyzed directly against Q6; third: if the rejection of the White noise hypothesis ($p<0.05$), means that the model is not well matched, the opposite usually means the model can be used normally.

Table 3. ARIMA (2,2,0) model parameter

Items	Statistics	p-value
Q1	2.570	0.109
Q2	4.862	0.088*
Q3	6.385	0.094*
Q4	6.633	0.157
Q5	8.312	0.140
Q6	15.686	0.016**
Q7	20.171	0.005***
Q8	20.756	0.008***
Q9	22.562	0.007***
Q10	22.589	0.012**
Q11	25.727	0.007***
Q12	29.171	0.004***
Q13	29.649	0.005***
Q14	29.649	0.009***
Q15	29.672	0.013**
Q16	29.842	0.019**
Q17	30.293	0.024**
Q18	31.103	0.028**

According to the ARIMA (2,2,0) model parameter table, the model formula is

$$y(t) = 6.932 + 1.711 * y(t-1) * y(t-2) \quad (6)$$

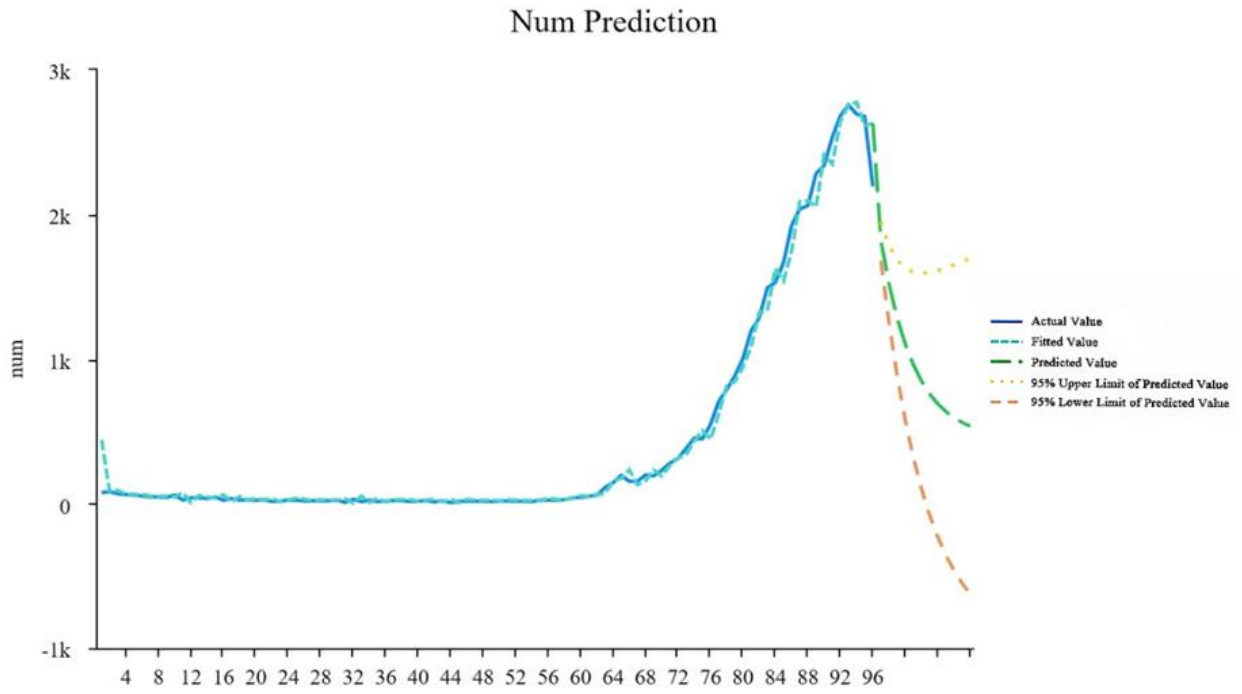

Fig. 3 Number Prediction

As can be seen from Fig 3 above, the model is effective in predicting the trips of shared bicycles in a short period of time. The fitting degree of the numerical model is high and close to the distribution state of the true value. Therefore, this model can meet the prediction requirements and can be used for the prediction of the quantity of shared bicycle usage in the plot in the next period. Table 4 shows the details

Table 4. Predicted Value (12 Phase)

Prediction	1	2	3	4	5	6	7	8	9	10	11	12
Values	1814	1517	1284	1101	958	846	758	689	636	593	560	534

The first analysis is Mean Square Error (MSE): 6143.5454, which is the average squared distance between the observed and predicted values (Table 4). Because it uses squared units rather than natural data units, the interpretation is less intuitive. This paper obtains the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE): the values are 78.3808 and 39.0501. RMSE measures the average difference between a statistical model's predicted values and the actual values, and quantifies how dispersed these residuals are, revealing how tightly the observed data clusters around the predicted values. In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. The values prove the model is effective

4. Conclusion

The study constructed an ARIMA model based on available data to predict the passenger traffic of shared bikes in a short time and obtained relatively accurate results. This demonstrates that it is feasible to build an analytical prediction model using predictions for future shared bikes, with the larger the input of early data, the higher the degree of matching between the model and actual data, and the more accurate the prediction results. But the need is that in the process of collecting data, to focus on the filtering of the data, for the ARIMA model, requires big data as the test object, the results will be more accurate and more relevant. By predicting future short-term passenger flows, sharing bicycle operators can help make decisions when placing bicycles, and improve the placement and use efficiency of shared bikes, while also facilitating people's travel needs.

References

[1] Wei Jin, An Shi, Zhang Yantang. Prediction of shared bicycle demand based on environment factor interactions Science Technology and Engineering, 2023, 23(26): 11424-11430.

[2] Kang Y, et al. Deep spatial-temporal modified-inception with dilated convolution networks for citywide crowd flows prediction International Journal of Pattern Recognition and Artificial Intelligence, 2020.

[3] Zhang X, et al. Urine sediment recognition method based on multi-view deep residual learning in the microscopic image, Journal of Medical Systems, 2019, 43(11): 325-410.

[4] Li R, Fu Z, Zhang L. Using LSTM and GRU neural network methods for traffic flow prediction, 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), 2016.

[5] Li Yaguang, et al. Deep Learning: A Generic Approach for Extreme Condition Traffic Forecasting, SIAM International Conference on Data Mining (SDM), 2017.

[6] Xu C, Ji J, Liu P. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets, Transportation Research Part C: Emerging Technologies, 2018.

[7] Cao Dandan, et al. Short-term demand forecasting of shared bicycles based on a long short-term neural network model. Science Technology and Engineering, 2020, 20(20): 8344-8349.

[8] Hua Y, et al. Deep learning with long short-term memory for time series prediction, IEEE Communications Magazine, 2019, 56(6): 114-119.

[9] Ma Yongjie, CHENG Shi-sheng, MA Yun-ting, MA Yide. Review of convolutional neural network and its application in intelligent transportation system, Journal of Traffic and Transportation Engineering, 2021, 21(4): 48-71.

[10] He Jiuran, Si Bingfeng. Application of an ARIMA-RBF model in the forecast of urban rail traffic volume. Shandong Science, 2013, 26(3): 75-81.