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

Research on the Urban Bike-sharing Usage based on ARIMA Model

Hao Cheng [1], Muze Li [2] and Haoting Zhang [3, *]

[1] School of Transportation Engineering, Chongqing Jiaotong University, Chongqing, 400074, China

[2] Institute of Smart City and Intelligent Transportation, Southwest Jiaotong University, Chengdu, 611756, China

[3] College of Transportation Engineering, Tongji University of Technology, Shanghai, 201800, China

* Corresponding Author Email: 2253447@tongji.edu.cn

Abstract. This study forecasts the demand for shared bicycles using the ARIMA model, incorporating insights from extensive literature on urban bike-sharing usage. Through in-depth analysis, first-order differencing was identified as crucial for achieving stationarity, leading to the recommendation of the ARIMA (312) model. This model effectively encapsulates the dynamics of shared bicycle usage, evident from the significant ADF test results and the careful selection of model parameters based on ACF and PACF plots. The research reveals the complex relationship between urban mobility and shared bicycle systems, providing a comprehensive framework for predicting usage trends. These findings make a significant contribution to the discussion on sustainable urban transportation and offer practical guidance for city planners and bike-sharing operators to enhance service efficiency and meet user demands effectively. The precise prediction of shared bicycle demand by the ARIMA (3, 1, 2) model highlights the effectiveness of advanced time series analysis in understanding and predicting bike-sharing usage patterns. According to the predictions, there are clear seasonal fluctuations, with a cycle of four quarters during which the predicted values gradually increase. This study emphasizes the potential of shared bicycles to enhance urban mobility and points out the need for more regulated development and management strategies to address challenges posed by the rapid growth of the bike-sharing industry. By providing a detailed understanding of the factors influencing shared bicycle usage, this research contributes to optimizing bike-sharing systems, thereby aiding in the sustainability of urban transportation networks.

Keywords: ARIMA model; urban bike-sharing; station planning.

1. Introduction

The rise of shared bicycles began in early 2016 with the launch of Mobike and Ofo, quickly growing due to high demand. Initially, demand forecasts used time series analysis and historical data, but these methods often missed changing user behaviors. Researchers have since adopted advanced models like activity-based Autoregressive Integrated Moving Average Model (ARIMA) and non-homogeneous Markov models to better understand the dynamic demand for shared bicycles. Recent progress in these predictive techniques has provided deep insights into shared bicycle usage across urban settings, employing various analytical methods to predict travel patterns and address specific challenges with advanced data analysis.

Yu et al. utilized shared bicycle travel data from the Boston area to analyze the spatiotemporal differentiation characteristics of shared bicycle riding. They employed GIS and Python, enhancing traditional built environment analysis with refined design indicators such as road direction entropy and section length. Through the use of the Generalized Additive Mixed Model (GAMM), they explored the interactive relationship between shared bicycle ridership and the built environment, offering insights into how urban design influences bike-sharing usage patterns [1]. In a separate study, Chen focused on Chengdu's central urban area, analyzing the spatiotemporal characteristics of shared bicycle usage. They applied Geographically Weighted Regression (GWR) to identify the factors influencing shared bicycle travel, providing a localized perspective on bike-sharing dynamics [2]. Yang and Jin concentrated on predicting the volume of shared bicycles around subway stations. They

developed a ridge regression-oriented model using relevant computer software, aiming to forecast bike-sharing volumes in subway station areas. This research offered a tool for planning and managing shared bicycle fleets, especially in transit-oriented urban sections [3]. Wang and Dai chose the DBSCAN clustering method for its suitability in spatial clustering over the K-means algorithm. Their study on the regional flow of shared bicycles, based on order data analysis, contributes to understanding the spatial distribution and movement patterns of bike-sharing services [4].

Wei et al. investigated how micro and macro-environmental factors within Beijing's fifth ring influence bike-sharing, using models like global regression (GR), geographically weighted regression (GWR), and multi-scale geographically regression (MGWR) for a detailed analysis of riding densities [5]. Zhang and Rao created a visual analysis system for Hangzhou's shared bikes, combining various predictive models with visual techniques for better decision-making [6]. Additionally, Ke and Wu used spatiotemporal models and combined KNN with Light Gradient Boosting Machine (LightGBM), enhanced by Principal Component Analysis (PCA) for feature extraction, to predict rental demand, offering insights into demand dynamics and improving upon traditional machine learning methods [7]. These studies collectively advance the field of urban bike-sharing research, offering diverse methodologies for analyzing and predicting shared bicycle usage patterns across different urban contexts. Influenced by shared bicycles, interdisciplinary research has emerged. He and Wang tackled high parking pressure by gathering data through various methods to propose parking optimizations based on user needs and urban design [8]. Liu and Ma analyzed choices between shared bicycles and feeder buses, using a model to enhance bus route design and efficiency, ultimately aiming to boost occupancy and lower costs [9].

However, despite the thriving development of shared bicycles, there are still some problems. As the number of investors grows, an increasing number of shared bicycle companies have emerged, with each company continuously adding more bicycles to the city. This has led to a mismatch between the regulatory authorities' management and the companies' own development, highlighting the drawbacks of shared bicycles. Wu et al. analyzed these drawbacks, noting that the development speed is too fast while management remains lax, shared bicycles often lack licenses, leading to insufficient supervision of user behavior, refunds of deposits are challenging, as the scale of development expands, operating companies frequently struggle to break even, and it affects the city's appearance [10]. These observations suggest the need for more regulated development and management strategies to address the challenges posed by the rapid expansion of the shared bicycle industry, ensuring its sustainable growth and integration into urban mobility systems.

Zhou discussed the current situation and strategies for the operation and development of internet shared bicycles. It highlights the success of shared bicycles in urban mobility but also addresses the challenges such as management difficulties, inadequate infrastructure, and product homogenization. The paper suggests improving technical means for vehicle management, enhancing cooperation with municipal departments, and innovating to overcome product similarity. It advocates for a "government + enterprise" operation model to enhance profitability and ensure sustainable development, emphasizing the long-term goal of serving society and advancing social progress [11].

Drawing on insights and expansions from relevant literature, this study opts for time series analysis prediction methods to forecast the demand for shared bicycles using their usage data. It aims to uncover users' travel patterns, identify the spatial and temporal distribution of shared bicycles in the region, and devise a planning and scheduling scheme for shared bicycle stations. This approach provides a data foundation for operators' scheduling activities and the recovery of faulty shared bicycles, ensuring efficient operation and maintenance.

2. Methods

2.1. Data Source

Firstly, this paper selected predicting the demand for shared bicycles as the research topic. To obtain a dataset with high feasibility and easy computation, Kaggle was accessed, and a dataset was found, including trip ID, bike ID, time, station name, and the longitude and latitude of the starting and ending points from 2014 to 2023. Moreover, this dataset is based on the Chicago Divvy website, containing all usage data for Divvy shared bikes in Chicago from 2014 to 2023.However, this data set is two-dimensional and involves a high degree of computational and programming difficulty. This paper decided not to use longitude and latitude as the modeling index, but instead of the usage amount of shared-bikes as the main indicator of the study.

2.2. Indicator Selection

The data was processed in the following manner: After removing invalid data, a quarter was used as a time node to count the recorded data amount for each period. A one-dimensional Excel spreadsheet was created, detailing the usage of shared bikes from the first quarter of 2014 to the fourth quarter of 2023.

In this paper, the total usage of bike-sharing per quarter is selected as the indicator for analysis, aiming to forecast its changes in the upcoming years. By utilizing the quarterly total usage of bike-sharing as the indicator, the seasonal variations of this data can be clearly observed. This information can then be used as a basis for optimizing the seasonal scheduling and allocation of bike-sharing resources in practice (Table 1).

Table 1. Data introduction.

Field	Instruction	Data type	Example
Ride id	Bicycle ID	Object	C2F7DD78E82EC875
Rideable type	Bicycle type	Object	Electric bike
Started at	Start time	Object	2022/1/13 11:59
Ended at	End time	Object	2022/1/13 12:02
Start station name	Start position	Object	Glenwood Ave & Touhy Ave
Start station id	Start position ID	Object	KA1504000151
End station name	End position	Object	Paulina St & Montrose Ave
End station id	End position ID	Object	TA1309000021
Member or casual	Member type	Object	casual

2.3. Method Introduction

In terms of methodology, the simple seasonal time series forecasting was utilized in this paper. After the model established, some technical and famous standards are used to evaluate the model, including: R-squared, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Bayesian Information Criterion (BIC). Based on the evaluation conducted on these standards, we will adjust the model until it is ultimately deemed to be precise, with minimal error, and effective.

The ARIMA model, standing for AutoRegressive Integrated Moving Average, combines three main components for time series forecasting: autoregression (AR), differencing (I) to make the series stationary, and moving average (MA). Mathematically, an ARIMA model is denoted as ARIMA (p, d, q), where 'p' is the order of the AR terms, 'd' is the degree of differencing needed to make the series stationary, and 'q' is the order of the MA terms. This model predicts future values based on past values (AR), the differences between past values (I), and forecast errors (MA).

$$Y(t) = c + \varphi_1 * Y(t-1) + \varphi_2 * Y(t-2) + \cdots + \varphi_p * Y(t-p) + \varepsilon(t) \quad (1)$$

3. Results and Discussion

3.1. ADF Test

The Augmented Dickey-Fuller (ADF) test results show that the original data (without differencing) has an ADF statistic of -1.000147 with a p-value of 0.753211, indicating the data is non-stationary. However, after first differencing, the ADF statistic is -3.613147 with a p-value of 0.005515, which is significantly below the 1% significance level, suggesting the differenced data is stationary (Table 2). Therefore, this time series data should undergo first-order differencing for subsequent ARIMA modeling.

Table 2. ADF test.

Differencing Order	t	p	Critical Value (1%)	Critical Value (5%)	Critical Value (10%)
0	-1.000147	0.753211	-3.632743	-2.948510	-2.613017
1	-3.613147	0.005515	-3.632743	-2.948510	-2.613017

3.2. ACF and PACF Test

The ACF and PACF plots suggest that the time series data is characterized by autoregressive and moving average components. The ACF shows significant correlations extending beyond the confidence interval, while the PACF declines towards zero after initial lags, indicating that a mixed ARMA model may be appropriate. SPSS software recommends an ARIMA (3, 1, 4) model where the data's autoregressive component is captured by three lags, differencing of order one is required for stationarity, and four lags are used in the moving average component (Figure 1 and 2).

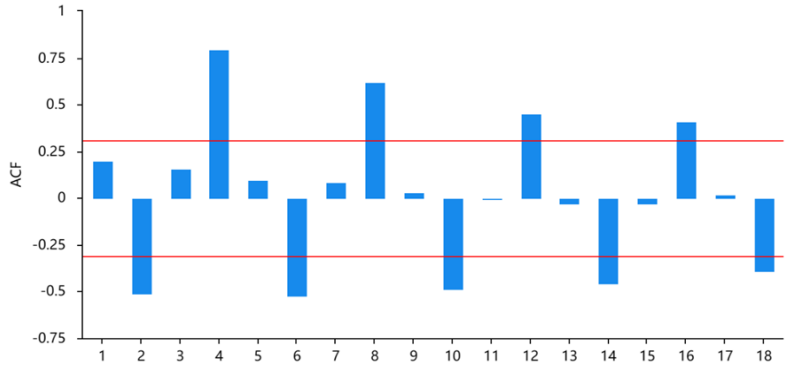

Fig. 1 ACF plot.

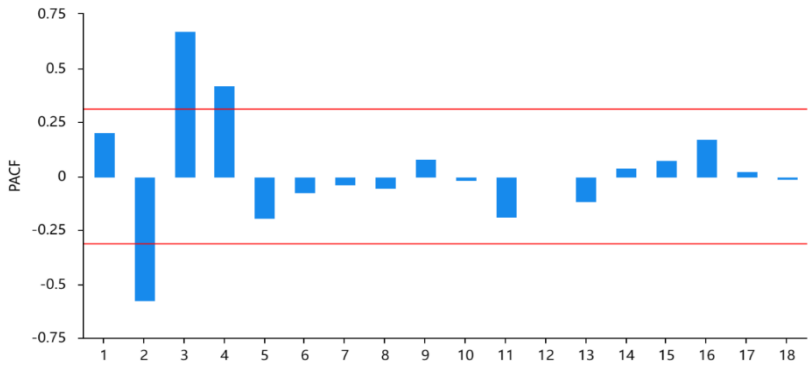

Fig. 2 PACF plot

3.3. ARIMA Model Results

The table 3 illustrates the results of the model construction, including regression coefficient values and p-values: Firstly, the model parameters table shows the results of the model construction, which typically do not require much attention, even if p-values are greater than 0.05. Secondly, the information criteria AIC and BIC values are used for comparing models across multiple analyses; lower values are preferable. If multiple analyses are conducted, one can compare the changes in these two values to comprehensively explain the optimization process of model construction.

Table 3. ARIMA (3, 1, 2) model parameter.

Item	Sign	Coefficient	Standard Error	Z value	P value	95%CI
Constant	c	43396.670	73100.110	0.594	0.553	-99876.913-186670.254
AR Parameter	α1	-0.293	0.258	-1.138	0.255	-0.798 ~ 0.212
	α2	-0.995	0.012	-84.926	0.000	-1.018 ~ -0.972
	α3	-0.302	0.246	-1.228	0.220	-0.783 ~ 0.180
MA Parameter	β1	-0.278	0.300	-0.926	0.354	-0.866 ~ 0.310
	β2	0.766	0.187	4.088	0.000	0.399 ~ 1.133

AIC value: 1080.088, BIC value: 1091.733

For the "Total shared bicycle usage," using the AIC information criterion (where a lower value is better), SPSSAU automatically models and compares several potential candidate models. It ultimately identifies the optimal model as ARIMA (3, 1, 2), with the formula:

$$y(t) = 43396.670 - 0.293y(t-1) - 0.995y(t-2) - 0.302y(t-3) - 0.278\varepsilon(t-1) + 0.766 * \varepsilon(t-2) \quad (2)$$

3.4. Model's Q Statistic

The table 4 displays the model's Q statistic information (specifically the Ljung-Box Q test statistic), including the statistic value and p-value; Firstly, the ARIMA model requires that the residuals are white noise, meaning there's no autocorrelation, which can be verified through the Q statistic test (null hypothesis: residuals are white noise); Secondly, for instance, Q6 tests whether the first six autocorrelation coefficients of residuals satisfy the white noise condition, with p-values greater than 0.1 typically indicating the condition is met (otherwise, it's not white noise). It's common to analyze Q6 directly; Thirdly, if the white noise assumption is rejected ($p<0.05$), it implies poor model fit, whereas acceptance generally means the model is fit for use.

Table 4. Model's Q statistic table

Item	Statistic	P value	Item	Statistic	P value
Q1	1.092	0.296	Q9	2.829	0.971
Q2	1.101	0.577	Q10	3.217	0.976
Q3	2.164	0.539	Q11	3.292	0.986
Q4	2.184	0.702	Q12	3.521	0.991
Q5	2.327	0.802	Q13	4.004	0.991
Q6	2.334	0.887	Q14	4.971	0.986
Q7	2.346	0.938	Q15	5.074	0.991
Q8	2.392	0.967	-	-	-

Based on the Q statistic results, with a p-value for Q6 greater than 0.1, the null hypothesis cannot be rejected at the 0.1 significance level. This indicates that the residuals of the model are white noise, and the model generally meets the requirements.

3.5. Model Fitting and Prediction

According to the fitting and prediction results of the numerical model shown in Figure 3, the degree of fit of the numerical model is high and close to the distribution state of the true values. Therefore, this model can meet the prediction requirements and can be used for predicting the usage of shared bicycles in this plot in the next period.

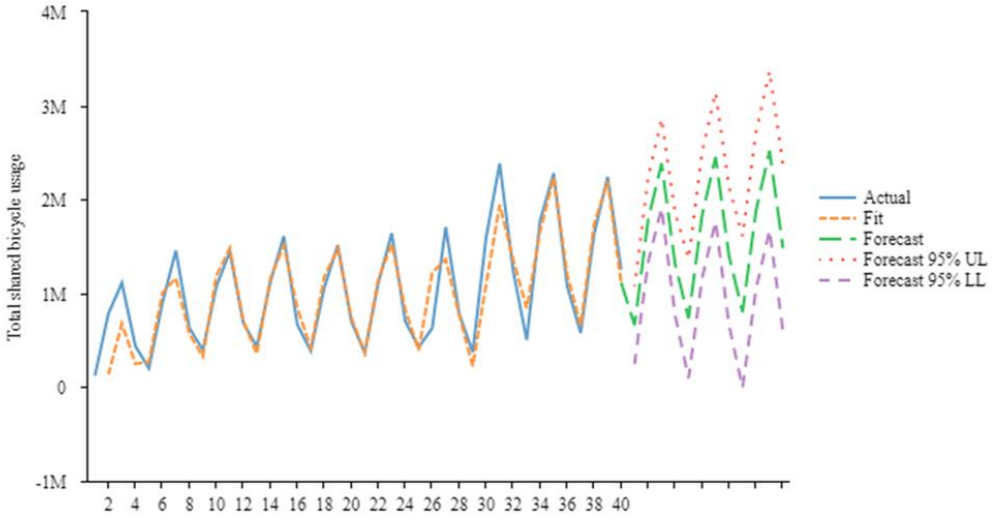

Fig. 3 Model fitting and prediction

Through calculation, we predict the value of shared bike usage in 12-time nodes in the future. According to the predicted values in Table 5, it can be seen that the predictions exhibit clear seasonal fluctuations, with a cycle consisting of four stages, during which the predicted values gradually increase.

Table 5. Prediction value

Prediction	Value	Prediction	Value
T=1	660270.176	T=7	2451622.490
T=2	1782770.116	T=8	1395883.242
T=3	2383721.109	T=9	800842.465
T=4	1312654.110	T=10	1881947.972
T=5	733840.874	T=11	2518600.285
T=6	1831282.323	T=12	1479221.781

4. Conclusion

In conclusion, this study leverages ARIMA modeling to analyze and forecast the demand for shared bicycles, incorporating insights from extensive literature on urban bike-sharing usage. Through meticulous analysis, it was determined that a first-order differencing is essential for stationarity, leading to the recommendation of an ARIMA (3, 1, and 2) model. This model adeptly captures the dynamics of shared bicycle usage, underscored by significant ADF test results and the discerning selection of model parameters based on ACF and PACF plots. The findings illuminate the complex interplay between urban mobility and shared bicycle systems, providing a robust framework for predicting usage patterns. This research not only contributes to the academic discourse on sustainable urban transportation but also offers practical insights for city planners and bike-sharing operators aiming to enhance service efficiency and meet user demand effectively.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Yu Bingjie, Liang Yuan, Yang Linchuan. Exploring the relationship between bike-sharing ridership and built environment characteristics: A case study based on GAMM in Boston. World Regional Studies, 2023, 32(2: 48-58.

[2] Chen Binglang. Study on the Use Characteristics and Influencing Factors of Shared Bicycles Based on Spatiotemporal Data-A Case Study of Chengdu City. Urban Construction, 2022, 17: 19.

[3] Yang Xinyu, Jin Qun. Demand Forecasting of Shared Bicycles in Subway Station Areas Based on Machine Learning. Journal of Shijiazhuang Tiedao University (Natural Science Edition), 2023, 36(3): 92-98.

[4] Wang Jianhua, Dai Yizhou. Study on the Distribution and Scheduling of Shared Bicycle Regional Flow Based on Density Clustering. Business and Management, 2023, 8: 46-53.

[5] Wei Jiaomin, Liu Zhuo, Chen Yanyan, et al. Analysis of the Influencing Factors of Shared Bicycle Riding Considering Macro and Micro Built Environment. Science Technology and Engineering, 2023, 23(9): 3904-3915.

[6] Zhang Qiqi, Rao Ning, Zhu Sujia, Cha Meng, Sun Guodao. Multi-model Visual Comparative Analysis for Shared Bicycle Demand Prediction. High Technology Letters, 2023, 33(12): 1323-1332.

[7] Ke R H, Wu S, Ke W W. A spatial-temporal model for identifying tidal shared-bicycle stops and a borrow-return demand prediction method based on KNN-LightGBM. Journal of Geo-information Science, 2023, 25(4): 741-753.

[8] He Jing, Wang Zhirui. Research on Urban Street Landscape and Parking Lot Design under the Development of Shared Bicycles. Industrial Design, 2022, 12: 94-96.

[9] Liu Lumei, Liu Zhengke, Ma Changxi, et al. Design and Vehicle Configuration Method of Feeder Bus Routes under the Impact of Shared Bicycles. Journal of Transportation Systems Engineering and Information Technology, 2023, 23(1): 165-175.

[10] Wu Yunfan. Ai Huiting, Zhu Daigen. Problems and Suggestions on Shared Bicycles. Industrial Innovation Research, 2023, 6: 78-80.

[11] Zhou Congcong. Current Operation Status and Countermeasures of Internet Shared Bicycles. Journal of Jinhua Polytechnic, 2023, 23(2): 23-29.