Spatial and Temporal Characteristics Analysis and Demand Forecasting based on ARIMA Model - An Example of Yellow Taxi in New York

Siyuan Wang*

Chang'an Dublin International College of Transportation, Changan University, Xi'an, 710064, China

* Corresponding Author Email: siyuan.wang@ucdconnect.ie

Abstract. With the development of technology and urbanization, cars have greatly facilitated people's travel, but they also bring problems such as road congestion, pollution and accidents, which restrict the healthy development of cities. How to predict traffic flow efficiently and accurately has become a hot topic. Taxi data is highly recommended in urban transportation research due to the characteristics of broad spatial coverage, long driving time and relatively complete data. In this paper, the data of yellow taxis in New York during March 2023 are used to present the demand changes by means of line chart thermal map, and divide the database of March into two parts, one is training set and the other one is test set, then simulation and analysis using ARIMA model. It reveals the changing relationship between taxi demand and time. By predicting the traffic flow, this study can alleviate the congestion of urban main roads and improve the quality of citizens' life. At the same time, forecasting commuting demand can balance supply and demand, provide more efficient services for operating companies, taxi drivers and passengers, and provide decision-making basis for government planning.

Keywords: ARIMA model, traffic flow prediction, New York yellow taxi.

1. Introduction

With the rapid technological and economic development brought about by the Industrial Revolution, urbanization is accelerating, and the automobile has become one of the necessary tools for people's daily life, which has greatly changed the way people travel [1]. Transportation is an important basis for keeping cities running and people living normally, but the rapid development of urban infrastructure construction and the continuous inflow of population gradually increase the pressure of traffic, making various traffic problems increasingly prominent. Traffic brings road congestion, pollution and accidents, which has become one of the bottlenecks of urban livelihood, economic and social development. It also restricts healthy development of cities [2].

Traffic flow in urban areas is typical spatio-temporal data, and forecasting and analyzing the demand and condition of historical traffic data is an important method to solve such traffic problems [3]. Traffic flow prediction uses intelligent computing methods to predict future traffic states based on historical and existing traffic data, and then recommends better routes for travelers and predicts travel times based on the results to achieve the goal of congestion avoidance and time saving. Short-time traffic flow prediction algorithms include two categories: basic algorithms and optimized combination algorithms [4, 5]. The basic algorithm is simple but low accuracy, while the optimal combination algorithm is high accuracy but long and complex [1]. In the traffic flow prediction problem, the current research results are mainly based on linear models, nonlinear models, deep learning models, traditional traffic assignment models and simulation models, and statistical theory methods such as time series method and regression analysis are used to predict short-term passenger flow [6, 7]. Using time series models, the characteristics of changes, trends and patterns of development of each variable can be found, and their future changes can be predicted accordingly.

Data of taxi is often used in urban traffic research due to broad spatial coverage, long driving time and relatively complete data. Those make them widely used to analyze traffic patterns and trends of cities [2, 5]. New York, for example, has one of the busiest taxi systems in the world. The majority of New York's population relies on public transportation. A research estimated that 54% of the

population does not have a car or any private vehicle, which results in nearly 200 million taxi trips per year [8]. There are approximately 13,000 taxis in New York, consisting of yellow and green taxis. Yellow taxis (Y-taxi) vehicles can pick up passengers anywhere in Bronx, Brooklyn, Manhattan, Queens, Staten Island of New York. Green taxis, on the other hand, are regulated to pick up passengers only in Upper Manhattan, the Bronx, Queens, and Staten Island [9]. As an important part of urban transportation, taxis can be used to analyze taxi trajectory data to understand traffic flow patterns and forecast traffic flow, which in turn can improve urban traffic conditions, solve the problem of congestion on main roads in cities, estimate the travel demand of residents, and improve the life standard of people [10,11]. Also predicting commuting demand can balance supply and demand, then a certain number of vehicles can be dispatched to the area in advance to increase the number of taxis in the area, facilitate drivers to find passengers faster, reduce unnecessary energy waste from idling for taxi drivers and other practitioners, and reduce air pollution [12, 13].

This paper will use a classical autoregressive integrated moving average (ARIMA) time series model to analyze the taxi trips demanded in the region through time correlation using public data of taxi demand situation in New York City in March 2023 to provide more efficiency for operating companies, taxi drivers and passengers, and provide a quality analytical decision basis for government planning.

2. Methods

2.1. Data Sources

This study used yellow taxi data provided by the Taxi and Limousine Commission (TLC) of New York City for analysis. This dataset includes 3.4 million pieces of information about yellow taxis in New York City. The time span covers traffic information from March 1 - March 31 in 2023, and covers the five boroughs of New York City. Includes information on each taxi's the date and time of boarding and alighting, the coordinates of the starting and ending locations, distance of the trip, itemized fares, rate types, payment types and the number of passengers reported by the driver. The analysis in this paper focuses on the analysis of traffic demand, so only five pieces of information are retained: the date and time of boarding and alighting, the coordinates of the starting and ending locations, and the distance of the trip for each taxi, as Table 1 shows.

Table 1. Data records dictionary of Y-taxi

Field Name	Description	Example
tpep_pickup_datetime	The date and time when the meter was engaged.	2023/3/1 0:08
tpep_dropoff_datetime	The date and time when the meter was disengaged.	2023/3/1 0:39
Trip_distance	The elapsed trip distance in miles reported by the taximeter.	12.4
PULocationID	TLC Taxi Zone in which the taximeter was engaged.	50
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged.	104

To ensure data accuracy and completeness, this paper first cleansed and pre-processed the data. Large amounts of redundant data increase memory consumption, raise costs and reduce model quality; incorrect data may make the final results inaccurate if not processed correctly. For the algorithm to be effective, it must be provided with clean, accurate and concise data. So unreasonable data and data that do not meet the requirements, such as travel distance less than 0 and data that are not from March 2023, are removed from the database, as Table 2 shows. All the demands from March 1 to March 31 are divided by every 5-minute slice, and the time series from March 1 to March 25 are selected as the training set and March 26 to March 31 as the test set.

Table 2. Example data after pre-processing

tpep_pickup_datetime	tpep_dropoff_datetime	trip_distance	PULocationID	DOLocationID
2023/3/1 0:00	2023/3/1 0:09	1.2	164	50
2023/3/1 0:00	2023/3/1 0:05	1.8	237	75
2023/3/1 0:00	2023/3/1 0:22	5.56	249	181
2023/3/1 0:00	2023/3/1 0:20	5.15	263	223
2023/3/1 0:00	2023/3/1 0:07	1.87	237	161

2.2. Preliminary Analysis of Data

After cleaning and pre-processing the taxi big data, the first step is to extract and analyze the operation and features of New York City yellow taxis for the dimensions of time and space. Figure 1 plots a line graph of taxi demand fluctuations for each day from March 1 to 31. The graph shows the fluctuation of taxi demand over time during this period. The whole month is roughly divided into four cycles, with peaks and troughs alternating in a cyclical manner, and the weekly trends and patterns are roughly the same.

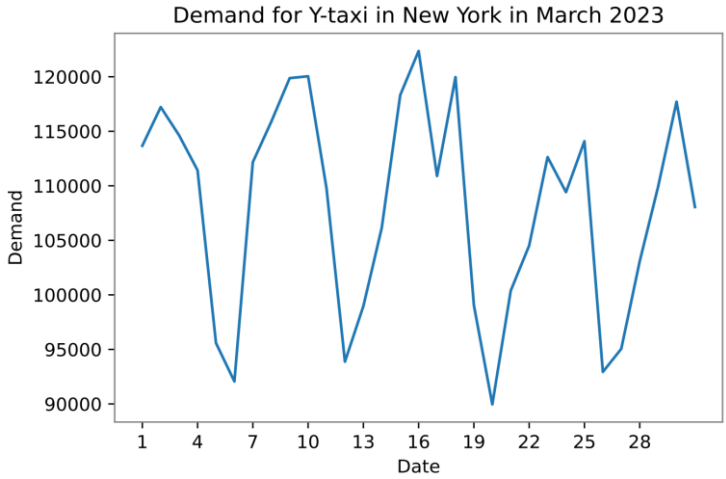


Fig. 1 Demand for Y-taxi in New York in March 2023

In Figure 2, a histogram of the frequency distribution of travel distance is shown, which shows the frequency of travel in different distance ranges. By looking at this histogram, the preferred travel distance of passengers in this area is generally within 3miles, with short trips being the main focus.

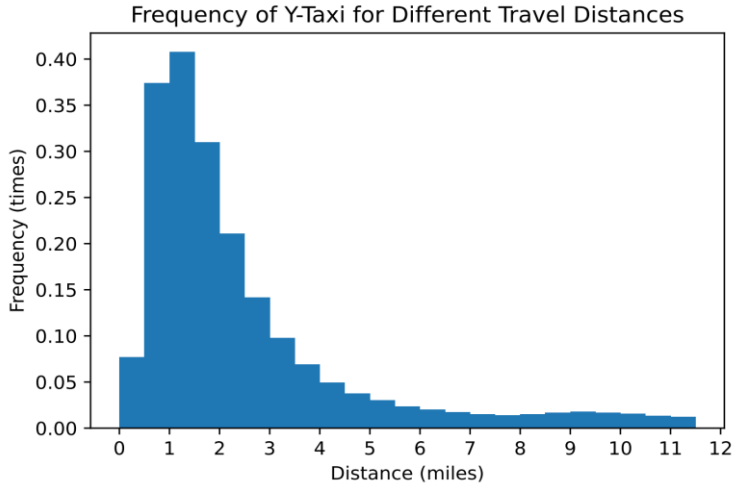


Fig. 2 Frequency of Y-taxi for different travel distances

The two graphs in Figure 3 show the origin and destination of trips. By identifying the popular origin and destination areas, we can find that the origin of passengers' boarding is scattered and difficult to concentrate. However, the destinations are mainly concentrated in a few popular areas

such as the northeast. This is very important information for taxi companies to adjust their operation strategy according to the demand of popular areas, such as increasing the distribution of vehicles in these areas to reduce passenger waiting time.

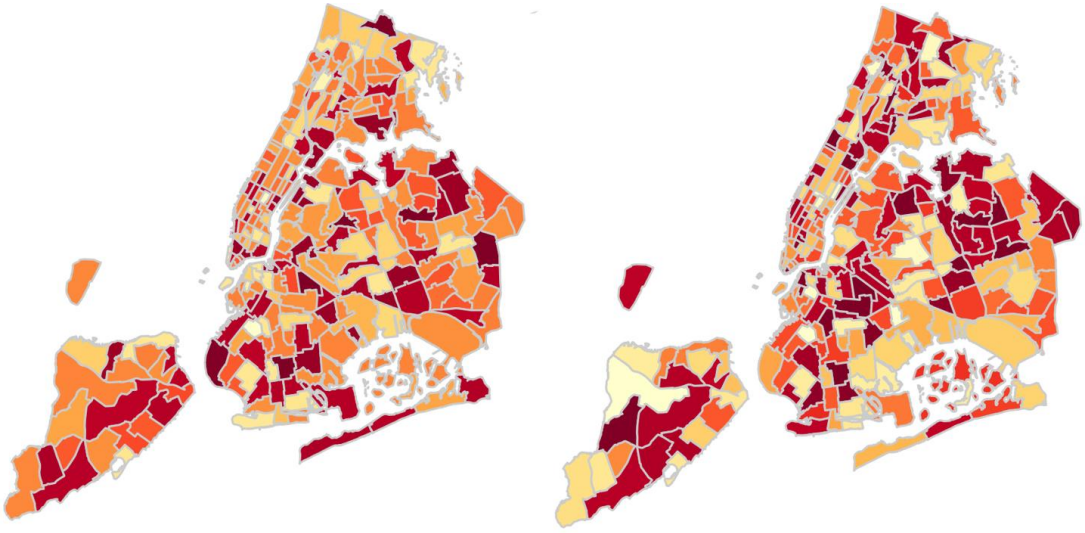


Fig. 3 Starting point and destination of Y-taxi demand

2.3. Model-related Theory

ARIMA model is a classical time series forecasting model, which is widely used in practical applications. It is simple, intuitive and easy to interpret, and can be applied to many different kinds of time series data. The model first transforms the predicted object sequence into a completely random sequence, and later simulates and approximates the sequence by combining different mathematical models [12].

The ARIMA model is fitted by combining the characteristics of the autoregressive (AR), integrated (I) and moving average (MA) models. Different parameters in ARIMA(p, d, q) represent different characteristics, in which "p" infers the autoregressive order, "d" shows the difference order and "q" means the moving average order. According to the trends of the time series data, the ARIMA model can be constructed by determining the appropriate values step by step according to the process (Fig. 4).

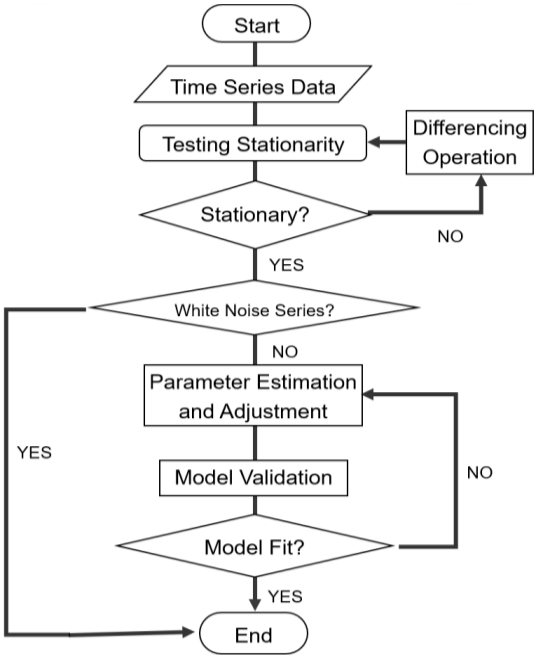


Fig. 4 Process of ARIMA model

The ARIMA model algorithm is to first perform the difference operation on the data to turn the unstable time series data into stable time series data. When the time series tends to increase or decrease, a differencing operation is required until the series passes the smoothness test. Where the number of differences is the d-order in the model ARIMA(p, d, q). The formulas for the difference operation are shown below:

$$\nabla x_t = x_t - x_{t-1} = x_t - Nx_t = (1-N)x_t \tag{1}$$

$$\nabla^2 x_t = \nabla x_t - \nabla x_{t-1} = (1-N)x_t - (1-N)x_{t-1} = (1-N)^2 x_t \tag{2}$$

$$\nabla^d x_t = (1-N)^d x_t \tag{3}$$

Where N is the backward shift operator, so the formula for the d-order difference sequence is:

$$S_t = \nabla^d x_t = (1-N)^d x_t \tag{4}$$

After conversion to a smooth series, the values of p and q in the model ARIMA(p, d, q) need to be determined. At this point the autoregressive moving average (ARMA) model is to process the differenced time series, which consists of AR model and MA model.

AR model is used to establish the relationship between the current and the past moment of the time series. The principle is to use the historical data to predict future of the variables themselves. In the AR(p) model, "p" denotes the steps of past time, which determines the number of past observations to be considered in the model.

The general p-order AR model is:

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + u_t + \delta \tag{5}$$

where X_t is the predicted value at the t time point, a_i is the information or perturbation at the i time point, ut is the mean, and δ is a constant.

When the random disturbance is equal to white noise εt, the AR(p) turns to a pure process:

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + \varepsilon_t + \delta \tag{6}$$

The p-order AR model is also expressed as

$$\varepsilon_t + \delta = X_t(1 - a_1 N_1 - a_2 N_2 - \cdots - a_P N_P) \tag{7}$$

The characteristic equation:

$$a(N) = (1 - a_1 N_1 - a_2 N_2 - \cdots - a_P N_P) \tag{8}$$

As can be seen from Equation 8, the autoregressive model requires a predetermined parameter p, indicating that the historical values are used to predict the current ones, so the time series data must be smooth.

MA model is used to deal with stochastic fluctuations in time series data by calculating the difference between the observed values and the moving average of the past observations. In the MA(q) model, "q" denotes the size of the moving average steps, which determines the number of past observations to be considered in the model.

In the AR model, when ut is not equal to white noise, it is usually seen as a moving average of order q:

$$u_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \tag{9}$$

where β is the moving regression coefficient and ε_t denotes the white noise at different time points.

In particular, if $X_t = u_t$, it means the present value and the historical value are irrelevant. The only thing that affects present value is the linear combination of historical white noise. MA model is as followed:

$$X_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \tag{10}$$

The combination of AR(p) with MA(q) is ARMA model.

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \tag{11}$$

When using ARMA model to forecast a set of trendless smooth stochastic time series by the historical moment data and random disturbances to predict future data.

The combination of AR model, MA model and difference operation is ARIMA model. The ARIMA (p, d, q) after the difference operation process is

$$S_t = a_1 S_{t-1} + a_2 S_{t-2} + \cdots + a_p S_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \tag{12}$$

3. Results and Discussion

3.1. Smoothing Treatment and White Noise Test

First of all, the original data is analyzed for smoothness, taking the taxi demand of the whole day on March 1 as an example. Observation of Figure 5 reveals that the time series data is not smooth, there are changes such as trend or seasonality. So it is necessary to perform the difference operation first. The result of the first-order differencing is shown in Figure 6, and there is no longer any obvious seasonal variation.

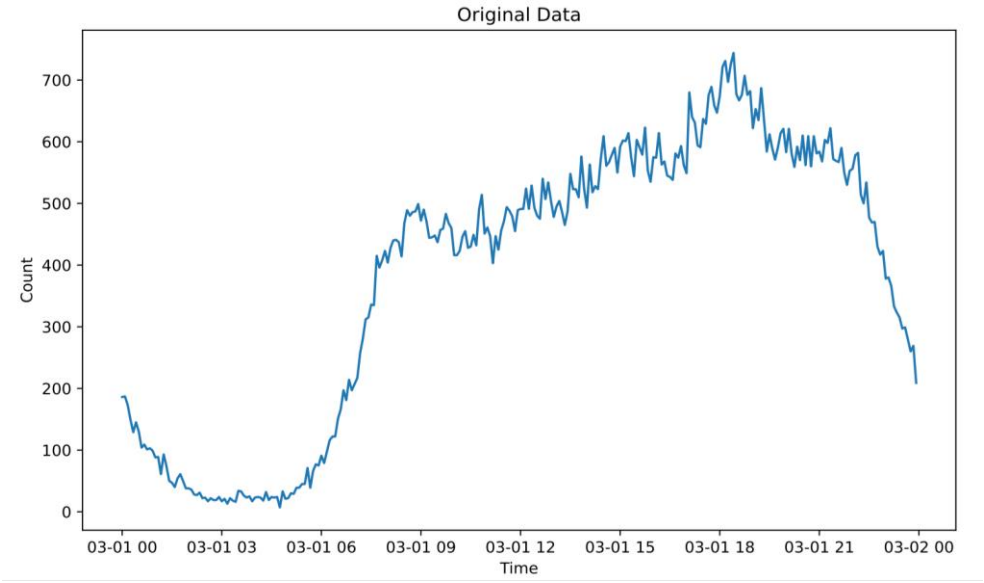


Fig. 5 Original data of daily demand in March 1st

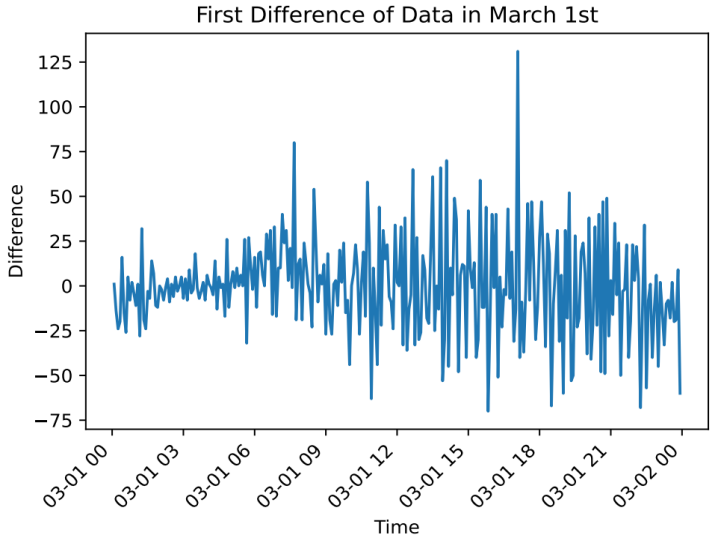


Fig. 6 First difference data of daily demand in March 1st

In order to verify whether the results of the first-order difference are consistent with a smooth series, this paper uses the Augmented Dickey-Fuller test (ADF test) to test the data smoothness. It is mainly used to determine whether the first-order difference of the data is smooth by testing whether there is a unit root in the time series data. If there is a unit root, it means that the time series data is not smooth, and vice versa. The result is as Table 3 showed.

Table 3. Results of ADF test

The ADF test results	Value
ADF Statistic	-3.126
P-value	0.025
Lags	8
Number of Observations	287
99% confidence interval	-3.454
95% confidence interval	-2.872
90% confidence interval	-2.572

From the ADF test results, we can see that the original time series data is less than the 95% confidence interval ADF test value after the first-order difference processing, so the data after the first-order difference is smooth data, so the parameter d = 1.

The white noise test is to check whether the data in the series are still correlated with each other. If all the information of the series has been extracted after the model calculation, the series will be transformed into a white noise series that can no longer be predicted, and no information available for further extraction will pass the white noise test. In this paper, the Ljung-Box test (LB test) is used to calculate the autocorrelation of the results in the time series by analyzing the information. The Q statistic of the LB test is expressed as:

$$Q_{(m)} = T(T+2)\sum_{t=1}^{m}\frac{\hat{\rho}_l^2}{T-l} \tag{13}$$

$$Q_{(k)} = n(n+2)\times\sum r^2\frac{j}{n-j} \tag{14}$$

$Q_{(k)}$ is the LB statistic to measure the autocorrelation of the time series with lag order k. n is the observation times of the time series. r(j) is the autocorrelation coefficient with lag order j, as Table 4 shows.

Table 4. Results of Ljung-Box statistic

0	lb_stat	lb_pvalue
1	25.721	3.94×10^{-7}
2	25.721	2.60×10^{-6}
3	25.794	1.05×10^{-5}
4	25.914	3.29×10^{-5}
5	27.606	4.35×10^{-5}
6	29.228	5.51×10^{-5}
7	29.521	1.16×10^{-4}
8	32.594	7.29×10^{-5}
9	32.628	1.55×10^{-4}
10	32.642	3.13×10^{-4}

lb_stat is the LB statistic, which is a statistic used to test autocorrelation, which is based on a series of lag-order autocorrelation coefficients. A larger LB statistic indicates more autocorrelation in the series. lb_pvalue is the p-value related to the LB statistic, which is used to measure the significance of the test result, which is the basis for determining whether the series is autocorrelated. A smaller p-value indicates that the series is significantly autocorrelated. According to the results in Table 4, the

larger LB statistic indicates a solid relationship of correlation between the lagged orders of the residual series, while the smaller p-value further supports this conclusion by indicating that the residual series of the posterior data have significant autocorrelation. This implies that the model fails to fully capture the full information and structure in the data, and there are unexplained dependencies that require further modeling.

3.2. Model Sizing

The initial approach for the ARIMA model sizing is to observe the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the time series. Parameter p and q range of these two functions can be roughly judged by the truncated and trailing characteristics. The trailing and truncated tails are used to analyze the autocorrelation of the residual series of the ARIMA model. The tailing refers to the autocorrelation coefficient of the residual series that remains high or decreases slowly at higher lag orders. The truncated tail means that the autocorrelation coefficient of the residual series becomes very small or close to zero after a certain lag order. According to the properties of truncated tails and trailing tails from Figure 7, the model takes p = 0,1,2, and q = 0,1,2

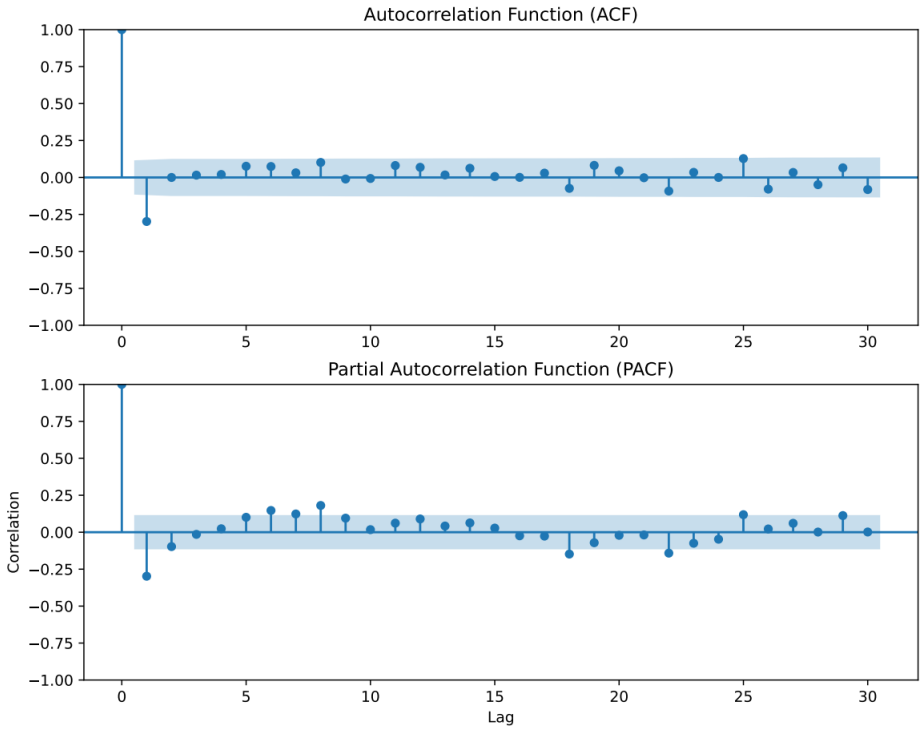


Fig. 7 Results of ACF and PACF

Different combinations of p and q are analyzed by Bayesian information criterion (BIC). n is the amount of data, k is the number of variables in the model, L is the maximum likelihood under this model, then we have:

$$BIC = -2\,ln(L) + ln(n)k \tag{15}$$

The BIC values are calculated as shown in Table 5, where p = 2 with q = 2 has the smallest BIC of 2689.616, so the ARIMA(2, 1, 2) model is finally used in this paper.

Table 5. Results of BIC values

	MA(0)	MA(1)	MA(2)	MA(3)
AR(0)	2734.676	2710.928	2716.321	2720.926
AR(1)	2713.313	2716.381	2698.541	2695.169
AR(2)	2716.153	2721.760	2689.616	2708.368
AR(3)	2721.726	2727.375	2693.529	2700.626

3.3. Prediction Results and Real Results

After determining the parameters of the ARIMA model as ARIMA(2,1,2), the training set from March 1st to 25th was used for training. Based on this model, the test set from 26th to 31st is predicted, and the results are shown in Figure 8.

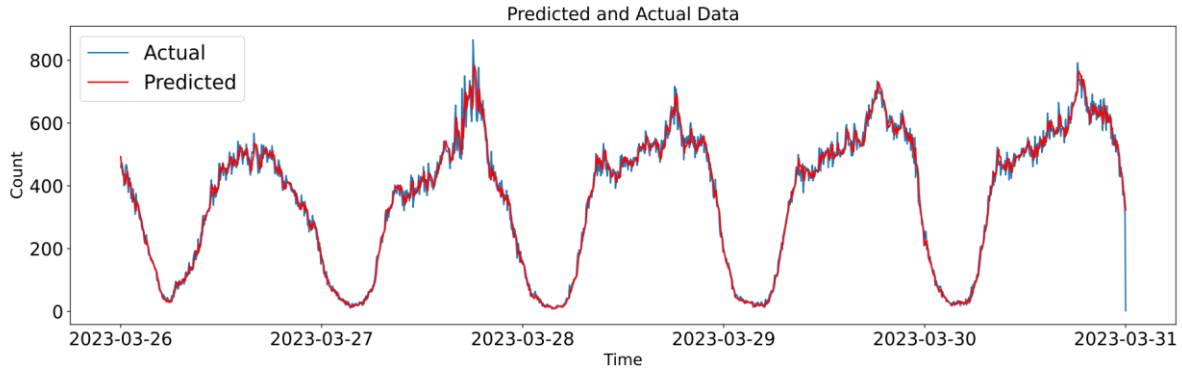


Fig. 8 Predicted data and true results of the test set

From Figure 8, we can see that the prediction results of the ARIMA (2,1,2) model constructed in this paper basically match the trend of the actual data change curve, and the fit is good. It indicates that after training, this algorithm has a high ability to simulate and predict the demand of yellow rental in New York City.

4. Conclusion

This paper summarizes preliminary information on taxi demand fluctuations, the distribution of ridership distances, and the distribution of origins and destinations by extracting information data on the date and time of yellow taxi boarding and alighting, coordinates of starting and ending locations, and trip distances in New York City in March 2023 for analysis.

The next focus is on analyzing the local taxi demand based on taxi boarding. The research methodology used the classical ARIMA time series analysis model. In order to determine the appropriate model parameters p, d, q in the ARIMA model according to the characteristics and trends of the series. The initial step is to convert the non-stationary series into a stationary series by difference operation. This step determined the value d. Then the characteristics of the stationary series are observed and calculated, p and q are determined by combining the requirements of AR model and MA model. Lastly by setting the value of p, d, q, the final suitable ARIMA model is determined. Then, the March taxi demand data were divided into training and testing sets, and after training, the final ARIMA model could perform better time correlation analysis and prediction of taxi demand numbers.

As a common and efficient time series forecasting model, ARIMA model can predict the distribution of taxi demand more accurately through training, which can provide more efficient services for operating companies, taxi drivers and passengers, and provide decision basis for government planning.

References

[1] Li J, He Y and Yang K L. Short-term Traffic Flow Prediction Based on Genetic Searching Algorithm, Journal of Pingdingshan College, 2023, 2(38): 47-52.

[2] Wu X P, Xu D S, Wu X Y and Jin J Q. A Predictive Visual Analytics Method for Taxi Routines. Journal of Computer-Aided Design & Computer Graphics, 2020, 32 (04): 520-530.

[3] Wang J, Gao Y, Zhang L, Ma L W and Feng J. Predicting Short-Term Urban Traffics Based on Causality Analysis Graph. Data Analysis and Knowledge Discovery, 2022, 6(11): 111-125.

[4] Du B, et al. Deep Irregular Convolutional Residual LSTM for Urban Traffic Passenger Flows Prediction. in IEEE Transactions on Intelligent Transportation Systems, 2020, 21: 972-985.

[5] Faghih S S, et al. Predicting short-term uber demand in New York City using spatiotemporal modeling. Journal of Computing in Civil Engineering, 2019, 33(3).

[6] Yao J B, Yuan X T, Yao B Z and Yu B. Optimization Model of Taxi Fleet Size Under Dynamic Demands. Transactions of Beijing Institute of Technology, 2012, 32(1).

[7] Riascos A P and Mateos J L. Networks and long-range mobility in cities: A study of more than one billion taxi trips in New York City. Scientific Reports, 2020, 10(1).

[8] Safikhani A, et al. Spatio-temporal modeling of Yellow Taxi demands in New York City using generalized star models. International Journal of Forecasting, 2020, 36(3): 1138–1148.

[9] Taxi & Limousine Commission. 2023. TLC Trip Record Data, Available at: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[10] Li X L. Research on Spatial and Temporal Characteristics Analysis and Demand Forecast Based on Taxi Data. Southwest Jiaotong University Master Degree Thesis, 2018.

[11] Poongodi M, et al. New York City taxi trip duration prediction using MLP and XGBoost. International Journal of System Assurance Engineering and Management, 2021, 13(S1): 16–27.

[12] Chang H. Research on Short-term Passenger Flow Forecast of Metro Based on LSTM Neural Network. the Master Degree of Electronic and Information Engineering of Xijing University, 2022.