ISSN 2616-7433 Vol. 4, Issue 6: 30-37, DOI: 10.25236/FSST.2022.040605

An Empirical Study on Stock Price Forecasting Based on ARIMA Model

Han Peng*, Zhou Yang

School of Management, University of Shanghai for Science and Technology, Shanghai, 200082, China
**Corresponding author*

Abstract: *The time series data in the financial market contains historical information, which can reveal the operation law of the system. This paper analyzes and predicts the stock price of China Merchants Bank based on the ARIMA model, selects 243 sample data of the daily stock closing price of China Merchants Bank from July 1, 2021 to July 1, 2022 as the research object, establishes the ARIMA model with Eviews, and uses Python for drawing. Finally, the closing price of the next 3 working days is predicted based on the model. The empirical results show that the ARIMA model has a good effect on the short-term prediction of stock prices, but in the long-term prediction, it can be considered to combine with other forecasting methods, and pay more attention to the stock market information and national macro policies, so as to improve the accuracy of the model prediction.*

Keywords: *time series; ARIMA model; Stock price forecasting*

1. Introduction

Time series is a series of data arranged in chronological order. These data are random, but there is a certain dependence between them. There are a lot of time series data in financial markets, such as changing stock prices, interest rates and so on. Different from cross-sectional data, time series data contains the operation law of the system. We can explore the law through research and predict the future trend, which is very necessary for financial workers.

Stock is the barometer of the economy, no matter for the country or for investors, stock price forecast is of great significance. ARIMA model is one of the most common statistical models used for time series prediction. In forecasting, it not only reduces the interference of random fluctuations, but also reduces the influence of time dependence on financial market indicators, and can accurately predict the short-term trend of stock prices.

In this paper, ARIMA model will be used to model the historical closing price data of China Merchants Bank stock, and the model will be used to predict the closing price of the next three days, so as to provide decision support for investors in stock investment.

2. Literature Review

The ARIMA model was first proposed by Box and Jinkins in the 1970s. In their book *Time Series Analysis Forecasting and Control*, the basic theoretical knowledge of ARIMA was briefly introduced. The future application fields are prospected[1]. Thomakos and Bhattacharya Fatai applied the ARIMA model to the economic field and predicted the inflation rate and industrial output rate of India, which achieved good forecasting results in the short term and laid a certain foundation for the future prediction in the financial field[2]. Jeffrey and Eric proved the effectiveness of ARIMA model in predicting financial indicators through empirical analysis of Shanghai A-share price index[3].

The study of ARIMA model in China is relatively late compared with foreign countries, but up to now, there has been a long-term development. Zhu Libin was the first to introduce ARIMA into our country's stock market prediction and discusses the practicality of the model, which has since become the theoretical basis of ARIMA application in our country[4]. Ai Xiaowei and Wang Youyuan used the ARIMA model to model the log return rate data of the Shenzhen Component Index in China, and tested the stability and rationality of the model, which proved that the model had a good effect in predicting short-term data[5]. Wu Yuxia and Wen Xin selected stock closing prices of "Huatai Securities" period 250 as the empirical analysis data of time series, and forecasted the law and trend of stock price changes in

ISSN 2616-7433 Vol. 4, Issue 6: 30-37, DOI: 10.25236/FSST.2022.040605

GEM market by establishing ARIMA model, and found that this model had good short-term dynamic and static forecasting effects[6]. By selecting A stock in the A-share stock market as the research object and using Python as the implementation tool, Liu Song and Zhang Shuai established the ARIMA model for testing and forecasting, and the maximum error between the predicted value of stock price and the real value in the short term was no more than 0.04[7]. This shows that the short-term prediction of stock prices by ARIMA model has a good effect and can provide help for stock market investors.

3. ARIMA Model Theory Introduction and Modeling Process

3.1. Theoretical Introduction of ARIMA Model

ARIMA model, full name of differential integrated moving average autoregressive model, is a famous time series forecasting method put forward by Box and Jenkins in the early 1970s, so it is also called Box-Jenkins model. It refers to the model established by transforming the non-stationary time series into stationary time series, and then regress the dependent variable on its lagged value, the present value of the random error term and the lagged value. ARIMA model includes moving average process (MA), autoregressive process (AR), autoregressive moving average process (ARMA) and ARIMA process according to whether the original series is stationary or not and the different parts contained in the regression.

ARIMA model includes three parameters: p, q and d. In the ARIMA(p, d, q) model, p is the number of autoregressive terms, q is the number of moving average terms, and d is the number of differences made to transform the original time series into stationary series.

The general ARIMA (p, q) model can be expressed as:

$$u_t = \phi_0 + \phi_1 u_{t-1} + \cdots + \phi_p u_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} (1)$$

Here, both p and q are non-negative integers, and $\{\varepsilon_t\}$ is a sequence of white noise. ARIMA model has the following two special cases: when q=0, ARIMA (p, 0) =AR (p) is called AR model; When p=0, ARIMA (0, q) =MA (q) is called MA model.

3.2 .Modeling Process of ARIMA Model

The modeling of ARIMA model consists of four steps: First, the original sequence is tested for stationarity, and the methods of testing stationarity include ADF test, autocorrelation diagram and partial autocorrelation diagram. If the sequence is not stationary, the difference operation is needed to make the sequence satisfy the condition of stationarity. Secondly, the ARIMA model was fitted and its parameters were estimated by the least square method. The best prediction model was selected by comparing the AIC, SC and HQ values of the model. Then, the residual sequence of the model is tested by white noise to judge the rationality of the model. If the white noise test is not passed, the ARIMA model needs to be fitted again. Finally, diagnostic analysis is carried out, and the established model is used to predict the future data, and the actual data is compared with it to test the accuracy. If the parameters are not accurately redetermined, a new model is established again.

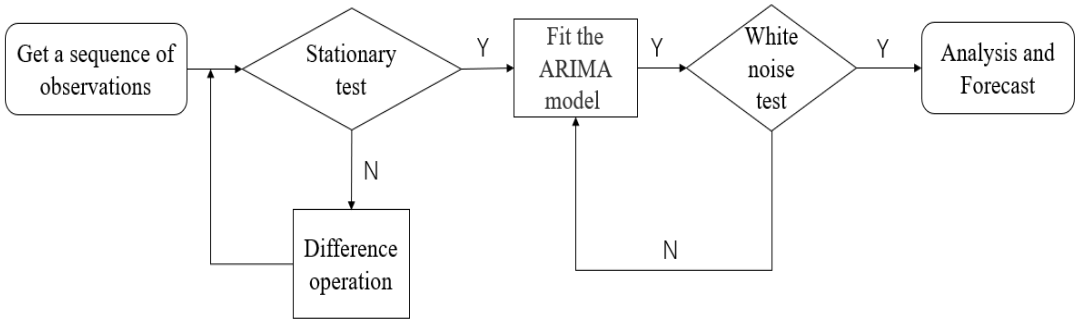

Figure 1: Modeling process of ARIMA model

ISSN 2616-7433 Vol. 4, Issue 6: 30-37, DOI: 10.25236/FSST.2022.040605

4. Empirical Study

4.1. Data Sources

The data in this paper comes from NetEase Finance database, and the stock closing price of China Merchants Bank from July 1, 2021 to July 1, 2022 is selected as the original data, with a total of 243 data samples. The sample size basically covers the closing price of all trading days of China Merchants Bank since July 2021.

4.2. Data Stabilization Processing and Unit Root Test

Through Python visualization, the original data of the stock price of China Merchants Bank is processed, and the time series image of the closing price is drawn. By observing the image, we find that the time series data of the closing price of China Merchants Bank presents the feature of non-stationary, but this is only a preliminary speculation, and it needs to be scientifically determined. Therefore, we conducted unit root test on the original data, and obtained that the value of T-statistic of ADF test was -1.777897, which was higher than the corresponding critical values of -3.457286, -2.873289 and -2.573106 respectively when the significance level was 1%, 5% and 10%. It can be concluded that the ADF test result of the original data falls within the interval of accepting the null hypothesis, that is, the closing price time series data of China Merchants Bank has a unit root and the data is non-stationary.

The non-stationarity of time series data can be solved by difference method. Therefore, we make first-order difference on the original data and conduct ADF test. At this time, the value of T-statistic obtained is -16.10265, which is smaller than the critical value of -3.996592 -3.428581 and -3.137711 when the significance level is 1%, 5% and 10%. In other words, after the first-order difference processing of the original data, the time series data is no longer non-stationary.

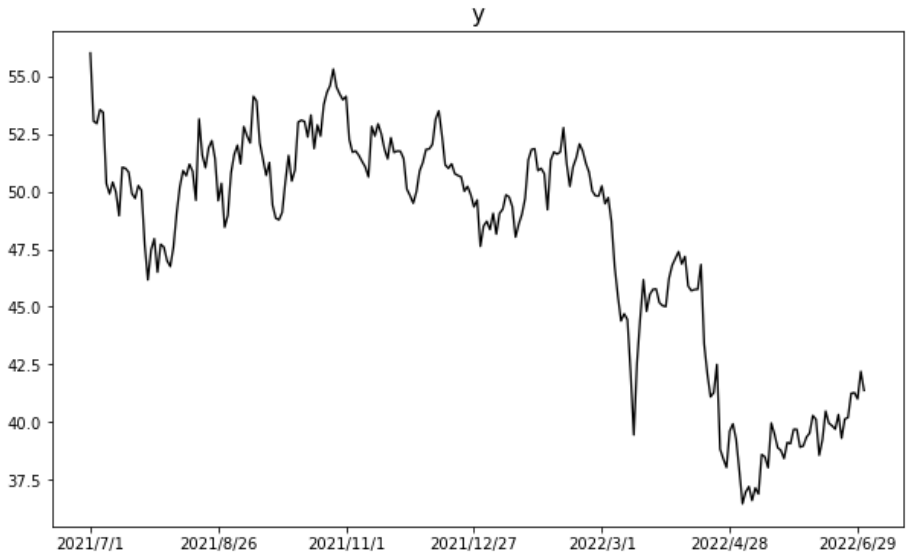

Figure 2: Sequence diagram of China Merchants Bank stock closing price

Null Hypothesis: Y has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.777897	0.3909
Test critical values:	1% level	-3.457286	
	5% level	-2.873289	
	10% level	-2.573106	

*MacKinnon (1996) one-sided p-values.

Figure 3: ADF test results of the original sequence

ISSN 2616-7433 Vol. 4, Issue 6: 30-37, DOI: 10.25236/FSST.2022.040605

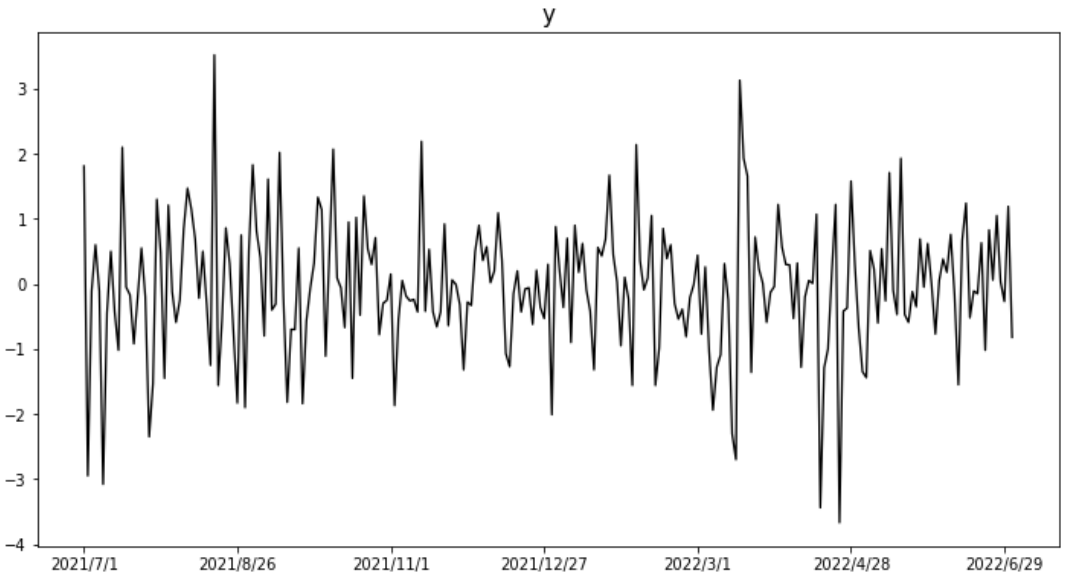

Figure 4: Sequence diagram after first difference

Null Hypothesis: DY has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-16.10265	0.0000
Test critical values:	1% level	-3.996592	
	5% level	-3.428581	
	10% level	-3.137711	

*MacKinnon (1996) one-sided p-values.

Figure 5: ADF test results after first difference

4.3. Draw Autocorrelation Diagram and Partial Autocorrelation Diagram

After the stationarity processing of the original sequence, we need to use ACF and PACF diagrams to identify the model form, and use the information criterion to determine the lag order p and q.

Through Python, the autocorrelation diagram ACF and partial autocorrelation diagram PACF after the first difference of the original sequence can be obtained. According to the observation method of ACF and PACF diagram (see Table 1), the p and q values of the tentative ARIMA model are 1,1, that is, the model is ARIMA (1,1,1). In order to ensure the optimum of the model, only naked eye observation is not enough to judge the optimal p and q values of the ARIMA model, and further judgment must be made through the information criterion, that is, to find the model that minimizes AIC, SC and HQ.

Table 1: ARIMA (P, D, Q) order determination

Model	ACF	PACF
AR（p）	Decay goes to zero	Censor to order p
MA（q）	Censor to order q	Decay goes to zero
ARMA（p,q）	Decay goes to 0 after order q	Decay goes to 0 after order q

Note: Censoring means that the value falls within the confidence interval, and 95% of the points need to meet this rule.

ISSN 2616-7433 Vol. 4, Issue 6: 30-37, DOI: 10.25236/FSST.2022.040605

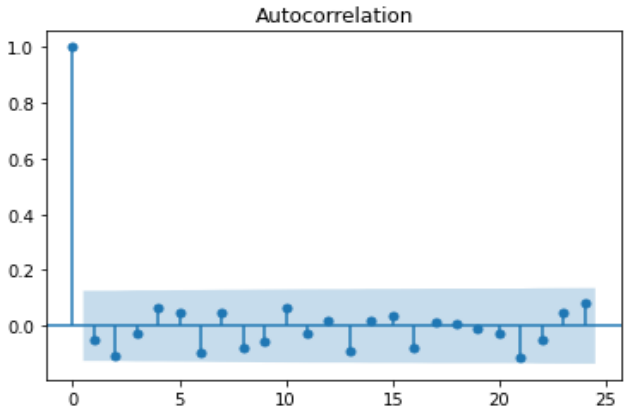

Figure 6: Autocorrelation plot after first difference

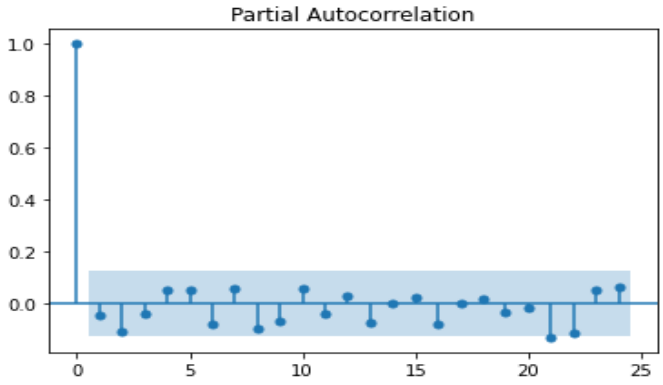

Figure 7: Partial autocorrelation plot after first difference

The p and q values determined by observing the Autocorrelograms and Partial Autocorrelograms after the first difference of the original sequence are only a rough estimate, and the exact values need to be compared with the nearby values. Four models, ARIMA(1,1,1), ARIMA(1,1,2), ARIMA(2,1,1) and ARIMA(2,1,2), were selected respectively. The AIC, SC and HQ values of the models were calculated by Eviews software and compared, and the optimal model was selected for prediction.

The significance of the coefficients and the values of AIC, SC and HQ in the four cases were observed and compared, and Table 2 was obtained.

Table 2: AIC, SC and HQ values of the four models

	Significance	AIC	SC	HQ
ARIMA(1,1,1)	MA(1) is not significant	2.9297	2.9872	2.9529
ARIMA(1,1,2)	AR(1), MA(1) and MA(2) are not significant	2.9420	3.0139	2.9710
ARIMA(2,1,1)	AR(1), AR(2) and MA(1) are not significant	2.9407	3.0126	2.9697
ARIMA(2,1,2)	signifinant	2.9204	3.0067	2.9552

It can be seen from Table 2 that the coefficients of ARIMA (2,1,2) model are very significant, and the values of AIC, SC and HQ are relatively minimal. Therefore, ARIMA (2,1,2) model is selected to model the daily closing price of China Merchants Bank.

4.4. Parameter Estimation

According to the above, ARIMA (2,1) is selected as the best prediction model. According to the results in the figure, the coefficients of AR(1), AR(2), MA(1) and MA(2) in the parameter estimates of model ARIMA(2,1) are statistically significant, while the p value corresponding to the constant term is

ISSN 2616-7433 Vol. 4, Issue 6: 30-37, DOI: 10.25236/FSST.2022.040605

0.43, which is not significant, so it is excluded. After elimination, the model ARIMA (2,1,2) is re-estimated and tested, and the results are shown in Figure 8. The results show that after removing the constant term, the coefficient of the model is significant, and the AIC value is smaller, which indicates that the modified model is more accurate and the fitting degree is better. Thus, the model expression can be written as follows:

$$y_t = 0.993138y_{t-1} - 0.936635y_{t-2} + \varepsilon_t - 1.072313\varepsilon_{t-1} + 0.979468\varepsilon_{t-2} (2)$$

Where ε_t is the sequence of residuals.

Dependent Variable: DY
Method: ARMA Maximum Likelihood (OPG - BHHH)
Date: 07/23/22 Time: 15:35
Sample: 0001 0243
Included observations: 243
Convergence achieved after 49 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	0.993138	0.045026	22.05699	0.0000
AR(2)	-0.936635	0.033071	-28.32188	0.0000
MA(1)	-1.072313	0.028106	-38.15298	0.0000
MA(2)	0.979468	0.026694	36.69232	0.0000
SIGMASQ	1.029934	0.080570	12.78303	0.0000

R-squared	0.046561	Mean dependent var	-0.052716
Adjusted R-squared	0.030537	S.D. dependent var	1.041487
S.E. of regression	1.025461	Akaike info criterion	2.915007
Sum squared resid	250.2740	Schwarz criterion	2.986881
Log likelihood	-349.1734	Hannan-Quinn criter.	2.943957
Durbin-Watson stat	1.970451		

Inverted AR Roots	.50-.83i	.50+.83i
Inverted MA Roots	.54+.83i	.54-.83i

Figure 8: Model building and parameter estimation

4.5. Residual Test

After parameter estimation, the adaptability of the fitted model is tested, that is, the residual sequence of the model is tested with white noise. If the residual sequence is not white noise, it means that important information has not been extracted, and the model should be reset. The specific determination method of residual test is to observe whether the autocorrelation coefficients are within the random interval. If so, the residual sequence is white noise.

As can be seen from Figure 9, p values are all greater than 0.05, indicating that the residual order of the model is listed as white noise, and the fitting degree of the model is good.

ISSN 2616-7433 Vol. 4, Issue 6: 30-37, DOI: 10.25236/FSST.2022.040605

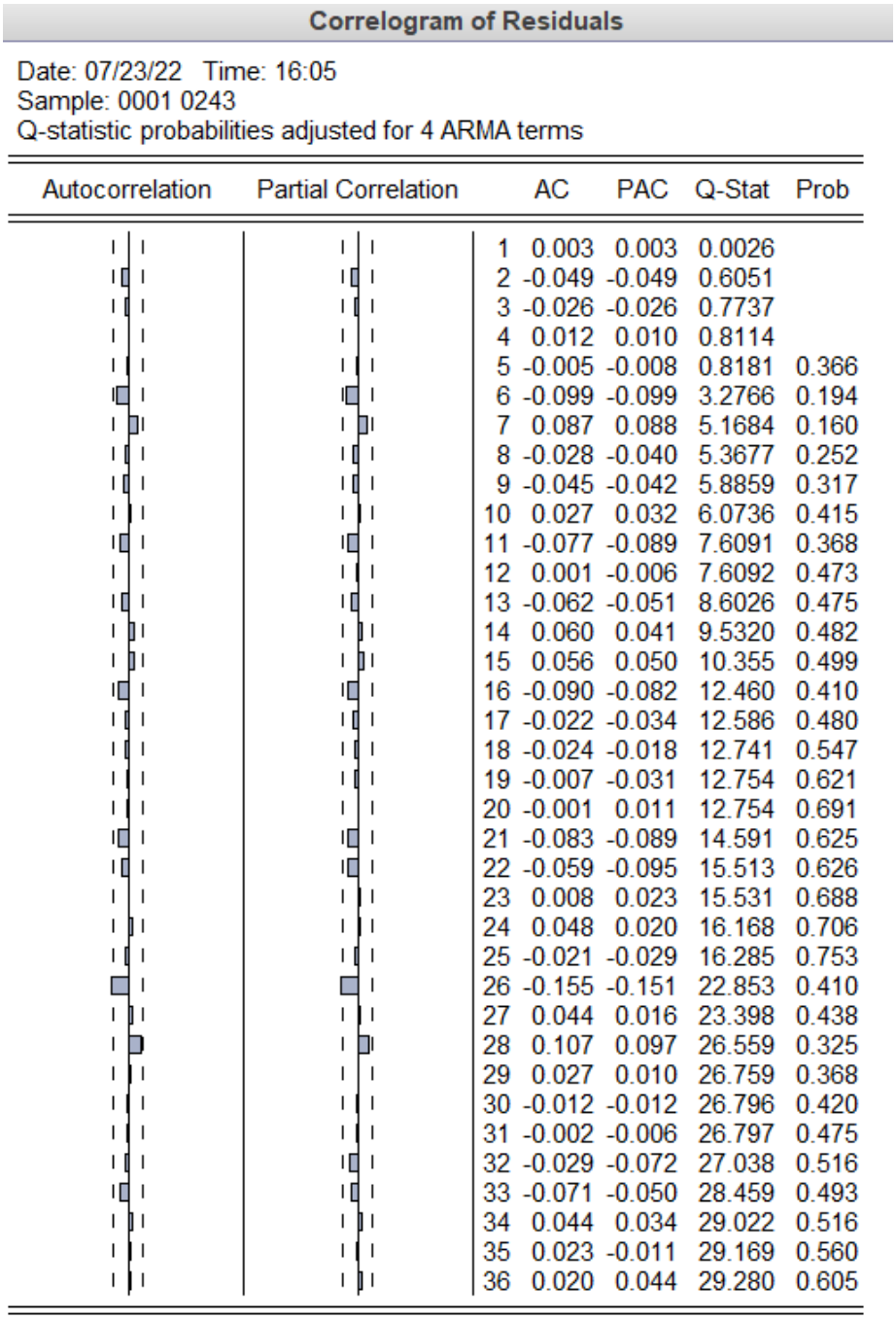

Figure 9: Residual examination

4.6. Analysis and Forecast

The established ARIMA model was used to predict the stock closing price of China Merchants Bank in the next three days, and the results were compared with the real data. The results are shown in Table 3.

Table 3: China Merchants Bank in the next three days of stock closing forecast results

	Predicted value	Actual value	Error
2022-07-04	41.33694	41.05	0.28694
2022-07-05	41.02176	40.98	0.04176
2022-07-06	39.95648	39.80	0.15648

According to Table 3, except for the large error between the predicted value and the actual value of the stock price on July 4, 2022, the predicted value of July 5, 2022 and July 6, 2022 are close to the real

ISSN 2616-7433 Vol. 4, Issue 6: 30-37, DOI: 10.25236/FSST.2022.040605

price with small error. The reason for the large error of the stock price forecast on July 4, 2022 may be that due to the influence of the weekend, the information on the stock market is relatively mixed and macro policy changes are relatively large at the beginning of the month. Therefore, when using ARIMA model for forecasting, it should be noted that ARIMA model is only suitable for short-term forecasting but not for long-term forecasting. At the same time, it is necessary to pay attention to the impact of macro policies in order to obtain more accurate data.

5. Conclusion

In this paper, 243 sets of closing price data of China Merchants Bank from July 1, 2021 to July 1, 2022 are modeled and predicted, and the closing price of the next three days is estimated, and the results are satisfactory. It shows that ARIMA model can solve the modeling problem of non-stationary time series well, and it can be used in the research and forecasting of financial time series problems. Investors can use the model, combined with tools such as Eviews and Python, to provide reasonable suggestions for their investment decisions.

Due to the limited time, this paper only conducts modeling analysis on the actual data of part of the closing price of China Merchants Bank. However, when the sample data changes greatly, the parameters of the model will change, and the prediction accuracy will also change, and the conclusion may lack universality. This requires investors to choose the historical data with relatively stable development and no large abnormal fluctuations caused by external factors such as unexpected events and policy introduction when using ARIMA model for stock price prediction.

To sum up, this paper establishes the ARIMA model, uses the historical closing price of China Merchants Bank as series data, and makes short-term forecasts of the price in the next three days, hoping to help investors reduce investment risks and discover investment opportunities. At the same time, we can combine other forecasting methods to pay more attention to stock market information and national macro policies, so as to improve the accuracy of model forecasting.

References

[1] G.E.P Box, G.M Jenkins. Time Series Analysis Forecasting and Control [M]. San Francisco: San Francisco, 1978.
[2] DIMITRIOS D. THOMAKOS, PRASAD S. BHATTACHARYA. Forecasting Inflation, Industrial Output and Exchange Rates: A Template Study for India[J]. Indian Economic Review, 2005,40(2).
[3] Jeffrey E Jarrett Ph.D., Eric Kyper Ph.D. ARIMA Modeling with Intervention to Forecast and Analyze Chinese Stock Prices [J]. International Journal of Engineering Business Management, 2011,3(3).
[4] Zhu Libin. Application of ARIMA model in stock market prediction [J]. Jiangsu Statistics, 1999(01):27-28.
[5] Ai Xiaowei, Wang Youyuan. Analysis of Shenzhen Component Index return rate based on ARIMA model [J]. Statistics and Decision, 2008(19):138-140.
[6] Wu Yuxia, Wen Xin. Based on ARIMA model to predict the short-term stock price [J]. Journal of statistics and decision, 2016(23):83-86.DOI:10.13546/j.cnki.tjyjc. 2016.23.051.
[7] Liu Song, Zhang Shuai. An empirical study on stock price prediction using ARIMA model [J]. Economic Research Guide, 2021(25):76-78.