

Prediction of the Fluctuation of the Shanghai Composite Index Based on the ARIMA Model

Pengyu Zhao

*School of Economics, Anhui University, Hefei, China
i42214013@stu.ahu.edu.cn*

Abstract. As one of the most influential emerging countries in the world, China's international influence and economic status are gradually strengthening globally. The Shanghai Composite Index is the core stock index of the Shanghai Stock Exchange, reflecting the overall performance of A-share and B-share stocks in the Shanghai market. This index serves as a key benchmark for China's stock market. In order to better analyze the stock market situation and explore the practicality and limitations of the ARIMA model, this paper selected the closing prices of the Shanghai Composite Index from January 2, 2020 to April 30, 2024 as the representative sample of the Chinese stock market, and based on the ARIMA model predicted the closing prices of the Shanghai Composite Index. The results show that due to the semi-strong efficiency of the Chinese market and the unpredictable short-term price trends caused by external events, the prediction accuracy of the ARIMA model will be affected by extreme fluctuations. This paper analyzes the effectiveness and limitations of the ARIMA model in prediction, providing a reference for investors on index prediction methods.

Keywords: Shanghai Composite Index, ARIMA, Forecast.

1. Introduction

After years of development, China's equity market has emerged as a globally significant financial hub. In the past decade, substantial advancements have been made in both primary and secondary markets, including the diversification of listed securities, trading volume growth, and overall market sophistication. After years of continuous development, the Chinese stock market has become one of the stock markets with significant global influence. In recent years, both the issuance market and the circulation market have witnessed remarkable progress in terms of the number and types of stock issuances, as well as the capacity and development level of the stock market. However, due to the significant impact of market risks on investors, the prediction of the Shanghai Composite Index has become particularly important.

He et al. predicted the opening price of stocks based on the AutoRegressive Integrated Moving Average Model (ARIMA Model), Long Short-Term Memory (LSTM) models and the combined models of ARIMA and ARIMA-LSTM [1]. They compared the accuracy of predictions of different time lengths and different models and concluded that the short-term prediction effect of the ARIMA model was better. The long-term prediction shows that the ARIMA-LSTM model and the combined

models of ARIMA and ARIMA-LSTM perform well. Liang et al. used the ARIMA model to make short-term predictions on the closing prices of the Shanghai Composite Index and the Treasury Bond Index, compared and analyzed the predicted values with the actual values, and concluded the impact of the COVID-19 pandemic on the returns of the stock and bond markets [2]. Guan used the ARIMA model, the Recurrent Neural Network (RNN) model and the ARIMA-RNN hybrid model to predict the China Securities Index 300 (CSI 300 Index), and concluded that the prediction accuracy of the ARIMA-RNN hybrid model was higher than that of the single recurrent neural network model [3]. Cao et al. predicted the stock price of Muyuan Food based on the ARIMA model and concluded that the ARIMA model can be used for stock price time series analysis, mainly providing a reference for short-term investment decisions, but the long-term prediction error is relatively large [4]. Li et al. conducted a trend analysis of the stock data of a well-known cosmetics company from 2010 to 2018 based on the ARIMA model, predicted the future development trend, and achieved good prediction results [5]. Liu et al. constructed the CNN-BiLSTM-AT deep neural network model based on the ARIAM-GARCH hybrid model to select the optimal investment decision for the investment portfolio [6]. This method has a relatively good result for small and medium-sized enterprise stocks. Su et al. predicted the CSI 300 Index based on the ARIMA-LSTM model, the multivariate LSTM model, the PCA-LSTM model and the double PCA-LSTM error correction model, and concluded that the double PCA-LSTM error correction model was more reliable and accurate than the other three models [7]. Huang's application of the ARIMA model to forecast China Merchants Bank's stock performance revealed its efficacy in identifying short-term price movement patterns within time series data. This modeling technique offers investors meaningful insights for making informed stock market investments [8]. Xiong et al. predicted the closing price of stocks based on the ARIMA-GARCH-M model and concluded that this model has a good effect and certain feasibility in the short-term prediction of stocks [9]. Ding conducted a predictive study on the CSI 300 Index based on the combined model of ARIMA and LSTM [10]. It is concluded that the combined model of ARIMA and LSTM has a better effect than their single models in predicting the indicators of MAE, RMSE and MAPE. Meanwhile, the combined model of ARIMA and LSTM also has certain practical value in the stock market.

The ARIMA Model is suitable for the prediction of time series. Based on the closing prices of the Shanghai Composite Index in China from January 2, 2020 to April 30, 2024, this paper predicts the closing prices of the Shanghai Composite Index in the coming days. Based on the prediction results of the model, evaluate the model and present its advantages and limitations.

2. Model construction

2.1. Mean equation

ARIMA is a statistical model for predicting time series by combining autoregressive, difference, and moving average methods. By means of the autocorrelation and difference of the data, the time series patterns hidden behind the data are extracted, and then these patterns are used to predict future data. Its expression is shown in Formula (1):

$$X_t = \sum_{i=1}^a \varphi_i X_{t-i} + \varepsilon_t + \sum_{j=1}^b \lambda_j \varepsilon_{t-j} + \varphi_0 \quad (1)$$

Where, φ_0 is the constant term, φ_i and λ_j are the coefficients of AR(a) and MA(b) respectively, a and b are the optimal order of AR and MA models respectively, ε_t represents the residual at time t,

that is, the uncertainty of the market, The expected value of the residuals is shown in Formulas (2) and (3):

$$E(\varepsilon_t) = 0 \quad (2)$$

$$E(\varepsilon_t \varepsilon_s) = 0 (t \neq s) \quad (3)$$

2.2. Data selection and sources

The daily closing price of Shanghai Stock Exchange (SSE) Composite index was selected as the target variable. Data from Investing.com and the sample frequency is daily data. The time Period is from January 2, 2020 to April 30, 2024. In the model building phase, the data set is divided chronologically into the training set (2020-2023, accounting for 80%) and the test set (2023-2024, accounting for 20%).

2.3. Data processing

In order to ensure the stationarity of the time series, Augmented Dickey-Fuller (ADF) test was used to verify the stationarity of the data. When the sequence is non-stationary, it is processed by log transformation and first - or second-order difference ($d=1$ or $d=2$) until it passes the stationarity test. Based on the training set data, Auto-ARIMA algorithm (from the forecast package) was used to automatically optimize ARIMA(p, d, q) parameter combinations, and Akaike information criterion (AIC) was used to compare the goodness of fit of models of different parameter combinations, and finally the optimal model with the lowest AIC value was selected for subsequent analysis. This standardized process ensures scientific modeling and reliable results.

The logarithmic difference is performed on the data to make it into a stable return series for the following modeling analysis.



Figure 1: Stock price trend of Shanghai Composite Index

Figure 1 shows the stock price trend of the Shanghai Composite Index from January 2, 2020 to April 30, 2024. It can be observed that the overall performance of the stock price of the Shanghai Composite Index during the sample period is an inverted U-shaped trend accompanied by the extension of the date and the overall rise of the stock price first and then the decline. The specific performance is that the closing price of the Shanghai Composite Index in 10 cases in 2020 is 3085.20. It rose to 3696.17 on February 19, 2021, and in 2021, although the Shanghai Composite Index fluctuated sharply in a small range, its closing price remained relatively stable, floating within 3400-3600. Since 2022, the price of the Shanghai Composite Index has shown a development trend of gradual reduction with the extension of time. This paper see an overall decline in index closing prices to the point where they reach the smaller value of the sample period in early 2024. During the

sample period, the stock price of Shanghai Composite Index fluctuated significantly in a short period of time, showing unstable characteristics.

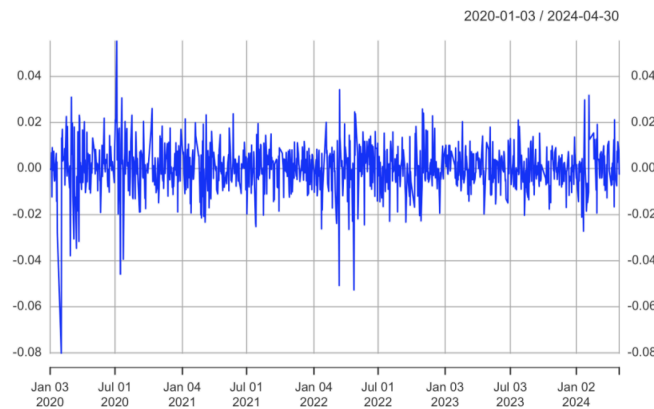


Figure 2: The trend of the logarithmic return of the Shanghai Composite Index

Figure 2 shows the trend of the logarithmic return series obtained after the logarithmic difference of the stock price of the Shanghai Stock Index. According to Figure 2, the logarithmic return series obtained after the logarithmic difference of the stock price of the Shanghai Stock Index fluctuated within a constant range of $(-0.02, 0.02)$ in most of the sample period, so the logarithmic return series of the Shanghai Stock Index has a certain degree of stability.

2.4. Descriptive statistics

According to Table 1, the mean value of the logarithmic return series of the Shanghai Stock Exchange Index is 0.000006, significantly greater than 0, indicating that the return situation is good during the sample period. The kurtosis coefficient of the return series is 5.641345, greater than 3, so the logarithmic return series of the Shanghai Stock Exchange index has a significant peak distribution feature. From the perspective of the skewness coefficient, the skewness coefficient of the return series is -0.671621, which is less than 0, indicating that the return series of the Shanghai Stock Exchange Index has a left-skewness distribution. Therefore, the yield series reflects the characteristics of "peak and thick tail" of financial time series data.

Table 1: Descriptive statistics

Statistic	Shanghai composite Index
Mean	0.000006
Median	0.0001
Maximum	0.0555
Minimum	-0.0804
Standard Deviation	0.0104
Skewness	-0.6716
Kurtosis	5.6413

2.5. Data check analysis

The stationarity test and normality test of the logarithmic return series of Shanghai Stock Exchange index must be carried out before using correlation for empirical research. The stationarity

examination of financial time series serves as a critical foundation for econometric modeling. When models are fitted using non-stationary financial data, the results become statistically invalid and may generate misleading regression relationships. In practice, unit root tests represent the standard methodology for assessing the stationarity properties of return sequences.

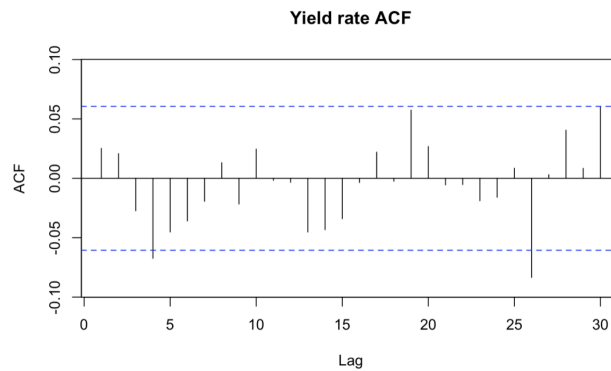


Figure 3: Autocorrelation function

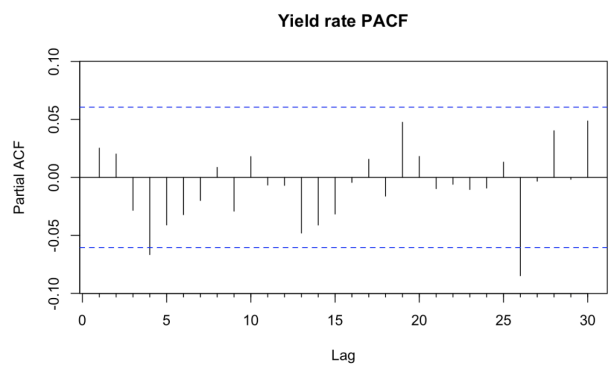


Figure 4: Partial autocorrelation function

In figure 3, most of the points fall in the blue interval, which shows that the autocorrelation is not significant. Then calculate and draw the partial autocorrelation function (PACF) graph. According to figure 3 and figure 4, judging the order of the AR model from the charts, it can be seen that it falls within the confidence interval after the 4th order. The fourth order tailing can be preliminarily judged, namely AR(4).

3. ARIMA model checking

According to auto ARIMA, by comparing the AIC and BIC criteria, the parameter was determined as ARIMA (3,1,1). The AIC value of the ARIMA(3,1,1) model is the smallest among the four models (see Table 2).

Table 2: The parameter estimation result table of the ARIMA (3,1,1) model

	AR(1)	AR(2)	AR(3)	MA(1)
	0.8593	0.0082	-0.0699	-0.8430
s.e.	0.1086	0.0407	0.0330	0.1055
$\sigma^2 = 1094$; log likelihood = -5146.69				
AIC=10303.39 AICc=10303.44 BIC=10328.15				

According to residual Figure 5, it is judged that this sequence is a white noise sequence and there is no unit root.

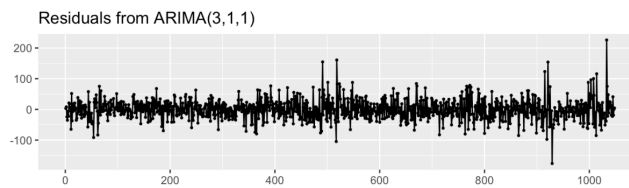


Figure 5: Residuals test from ARIMA(3,1,1)

As shown in Table 3, through the Ljung-Box test, P-value > 0.05, the difference is a sequence of white noise, it can be seen that it conforms to the normality.

Table 3: Test for heteroscedasticity

	X-squared	df	p-value
Jarque Bera Test	819.89	2	0.0000
Box-Ljung Test	0.0037985	1	0.9509

4. Model prediction

Next, this paper will use the ARIMA model to predict the closing price of the Shanghai Composite Index.

Figure 6 divides the data into a training set (2020.1-2023.6) and a test set (2023.6-2024.4), and makes predictions for the test set. Red represents the original data of the test set, and blue represents the predicted values. By comparing the two, it is found that the model error is relatively large. For the results in Figure 6, the prediction error is relatively large. The analysis may have the following reasons.

First, according to the efficient market hypothesis, China's current capital market is a semi-strong efficient market. It can be seen that market prices mainly reflect historical price information, but cannot fully reflect all public information. In such a market, due to the incomplete disclosure of information, there is uncertainty in the fluctuation of stock indices. The fluctuations of the Shanghai Composite Index today are not only influenced by the macroeconomy, but also by factors such as market sentiment and policy intervention. And the fluctuations that lead to market sentiment are even more unpredictable. A large proportion of Chinese stock investors are individual investors, and in most cases, individual investors lack understanding and judgment of the market and are also prone to being led by others. Therefore, the Shanghai Composite Index reflects more of an unpredictability.

Secondly, for the model, the ARIMA model is suitable for predicting small samples, data with current trends or seasonality. The closing price of the Shanghai Composite Index, as a daily non-seasonal data, is not accurately predicted by the ARIMA model.

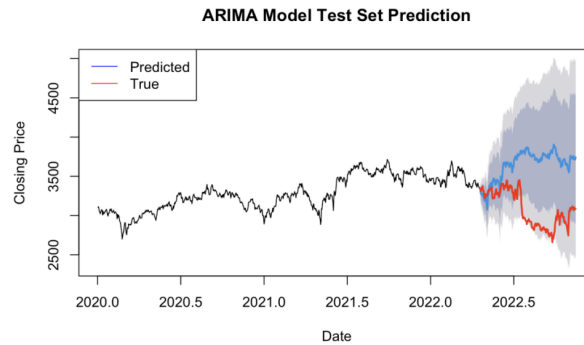


Figure 6: ARIMA model test set prediction

Due to the excessive time span and data volume of the test set causing inaccurate predictions, the test set is now redefined as the data of the last ten days of the data set. The prediction results are shown in Table 4. According to the error ratio, it can be seen that the prediction accuracy is relatively high.

Subsequently, predictions were made outside the dataset, as shown in Table 4. Table 4 presents the prediction results of the closing price of the Shanghai Composite Index for the next ten days outside the dataset.

Table 4: SSE index closing price forecast for the next 10 days

Date	Forecast	Lo95	Hi95	True value	Error ratio
2024-05-06	3084.200	3019.385	3149.016	3140.72	-0.017995874
2024-05-07	3083.326	2990.918	3175.735	3147.74	-0.020463571
2024-05-08	3082.469	2968.151	3196.786	3128.48	-0.014707142
2024-05-09	3081.794	2950.776	3212.812	3154.32	-0.022992594
2024-05-10	3081.269	2936.694	3225.844	3154.55	-0.023230255
2024-05-13	3080.872	2924.904	3236.840	3148.02	-0.021330233
2024-05-14	3080.574	2914.693	3246.454	3145.77	-0.020724974
2024-05-15	3080.351	2905.617	3255.084	3119.90	-0.012676368
2024-05-16	3080.185	2897.376	3262.993	3122.40	-0.013520049
2024-05-17	3080.061	2889.766	3270.355	3154.03	-0.023452218

As shown in Table 4, the deviation between the actual values and the predicted values is controlled within 5%, indicating that the model has a good fitting effect. This also verifies the reliability and accuracy of the ARIMA model in short-term predictions.

5. Conclusion

The Chinese market is not fully efficient, information is not completely transparent, and technical analysis is difficult to remain effective continuously. Policy and news are sensitive. The A-share market is greatly influenced by policy regulation (such as IPO pace, monetary policy, and regulatory policies), and these factors are difficult to be accurately predicted in advance. Market sentiment

fluctuates greatly, with a high proportion of retail investors, and is easily driven by emotions, leading to irrational fluctuations such as "herd effect". The trend of the Shanghai Composite Index is affected by multiple factors such as policy, economy, sentiment, international environment, and capital game, and these factors themselves have uncertainty and time-varying characteristics. The ARIMA model is more inclined to predict data with seasonal changes. For daily frequency data, the ARIMA model has some errors and cannot well reflect outliers. The accuracy of short-term prediction is relatively high.

According to the research in this article, it can be found that the ARIMA model is relatively accurate for short-term prediction of the closing price of the Shanghai Composite Index, but has a large error for long-term prediction. Further research can combine the GARCH model to analyze the volatility of the Shanghai Composite Index's closing price, analyze the fluctuation situation by combining major political or public events, or provide better assistance for prediction.

References

- [1] He, J. , Li, S. , He, Y. , et al. (2024). Stock analysis based on ARIMA and LSTM models. *Modern Information Technology*, 8(21), 41-45.
- [2] Liang, S. , & Wang, Y. (2024). Research on the impact of COVID-19 on China's stock and bond markets: An empirical analysis based on ARIMA model. *China Securities & Futures*, 2024(4), 59-66+89.
- [3] Guan, X. (2024). Stock price prediction based on ARIMA-RNN hybrid model. *Journal of Harbin University of Commerce (Natural Sciences Edition)*, 40(2), 250-256.
- [4] Cao, L. , & Zhang, R. (2024). Stock price prediction based on ARIMA model: A case study of Muyuan shares. *Fortune Times*, 2024(2), 54-56.
- [5] Li, L. , & Xin, H. (2024). Application of time series model ARIMA in data analysis. *Fujian Computer*, 40(4), 25-29.
- [6] Liu, Q. , Shi, S. , Lou, L. , et al. (2024). Stock price prediction and decision-making based on ARIMA-GARCH deep learning. *Journal of Changchun University of Science and Technology (Natural Science Edition)*, 47(1), 119-130.
- [7] Su, S. (2023). A comparative study of stock price prediction models based on LSTM and ARIMA, Chongqing University.
- [8] Huang, S. (2022). Stock price analysis and prediction based on ARIMA model: A case study of China Merchants Bank. *Management and Technology of SME*, 2022(11), 184-187.
- [9] Xiong, Z. , & Che, W. (2022). Application of ARIMA-GARCH-M model in short-term stock prediction. *Journal of Shaanxi University of Technology (Natural Science Edition)*, 38(4), 69-74.
- [10] Ding, M. (2022). Application of ARIMA and LSTM combined model in the stock market, Shanghai University of Finance and Economics.