Proceedings of the 2nd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/14/20230846

Stock Index Prediction Based on ARIMA Model

Junhao Hou[1,a,*]

[1]*School of Economics and Finance, Queen Mary University of London, London, UK*
a. te22175@qmul.ac.uk
*corresponding author

Abstract: Stock index price forecasting is a very important thing for financial markets. Stock indices select the most representative stocks in the stock market, which are the most favorable representatives of the industry, sector or market. Successful prediction of them can guide investors to a good payoff and allow researchers to understand the workings of the market economy. Time series models are widely used in stock price forecasting precisely because they have the advantage of being able to predict in complex environments such as large shocks. Therefore, this paper introduces the process of using time series model - ARIMA model to establish the prediction of index price fluctuation. And the five-year-long stock index data from May 2017 to April 2022 of the CSI 300 index is used as the research sample to successfully predict the volatility trend of the CSI 300 index in May 2022. The outcomes demonstrate that the ARIMA model has excellent capability for short-term forecasting. and is capable of forecasting complex situations that cannot be accomplished by linear models.

Keywords: ARIMA, stock index, prediction

1. Introduction

As a barometer of the market economy, the stock market not only indicates whether the socio-economic system is functioning well, but also affects the real income of stockholders [1]. If the stock market has been trending steadily upward, it can be concluded that the market economy is in a prosperous period, and stockholders can benefit from this prosperity of the market, and their shares are increasing in value. On the other hand, the market economy is likely in a period of recession or depression and stockholders face a greater risk of losing their investments if the stock market is in a startling or declining trend. Therefore, if investors can correctly forecast the trend of stock fluctuations, they can buy stocks at the bottom of the cycle and sell them at the top to gain benefits. The CSI 300 index is a "barometer" reflecting the complete functioning of the stock market and was chosen as the sample data for this study because it is the first index to describe the entire image of the Shanghai and Shenzhen stock markets since the formation of China's securities market [2].

The two most popular analytical techniques employed by academics to forecast stock values are time series models and linear regression models. Although the linear regression model is a widely used statistical tool for explaining one or more dependent variables using a collection of independent variables, it has substantial drawbacks [3]. According to Junhao Li, linear regression models can be used to anticipate when the stock market is steady, but they are unable to forecast the proper trend during times of turbulence and oscillation [4]. Complex time series models, which feature traits like volatility and non-stationarity compared to simpler linear regression models, can make up for the

DOI: 10.54254/2754-1169/14/20230846

drawback that linear regression models cannot anticipate in oscillating markets [5]. With the exception of the medical sector, other sectors of the stock index have experienced a significant amount of shock over the past few years as a result of the COVID-19. And in accordance with the characteristics of the time series model analysis, this paper uses a time series model for forecasting.

In order to make short-term forecasts of the performance of stocks in various nations and sectors, researchers have built ARIMA models in a variety of scenarios. In the academic community, there have been many successful cases thus far. Yingchao Zhang and Yingjun Sun selected the closing price of SSE index from February 1, 2016 to October 16, 2018 as the research data to construct an ARIMA model to analyze and forecast the SSE index, which successfully confirmed that time series analysis can predict the trend of stock price fluctuations for a brief amount of time [6]. Ariyo A A, Adewumi A O and Ayo C K correctly predicted the stock price fluctuations of Nokia and Zenith Bank using the ARIMA model [7]. Mondal P, Shit L and Goswami S constructing ARIMA models to successfully predict the short-term movements of fifty-six stocks in different sectors in India [8]. Almasarweh M and Alwadi S confirm the ARIMA model's ability to anticipate the trend of banking stocks over the short term using a large amount of data [9].

2. Methodology

In the early 1970s, statisticians Box and Jenkins proposed the ARIMA method, which has the ability to precisely predict non-stationary time series. Time series' dynamic, continuous properties are examined by ARIMA models, which also show how the past, present, and future of time series are related [10]. The difference autoregressive moving average model is another name of ARIMA model. There are three parts in the model, the moving average process (MA), autoregressive process (AR), autoregressive moving average process (ARMA). According to the different situations, like whether the initial sequence is stable, researchers can take different process to establish model. And ARIMA model transforms non-stationary time series into stationary time series and regresses only the lag value of dependent variable and the present value of random error factor.

2.1. Data

The data's time window selected for the first time in this paper is from May 2017 to April 2022. The CSI 300 index closed every day except holidays is taken as the sample data of this study, with a total of 1217 valid data, as well as the daily yield of the CSI 300 index during this period. All data in this paper are from GuoTaiAn database. During the research process, it was discovered during the research process that the closing price of the CSI 300 index under this time period was not suitable for building an ARIMA model, and the specific causes will be discussed later. Then, the daily return of the CSI 300 index was used to build an ARIMA model to predict the fluctuation trend of that index's daily return in May 2022, which was then converted into closing price. The yield and share price are defined in the formula (1) and (2), as follows.

$$Yield = (share\ price\ of\ current\ period - share\ price\ of\ previous\ period) / share\ price\ of\ previous\ period \quad (1)$$

$$Share\ price\ of\ current\ period = yield * share\ price\ of\ previous\ period + share\ price\ of\ previous\ period \quad (2)$$

2.2. Model

BOX and Jenkins used 3 simple and practical steps to build the ARIMA model, namely (1) Identification (2) Estimation (3) Diagnostic checking. Obviously, this is only a simple step and it is still necessary to decide how to build the model on a case-by-case basis. So, the ARIMA model is constructed in five steps as follows.

DOI: 10.54254/2754-1169/14/20230846

(1) The smoothness of the series was identified based on the line graph, autocorrelation function and partial autocorrelation function of the time series, and the variance was tested by ADF unit root.

(2) The non-stationary series should be rounded. The data must be differentiated if the data series is non-stationary and has an increasing or declining trend.

(3) Through the above steps, choose the appropriate model from AR, MA and ARMA models; the requirements of AR model are PACF p-order post-truncated and ACF trailing; MA model is PACF trailing and q-order post-truncated; ARMA model is PACF first p-bar irregular, followed by trailing, ACF first q-bar irregular, followed by trailing.

(4) After the model is selected, the time series model estimation is started, where the p-value corresponding to the selected model is less than 5, and it can be concluded the effect of the variable on the dependent variable is significant at the 5% level of significance.

(5) Get the results after forecasting and compare with the actual situation. The specific formula is as follows:

$$Y_t = \varphi_0 + \varphi_2 Y_{t-1} + \varphi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (3)$$

In the above equation, the actual value is expressed in terms of Y_t and the random error is represented by ε_t at t. φ_i and θ_j are the coefficients, p and q are integers commonly called autoregressive and moving averages, respectively.

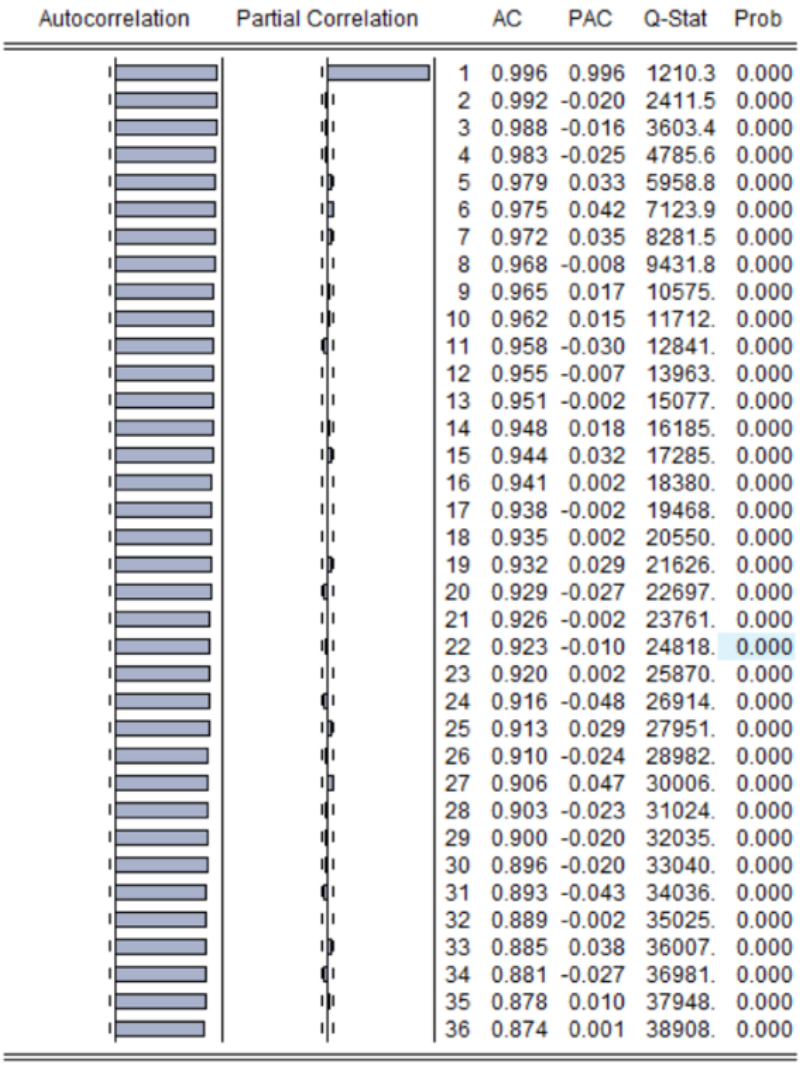
3. Results

3.1. Analysis on the Time Series Trend and Correlation of CSI 300 Index

As Figure 1 shown, it can be seen that the CSI 300 Index is showing a rising trend and then a downward trend from 2017 to 2018. In 2019, it shows a slow upward trend. 2020 shows an upward trend in the first half of the year, and the upward trend in the second half of the year is not as obvious as the first half of the year. 2021 shows a slow downward trend. Overall, from 2017 to 2022, the fluctuation of the CSI is amplitude and unstable.

Figure 1: The closing price of CSI 300 index from May 2017 to April 2022.
(Photo credit: Original)

DOI: 10.54254/2754-1169/14/20230846



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.996	0.996	1210.3	0.000
		2	0.992	-0.020	2411.5	0.000
		3	0.988	-0.016	3603.4	0.000
		4	0.983	-0.025	4785.6	0.000
		5	0.979	0.033	5958.8	0.000
		6	0.975	0.042	7123.9	0.000
		7	0.972	0.035	8281.5	0.000
		8	0.968	-0.008	9431.8	0.000
		9	0.965	0.017	10575.	0.000
		10	0.962	0.015	11712.	0.000
		11	0.958	-0.030	12841.	0.000
		12	0.955	-0.007	13963.	0.000
		13	0.951	-0.002	15077.	0.000
		14	0.948	0.018	16185.	0.000
		15	0.944	0.032	17285.	0.000
		16	0.941	0.002	18380.	0.000
		17	0.938	-0.002	19468.	0.000
		18	0.935	0.002	20550.	0.000
		19	0.932	0.029	21626.	0.000
		20	0.929	-0.027	22697.	0.000
		21	0.926	-0.002	23761.	0.000
		22	0.923	-0.010	24818.	0.000
		23	0.920	0.002	25870.	0.000
		24	0.916	-0.048	26914.	0.000
		25	0.913	0.029	27951.	0.000
		26	0.910	-0.024	28982.	0.000
		27	0.906	0.047	30006.	0.000
		28	0.903	-0.023	31024.	0.000
		29	0.900	-0.020	32035.	0.000
		30	0.896	-0.020	33040.	0.000
		31	0.893	-0.043	34036.	0.000
		32	0.889	-0.002	35025.	0.000
		33	0.885	0.038	36007.	0.000
		34	0.881	-0.027	36981.	0.000
		35	0.878	0.010	37948.	0.000
		36	0.874	0.001	38908.	0.000

Figure 2: Autocorrelation and skewness analysis of CSI 300 index.
(Photo credit: Original)

Figure 2 shows the results of the autocorrelation and bias correlation analysis of the CSI 300 index. It is evident that by the 36th period, the ACF of the CSI 300 Index is still not eliminated, and by the 2nd period PACF of the CSI 300 Index is eliminated. This situation does not fit any of the models stated in 2.2, indicating that the time series variable is unstable and needs to be differenced to eliminate the unstable phenomenon.

3.2. Differential

From Figure 3, it can be seen that the overall fluctuations of the CSI indices are leveling off after the first-order differential. It meets the requirements of constructing ARIMA model. In addition, Figure 5 demonstrates the result of autocorrelation and partial correlation analysis on the CSI after first-order differencing. it can be seen that the ACF of the first-order differential CSI is all eliminated and the trailing phenomenon of PACF is all absent, indicating that the time series variable is not suitable for ARIMA model.

DOI: 10.54254/2754-1169/14/20230846

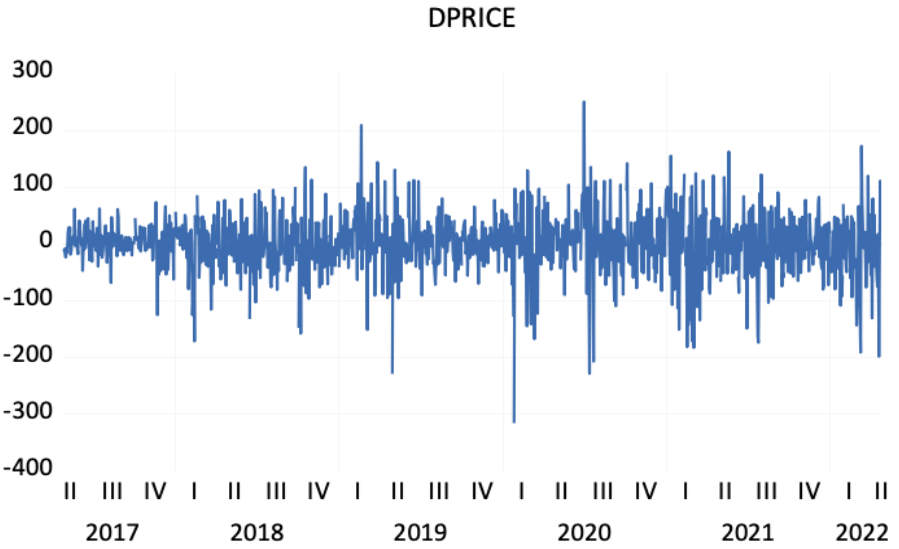
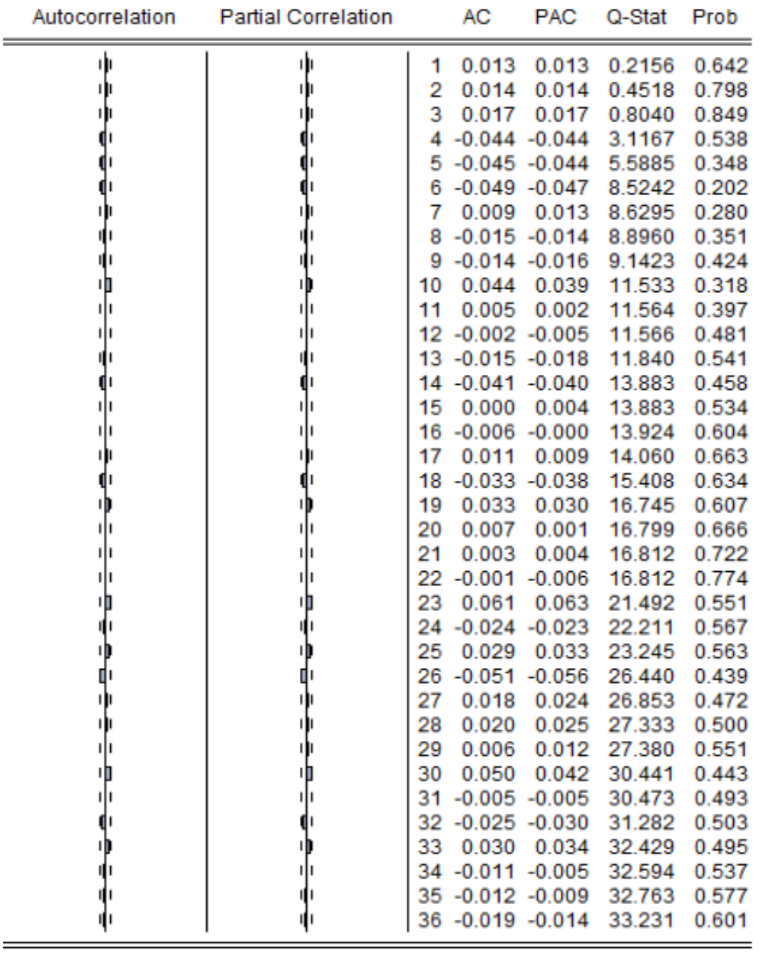


Figure 3: First-order differential stock price chart of CSI 300 index from May 2017 to April 2022. (Photo credit: Original)



	AC	PAC	Q-Stat	Prob
1	0.013	0.013	0.2156	0.642
2	0.014	0.014	0.4518	0.798
3	0.017	0.017	0.8040	0.849
4	-0.044	-0.044	3.1167	0.538
5	-0.045	-0.044	5.5885	0.348
6	-0.049	-0.047	8.5242	0.202
7	0.009	0.013	8.6295	0.280
8	-0.015	-0.014	8.8960	0.351
9	-0.014	-0.016	9.1423	0.424
10	0.044	0.039	11.533	0.318
11	0.005	0.002	11.564	0.397
12	-0.002	-0.005	11.566	0.481
13	-0.015	-0.018	11.840	0.541
14	-0.041	-0.040	13.883	0.458
15	0.000	0.004	13.883	0.534
16	-0.006	-0.000	13.924	0.604
17	0.011	0.009	14.060	0.663
18	-0.033	-0.038	15.408	0.634
19	0.033	0.030	16.745	0.607
20	0.007	0.001	16.799	0.666
21	0.003	0.004	16.812	0.722
22	-0.001	-0.006	16.812	0.774
23	0.061	0.063	21.492	0.551
24	-0.024	-0.023	22.211	0.567
25	0.029	0.033	23.245	0.563
26	-0.051	-0.056	26.440	0.439
27	0.018	0.024	26.853	0.472
28	0.020	0.025	27.333	0.500
29	0.006	0.012	27.380	0.551
30	0.050	0.042	30.441	0.443
31	-0.005	-0.005	30.473	0.493
32	-0.025	-0.030	31.282	0.503
33	0.030	0.034	32.429	0.495
34	-0.011	-0.005	32.594	0.537
35	-0.012	-0.009	32.763	0.577
36	-0.019	-0.014	33.231	0.601

Figure 4: Autocorrelation and partial correlation analysis of the CSI after first-order differencing. (Photo credit: Original)

DOI: 10.54254/2754-1169/14/20230846

3.3. Constructing an ARIMA Model Using Returns and Forecasting

Obviously, the direct use of daily closing prices of CSI 300 index to construct ARIMA model will have a large error because its autocorrelation and partial autocorrelation tests cannot satisfy any of the three models, AR, MA, and ARMA. Therefore, this paper uses the relationship between yield and daily closing price, take the daily yield of the same period as the research data, and construct an ARIMA model to predict the data of May 2022, and then transform the yield into the closing price to achieve the final goal of predicting the closing price fluctuation. The specific calculation process is as previously mentioned Formula (1) and (2).

3.3.1. Time Series Trend Analysis and Testing on the CSI Index Returns

Figure 5 demonstrates that the overall volatility tends to be stable following the first order difference in the daily yield of the CSI 300 Index, which is necessary for the construction of the ARIMA model.

Figure 5: Time series trend of CSI returns.
(Photo credit: Original)

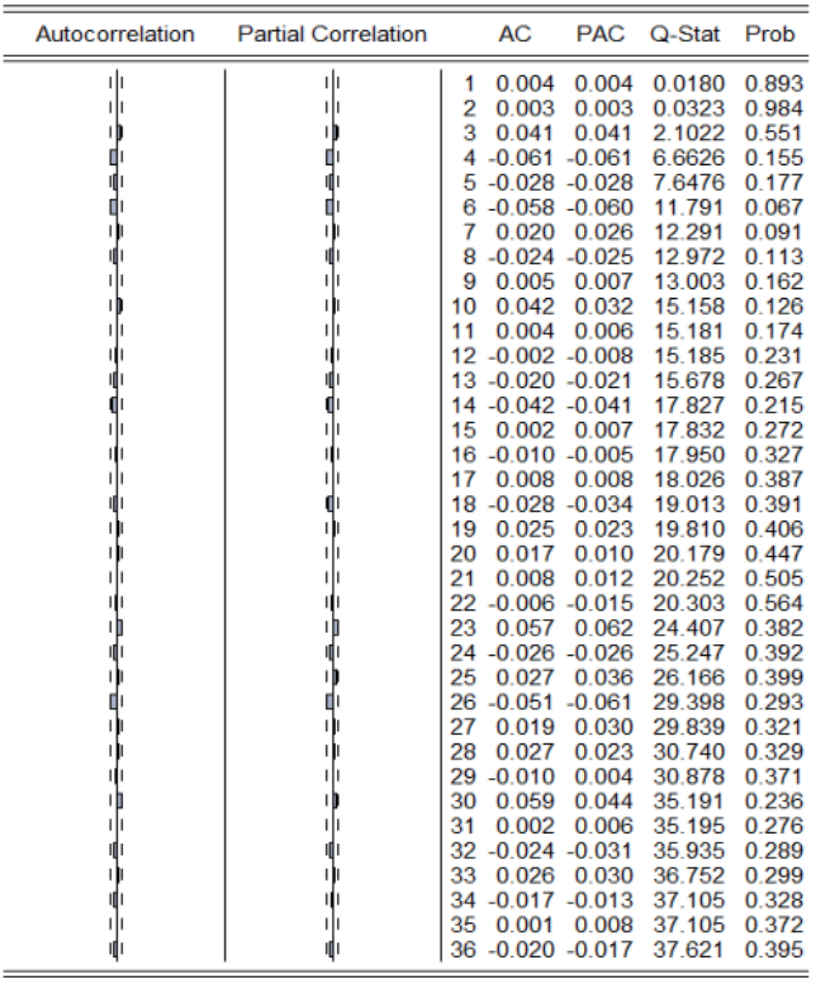
According to the Augmented Dickey-Fuller test statistic in Table 1, the t-value is -34.95884, and the associated p-value is below the significance threshold of 0.01, which means that the CSI index return rejects the original hypothesis "the variable is not smooth", indicating that the CSI index return fluctuates smoothly and the ARMA model can be constructed.

Table 1: Augmented Dickey-Fuller test statistic.

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-34.65803	0.0000
Test critical values	1% level	-3.435514	
	5% level	-2.863708	
	10% level	-2.567974	

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

DOI: 10.54254/2754-1169/14/20230846



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.004	0.004	0.0180	0.893
		2	0.003	0.003	0.0323	0.984
		3	0.041	0.041	2.1022	0.551
		4	-0.061	-0.061	6.6626	0.155
		5	-0.028	-0.028	7.6476	0.177
		6	-0.058	-0.060	11.791	0.067
		7	0.020	0.026	12.291	0.091
		8	-0.024	-0.025	12.972	0.113
		9	0.005	0.007	13.003	0.162
		10	0.042	0.032	15.158	0.126
		11	0.004	0.006	15.181	0.174
		12	-0.002	-0.008	15.185	0.231
		13	-0.020	-0.021	15.678	0.267
		14	-0.042	-0.041	17.827	0.215
		15	0.002	0.007	17.832	0.272
		16	-0.010	-0.005	17.950	0.327
		17	0.008	0.008	18.026	0.387
		18	-0.028	-0.034	19.013	0.391
		19	0.025	0.023	19.810	0.406
		20	0.017	0.010	20.179	0.447
		21	0.008	0.012	20.252	0.505
		22	-0.006	-0.015	20.303	0.564
		23	0.057	0.062	24.407	0.382
		24	-0.026	-0.026	25.247	0.392
		25	0.027	0.036	26.166	0.399
		26	-0.051	-0.061	29.398	0.293
		27	0.019	0.030	29.839	0.321
		28	0.027	0.023	30.740	0.329
		29	-0.010	0.004	30.878	0.371
		30	0.059	0.044	35.191	0.236
		31	0.002	0.006	35.195	0.276
		32	-0.024	-0.031	35.935	0.289
		33	0.026	0.030	36.752	0.299
		34	-0.017	-0.013	37.105	0.328
		35	0.001	0.008	37.105	0.372
		36	-0.020	-0.017	37.621	0.395

Figure 6: Autocorrelation and skew-correlation analysis of CSI index returns.
(Photo credit: Original)

Figure 6 shows the result of the autocorrelation and partial correlation analysis on the CSI index returns. It can be seen that the autocorrelation coefficient of the CSI index returns is at the 3rd order truncated tail and the partial autocorrelation coefficient is at the 3rd order truncated tail, which is in line with the requirements of constructing the ARMA model mentioned in the previous section, so it is initially decided to construct the model in ARMA (3,3).

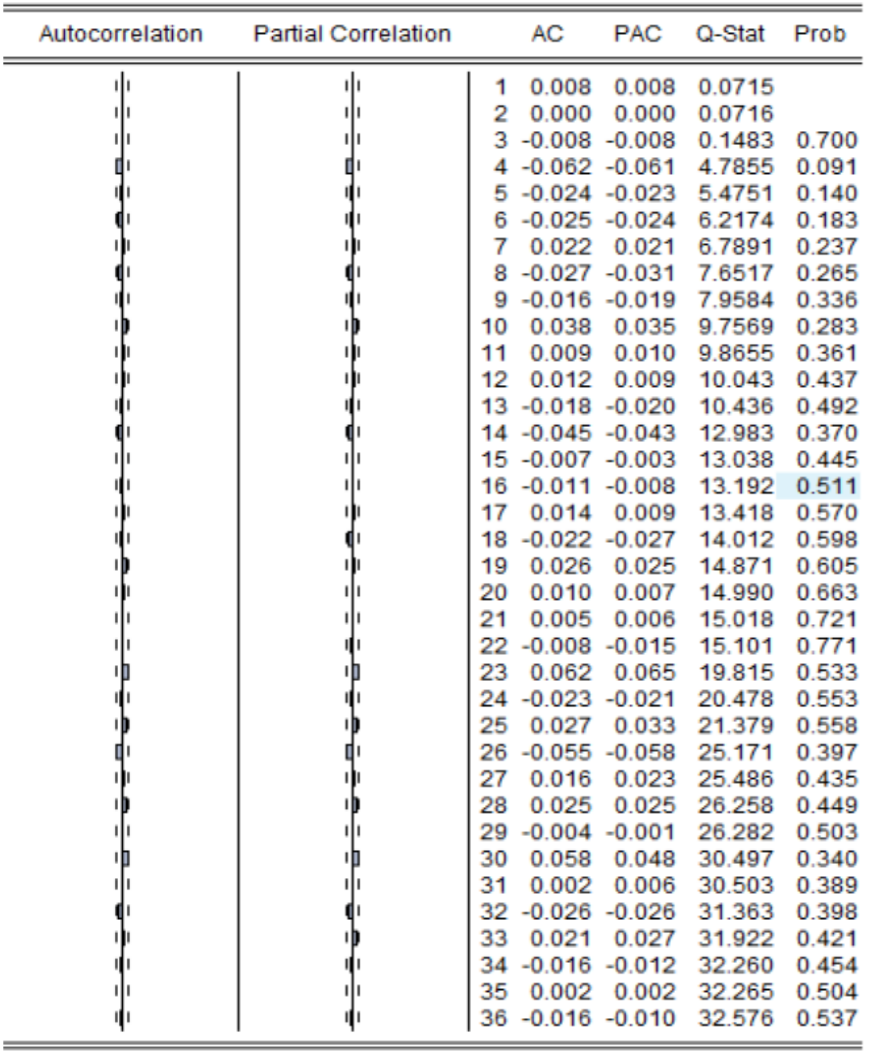
3.3.2. Constructing ARMA Model

Table 2 shows the empirical results based on ARMA model, and Figure 7 indicates the testing results of autocorrelation and partial autocorrelation of the residual series. It can be learned that the residual series of the CSI index return equation is basically a smooth series with zero mean, and the autocorrelation coefficients and partial autocorrelation coefficients of the residual series of the CSI index return equation are not serially correlated, therefore, it can be predicted with the help of ARMA(3,3) model.

DOI: 10.54254/2754-1169/14/20230846

Table 2: ARMA model.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(3)	-0.659917	0.198458	-3.325226	0.0009
MA(3)	0.713529	0.185972	3.836764	0.0001
SIGMASQ	0.000158	4.01E-06	39.37223	0.0000
R-squared	0.004783	Mean dependent var		0.000207
Adjusted R-squared	0.003144	S.D. dependent var		0.012606
S.E. of regression	0.012586	Akaike info criterion		-5.909942
Sum squared resid	0.192313	Schwarz criterion		-5.897360
Log likelihood	3599.200	Hannan-Quinn criter.		-5.905206
Durbin-Watson stat	1.981123			



	AC	PAC	Q-Stat	Prob
1	0.008	0.008	0.0715	
2	0.000	0.000	0.0716	
3	-0.008	-0.008	0.1483	0.700
4	-0.062	-0.061	4.7855	0.091
5	-0.024	-0.023	5.4751	0.140
6	-0.025	-0.024	6.2174	0.183
7	0.022	0.021	6.7891	0.237
8	-0.027	-0.031	7.6517	0.265
9	-0.016	-0.019	7.9584	0.336
10	0.038	0.035	9.7569	0.283
11	0.009	0.010	9.8655	0.361
12	0.012	0.009	10.043	0.437
13	-0.018	-0.020	10.436	0.492
14	-0.045	-0.043	12.983	0.370
15	-0.007	-0.003	13.038	0.445
16	-0.011	-0.008	13.192	0.511
17	0.014	0.009	13.418	0.570
18	-0.022	-0.027	14.012	0.598
19	0.026	0.025	14.871	0.605
20	0.010	0.007	14.990	0.663
21	0.005	0.006	15.018	0.721
22	-0.008	-0.015	15.101	0.771
23	0.062	0.065	19.815	0.533
24	-0.023	-0.021	20.478	0.553
25	0.027	0.033	21.379	0.558
26	-0.055	-0.058	25.171	0.397
27	0.016	0.023	25.486	0.435
28	0.025	0.025	26.258	0.449
29	-0.004	-0.001	26.282	0.503
30	0.058	0.048	30.497	0.340
31	0.002	0.006	30.503	0.389
32	-0.026	-0.026	31.363	0.398
33	0.021	0.027	31.922	0.421
34	-0.016	-0.012	32.260	0.454
35	0.002	0.002	32.265	0.504
36	-0.016	-0.010	32.576	0.537

Figure 7: Correlation plots of the residual series of the CSI index return equations.
(Photo credit: Original)

3.3.3. Forecasting Results

Figure 8 shows the comparison between the predicted results and the actual data. As it can be clearly see in the graph, the blue line indicates the forecast index price and the yellow line indicates the real

DOI: 10.54254/2754-1169/14/20230846

situation. Because of the long time span, the horizontal coordinate is one unit per 2 days for the coordinate scale. The blue and yellow lines in the entire table, the initial period from May 5, 2022 to May 11, 2022, the real situation by the holiday effect than the forecast situation price index is low, the specific reasons will be explained in detail in later chapters. But then the yellow curve quickly returned to 4000 points, always maintaining the 4000 points fluctuations, and the two lines are very close to each other. From the trend of the two lines in this graph, the situation predicted by ARIMA in this paper is very close to the real data, which is a strong evidence to prove that ARIMA has short-term forecasting ability.

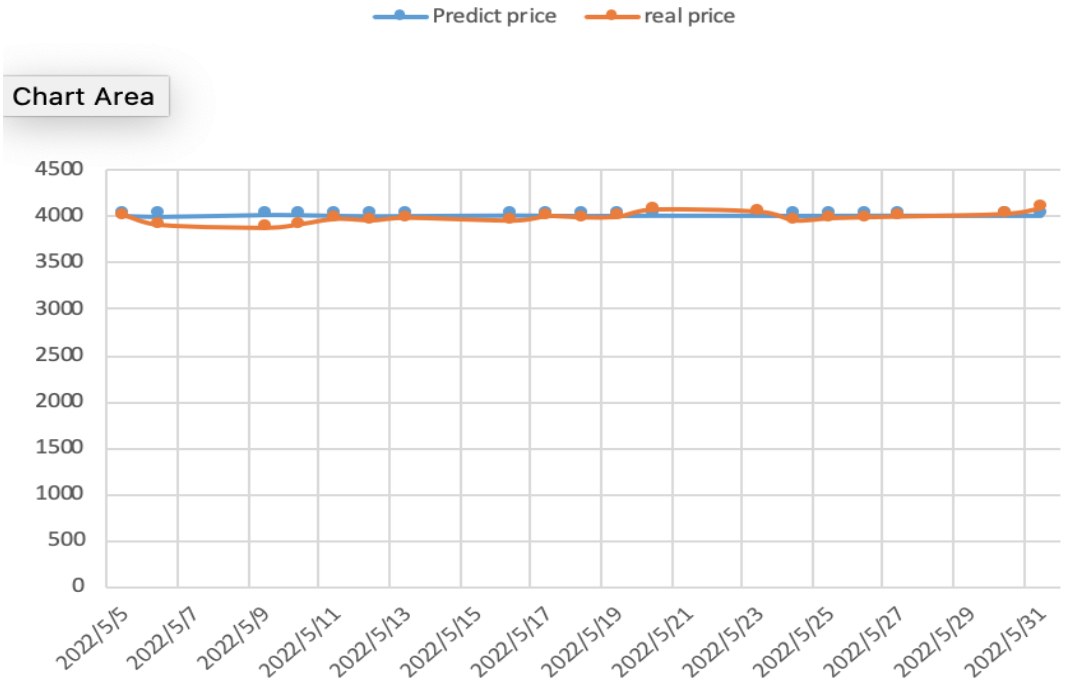


Figure 8: Comparison of forecast results and actual trend forecast.
(Photo credit: Original)

4. Discussion

The current sample data includes the extreme data of these years affected by the COVID-19, for example, the world was hit by the epidemic, in which the stock markets of several countries, including the United States, Canada, Brazil and South Korea, experienced multiple meltdowns. Within two weeks in March 2020, the U.S. stock market suffered four meltdowns within this short period of ten days, and the U.S. stock market collapsed in response. The Chinese stock market was also greatly affected, with all industry sectors except the medical sector beginning a prolonged downturn and severe fluctuations in stock market values. if the prediction can still meet the expected results under the influence of such external factors, it can prove that the time series model can make correct predictions than the linear regression model under the volatile and oscillating market; secondly, at the beginning of May 2022, the real CSI 300 index closing price appeared to be down, while the predicted index is more stable. This phenomenon can be explained by the holiday effect: either international or domestic holidays will have a negative impact on stock market returns [11]. This is a typical phenomenon in famous Chinese festivals, including the Spring Festival, Mid-Autumn Festival, National Day, and May Day. The stock prices around these holidays are volatile and cannot be predicted precisely. This is also an obvious drawback of this article, which cannot take into account too many complex situations, including unexpected public events and the inner thoughts of investors.

DOI: 10.54254/2754-1169/14/20230846

5. Conclusion

From the above graph, it is easy to conclude that the predicted and actual index prices obtained satisfactory results in terms of fluctuation trends and specific values, except for the holiday effect received at the beginning of the month. This paper puts forward the goal of establishing ARIMA model to predict the price of CSI 300 index. And the outcomes of the experiments using the ARIMA model show that the model has the ability to forecast short-term changes in stock values by using long-term time sample data. In this way, the construction of the ARIMA model can assist stock market investors in making wise investment choices. Also, research scholars can compare the data obtained from ARIMA simulations with the displayed data, and when large differences occur between the two, further research can be conducted to gain insight into what complex reasons are affecting the price of the stock. According to the results obtained, the ARIMA model can be more competitive than the linear regression model in terms of short-term forecasting. At the same time, this paper has some limitations. Forecasting fluctuations in stock price indices cannot fully take into account the effects of artificial factors such as holiday effects.

References

[1] *Xianzhong Tian, Anna Hu, Siyi Gu. A time series chain-based stock market forecasting method [J]. Journal of Zhejiang University of Technology,2021,49(5):503-510,563.*

[2] *Wang, Nong-Shi. Simulation study on volatility of CSI 300 index based on GARCH family model [J]. China Business Journal,2022(1):100-102.*

[3] *Li Y. Multiple linear regression model and stock sector index prediction [J]. National Circulation Economy,2017(10):62-63.*

[4] *Li, J. H.. An improved multiple linear regression based stock price forecasting model [J]. Science and Technology Economic Market,2019(8):61-62,64.*

[5] *Wang Xiaohong,Wang Mengyao,Hao Ting. Research on improved time-correlated serial stock price hybrid forecasting model [J]. Science and Technology for Development,2020,16(6):672-678.*

[6] *Zhang Yingchao,Sun Yingjun. An empirical study on the analysis and forecasting of SSE index based on ARIMA model [J]. Journal of Economic Research,2019(11):131-135.*

[7] *Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model [C]//2014 UKSim-AMSS 16th international conference on computer modelling and simulation. IEEE, 2014: 106-112.*

[8] *Mondal P, Shit L, Goswami S. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices[J]. International Journal of Computer Science, Engineering and Applications, 2014, 4(2): 13.*

[9] *Almasarweh M, Alwadi S. ARIMA model in predicting banking stock market data [J]. Modern Applied Science, 2018, 12(11): 309.*

[10] *Wang Yiting,Wei Jiangying. An empirical study of SSE 50 full return quotes based on ARIMA model [J]. Global Market,2018(16):55,57.*

[11] *Jiang Yitao,Yang Linyan. A study on the holiday effect of Chinese stock market [J]. Economic Economics,2009(6):135-138.*