International Journal of World Economic Research
ISSN: 3080-9886 (Print), ISSN: 3080-9894 (Online) | Vol. 1, No. 1, 2026

An empirical analysis of China's Shenzhen Composite Index based on ARIMA-ARCH model

Jinqing Zhang

Nanjing University of Science & Technology, Nanjing 210094, China

Abstract: Since the uncertainty of the stock market will affect the returns of investors, and the China Shenzhen Composite Index can basically reflect the fluctuations of the Chinese stock market, it has certain research significance to take the Shenzhen Composite index as the research object. In this paper, the stock composite price index of Shenzhen Composite Index from April 30, 1991 to May 31, 2023 (considering holidays) is selected as the sample, and the ARIMA (2,1,2) model is fitted after first-order difference for this unstable time series. Aiming at the ARCH effect in the residual series of the model, the GARCH (1,1) model is fitted, and the mean value forecast and volatility forecast of the price index of Shenzhen Composite Index from June to the end of October 2023 are carried out according to the fitted model. The final results show that the fitted model has strong short-term prediction ability, but weak long-term prediction ability. However, in general, the prediction results of ARIMA-GARCH model still have great reference value for Chinese investors.

Keywords: ARIMA model; ARCH effect; GARCH model; Time series; China Shenzhen Composite Index.

1. Introduction

The stock market is an important part of the financial market, which is crucial to the whole financial market. Therefore, by studying the price changes of the representative Shenzhen Composite Index, investors can have a deeper understanding of the trend and volatility risks of China's stock market, and provide the future trend prediction and volatility prediction of the stock index within a certain error range. On the one hand, it can help investors make better investment decisions and be vigilant and prevent the volatility risk of stocks. On the other hand, it can provide some necessary reference for China's stock market regulation and intervention.

Time series is a series of data arranged in time order, which is random but dependent on each other [1]. As a typical representative of the financial market, the price trend of the stock market is the time series, so it is desirable to use the time series to study the trend of the stock index. At present, ARIMA model has become an important tool for prediction due to its simplicity and applicability [2], and has been widely used in the study of stock price trend. However, the simple ARIMA model can only extract the level information of stock price trend, while the volatility risk information that investors are most concerned about cannot be fully extracted. For example, some stocks will rise or fall sharply in one session, and fluctuate smoothly in other sessions. Therefore, in the case of conditional heteroscedasticity in the residual sequence of the series, this paper uses GARCH model to extract volatility related information.

2. Basic theory and modeling steps of the model

2.1. ARMIMA model

Difference operation has a strong ability to extract deterministic information, and many non-stationary sequences will show the properties of stationary sequences after difference. In this case, we call this non-stationary sequence differential stationary sequence, and ARIMA model can be used to fit the differential stationary sequence.

ARIMA Model, called Autoregressive Integrated Moving Average Model (ARIMA), is a time series forecasting method proposed by Box and Jenkins in the early 1970s. Therefore, it is also called Box-Jenkins model.

ARIMA model comes into being on the basis of ARMA model. Both models include AR model and MA model. The difference between the two models is that ARMA model can only be used to analyze stationary time series, while ARIMA model can be used to analyze non-stationary series. In practical application, ARIMA model is more widely used because time series are mostly non-stationary series [3].

In the ARIMA(p,d,q) model, AR represents autoregression, and p is the order of autoregression; MA represents moving average,q is the number of moving average items, and d is the difference times (order) when the time series is stabilized. The model structure of ARIMA(p,d,q) is as follows:

$$\begin{cases} \Phi(B)\nabla^d x_t = \Theta(\mathrm{B})\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s) = 0, s \neq t \\ E(x_s\varepsilon_t) = 0, \forall s < t \end{cases}$$

Where, $\nabla^d = (1-B)^d$; $\Phi(\mathrm{B}) = 1 - \phi_1 B - \cdots - \phi_p B^p$, is the autoregressive coefficient polynomial of the stationary and reversible ARMA(p,q) model; $\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$, is the moving smoothing coefficient polynomial of the stationary and reversible ARMA(p,q) model [4].

In particular:

When d=0, the ARIMA(p,d,q) model is in fact the ARMA(p,q) model.

When p=0, the ARIMA(p,d,q) model can be abbreviated as the IMA(d,q) model.

When q=0, the ARIMA(p,d,q) model can be abbreviated as the ARI(p,d) model.

When d=1 and p=q=0, the ARIMA (0,1,0) model is the random walk model, or the drunkard model.

2.2. GARCH model

After Engle (1982) proposed the ARCH model, it aroused a strong response in the financial sector.

The essence of ARCH model is to use the Q-order moving average of the squared residual sequence to fit the value of the

current heteroscedasticity function:

$$h_t = \lambda_0 + \sum_{i=1}^{q} \lambda_i \varepsilon_{t-i}^2$$

Since the moving average model has Q-order censoring of the autocorrelation coefficient, the ARCH model is actually only applicable to the short-term autocorrelation process of heteroscedastic functions. However, in practice, the heteroscedasticity function of some residual series has long-term autocorrelation.

To fix this problem, Engle's student Tim Bollerslev proposed the Generalized Autoregressive Conditional Heteroskedastic (GARCH) model in 1985. The structure of GARCH(p,q) model is as follows:

$$\begin{cases} x_t = f(t, x_{t-1}, x_{t-2}, \cdots) + \varepsilon_t \\ \varepsilon_t = \sqrt{h_t e_t} \\ h_t = \omega + \sum_{i=1}^{p} \eta_i h_{t-i} + \sum_{j=1}^{q} \lambda_j \varepsilon_{t-j}^2 \end{cases}$$

GARCH model is actually based on ARCH model, adding the p-order autocorrelation of heteroscedasticity function. It can also effectively fit heteroskedastic functions with long-term memory.

2.3. Step of modeling

(1) Time series data preprocessing: Firstly, whether the series is stationary or not is determined by unit root test. If there is unit root, it indicates that the series is non-stationary and needs to be stationary. There are two ways to make non-stationary data stationary, one is to make it appear stationary in trend by taking the logarithm of the series, and then perform binary or multivariate time regression analysis on this time series. The second is through difference. If the first-order difference cannot reach the stationary sequence, the second-order difference is needed until all the data pass the stationarity test [5]. Moreover, for the stationary sequence or the stationary sequence, the Ljung-Box method should be used to determine whether the sequence is a white noise sequence. Only when the final sequence is a stationary non-white noise sequence can the model fitting be carried out in the next step.

(2) Model order determination and parameter estimation: the appropriate model is selected for fitting through the tailing or censoring of the autocorrelation coefficient map (ACF) and partial autocorrelation coefficient map (PACF), and then the optimal model order is determined according to the censoring order of ACF and PACF and the difference times of sequence stationarity, combined with AIC, AICc and BIC and other information criteria. Finally, the parameters of the model can be estimated by the maximum likelihood estimation method or the least squares estimation method.

(3) Significance test of parameters and models: whether it is significant and non-zero is judged by the P value of parameters and models.

(4) Residual sequence analysis: Ljung-Box test is performed on the residual sequence after model fitting to judge whether the residual sequence is white noise sequence, so as to judge whether the model has fully extracted the level information of the sequence.

(5) ARCH effect test: that is, to judge whether the residual series has the conditional heteroscedasticity property. The two commonly used ARCH test statistics are the Portmanteau Q test statistic and the LM (Lagrange multiplier method) test statistic.

(6) Fitting ARCH or GARCH model: For the series with ARCH effect, it is necessary to fit ARCH model to fully extract the fluctuation information of the series. When there is high-order ARCH effect, GARCH model can be used to fit the series instead.

(7) Use the fitted model to predict.

3. An Empirical study

3.1. Selection of data

The data of this study come from the flush iFinD data terminal, and the stock composite price index (386 data) of Shenzhen Composite Index from April 30, 1991 to May 31, 2023 (considering holidays) is selected. Finally, the stock composite index from June to October 2023 is predicted by the fitted model. The prediction accuracy of the model is evaluated by comparing it with the actual index.

3.2. Stationarity test and stationarity treatment

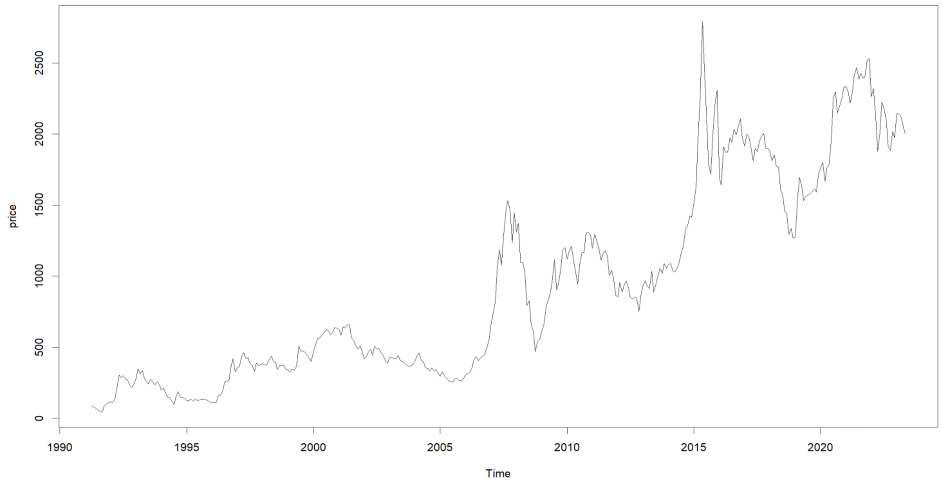

Fig. 1 Time series of monthly price index of Shenzhen Composite Index from April 30, 1991-May 31, 2023-May 31

Figure 1 shows the time series chart of the price at the end of the month of the Shenzhen Composite Index. It can be seen from the figure that the Shenzhen Composite price Index was in a relatively stable state before 2006, but from the perspective of the whole period, it was in an upward trend, and there were several relatively high peaks, especially in

2008 and 2015. This fully shows that the Shenzhen Composite index is not a stationary time series in this interval. A more objective judgment method is to use ADF to perform unit root test, and the specific results are shown in Table 1. It can be seen that the ADF test statistic is t= − 0.1251, and its absolute value is significantly less than the absolute value of the critical value at the significance levels of 1%, 5% and 10%, while the corresponding probability value p=0.90053>α=0.05. This shows that the null hypothesis cannot be rejected (there is a unit root, that is, the series is not stationary), which further proves that the series of the SZSE Composite index with monthly intervals from 1991 to 2023 is a non-stationary time series.

For unstable time series, the mainstream processing method is differential stationarity. The time sequence diagram after first-order difference of the original series is shown in 2. It can be seen from the figure that the series after first-order difference fluctuates around the zero mean value and is in a relatively stable state. The ADF unit root test is re-conducted on the sequence after first-order difference, and the specific results are shown in Table 2. It can be seen that the ADF test statistic is t=-14.0303, and its absolute value is significantly greater than the absolute value of the critical value at the significance levels of 1%, 5% and 10%, while the corresponding probability value p=2×10^(-16) < α=0.05. The illustration requires rejection of the null hypothesis that the series is stationary.

Table 1. ADF test results of time series of Shenzhen Composite Index

		t Statistic	P value
Augmented Dickey-Fuller test statistic		-0.1251	0.90053
Critical values for test statistics:	1% level	-2.58	
	5% level	-1.95	
	10% level	-1.62	

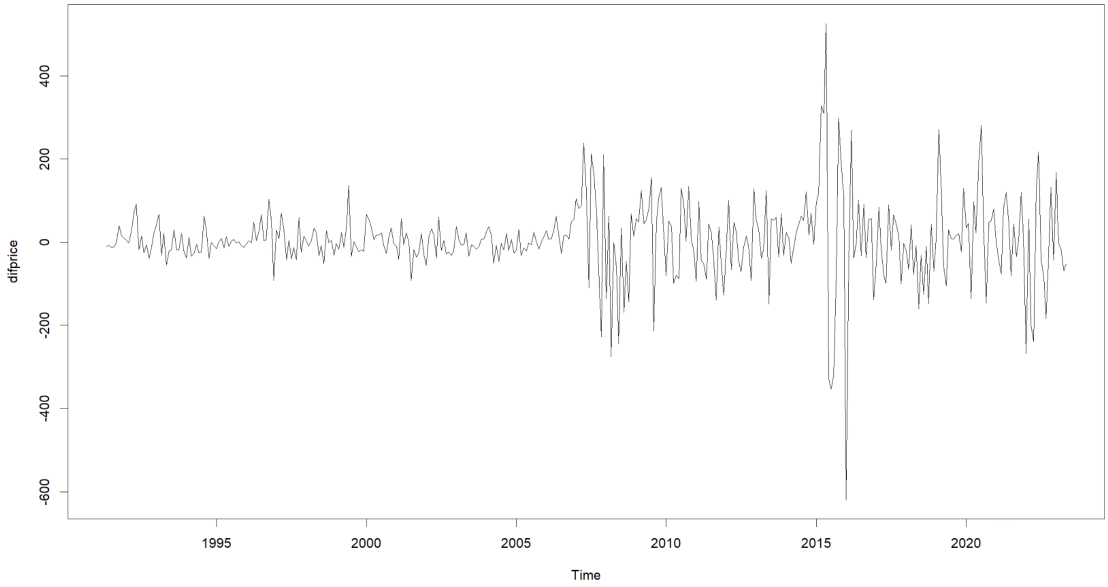

Fig. 2 Time series diagram of Shenzhen Composite Index after first-order difference

Table 2. ADF test results of the series after first difference

		t Statistic	P value
Augmented Dickey-Fuller test statistic		-14.0303	2×10^{-16}
Critical values for test statistics:	1% level	-2.58	
	5% level	-1.95	
	10% level	-1.62	

3.3. White noise test

The Ljung-Box method is used to perform the white noise test on the series after difference, and the specific test results are shown in Table 3. Based on the results of stationarity test, the sequence after first difference can be considered as stationary non-white noise sequence.

Table 3. Ljung-Box test results of the series after first difference

df	p_value
6	0.0001154
12	2.677×10^{-5}

3.4. Model order determination and parameter estimation

For the sequence after first-order difference, the corresponding autocorrelogram ACF and partial autocorrelogram PACF are drawn respectively, as shown in Figure 3 and Figure 4. It can be seen from Figure 3 that the ACF is in a trailing state, and the PACF in Figure 4 is also in a trailing state, so the ARIMA(p,d,q) model can be selected to fit the sequence. Since the original sequence is stationary by first-order difference,d =1 in ARIMA(p,d,q) model. In order to determine the order of p and q in ARIMA(p,d,q) model, The judgment can be made by observing the ACF and PACF graphs, integrating Akaike information criterion (AIC), Akaike Information criterion revised version (AICc) and Bayesian information criterion (BIC), and considering the significance of parameters. To fit the optimal model, the indicators of the ARIMA(p,d,q) model with different p and q values are shown in Table 4.

According to Table 4, all parameters of ARIMA (0,1,1), ARIMA (2,1,1) and ARIMA (2,1,2) are significant, but ARIMA (2,1,2) is optimal based on AIC, AICc and BIC, so ARIMA (2,1,2) is finally selected as the fitting model of this series.

The maximum likelihood estimation method is used to estimate the parameters of the model, and the specific results are shown in Table 5. It can be seen from Table 5 that the P values of AR (1), AR (2), MA (1) and MA (2) are all close to 0, indicating that these parameters are significant. However, the P value of the constant term is greater than 0.05, indicating that the constant term is not significant, so it can be removed

from the model ARIMA (2,1,2), and then the model parameters are re-estimated and tested. The pairs of AIC, AICc and BIC of models with constant terms are shown in Table 7, and it can be seen that the AIC, AICc and BIC after the constant term is removed are smaller, indicating that the model without constant term is more accurate and has better fitting.

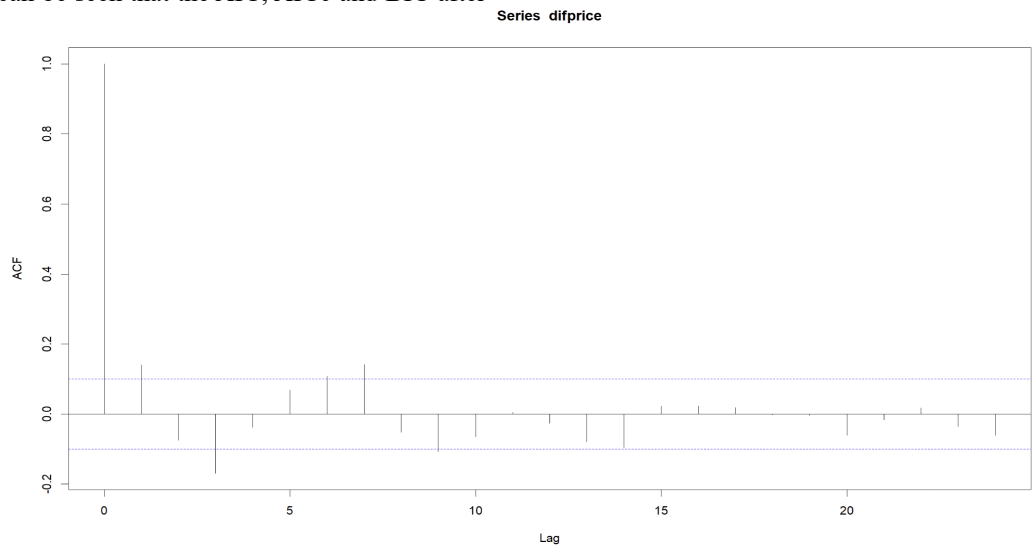

Fig 3 Autocorrelogram of the series after first difference

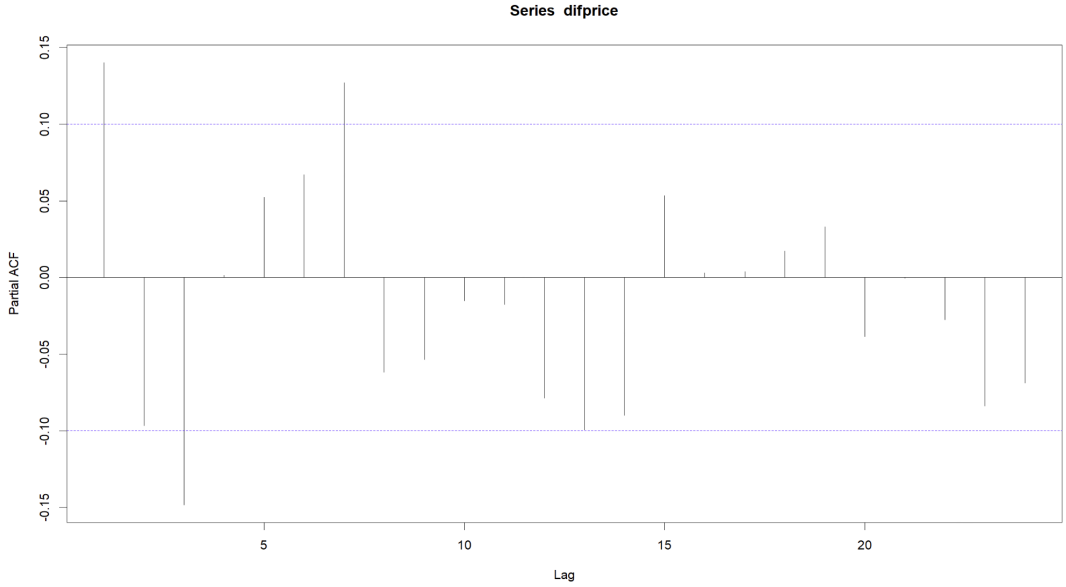

Fig 4 Partial autocorrelation plot of the series after first difference

Table 4. ARIMA comparison for different p and q

Model	significance	AIC	AICc	BIC
ARIMA (0,1,1)	MA (1) is significant	4592.52	4592.58	4604.38
ARIMA (1,1,0)	AR (1) is not significant	4593.58	4593.64	4605.44
ARIMA (1,1,1)	AR (1) is not significant MA (1) is not significant	4594.42	4594.52	4610.23
ARIMA (2,1,0)	AR (1) is significant AR (2) is not significant	4591.99	4592.1	4607.81
ARIMA (2,1,1)	AR (1) is significant AR (2) is significant MA (1) is significant	4590.29	4590.45	4610.06
ARIMA (2,1,2)	AR (1) is significant AR (2) is significant MA (1) is significant MA (2) is significant	4583.06	4583.28	4606.78
ARIMA (0,1,2)	MA (1) is significant MA (2) is not significant	4594.16	4594.26	4609.97
ARIMA (1,1,2)	AR (1) is significant MA (1) is not significant MA (2) is significant	4593.31	4593.47	4613.08

Table 5. Estimation results of model parameters with constant terms

	Coefficient	Std_Error	t_Statistic	p_value
AR (1)	0.8167112	0.1117470	7.308574	0.00000
AR (2)	-0.7729223	0.1213154	-6.371180	0.00000
MA (1)	-0.6905470	0.1288739	-5.358314	0.00000
MA (2)	0.6267366	0.1643109	3.814333	0.00016
drift	5.0509899	4.5722045	1.104717	0.26998

Table 6. Estimation results of model parameters without constant terms

	Coefficient	Std_Error	t_Statistic	p_value
AR (1)	0.8095145	0.1087939	7.440810	0.00000
AR (2)	-0.7789152	0.1188439	-6.554106	0.00000
MA (1)	-0.6820876	0.1251442	-5.450413	0.00000
MA (2)	0.6374978	0.1614055	3.949666	0.00009

Table 7. Comparison of criteria in models with and without constant terms

Model	AIC	AICc	BIC
ARIMA (2,1,2) with constant term	4583.06	4583.28	4606.78
ARIMA (2,1,2) without constant term	4582.26	4582.42	4602.03

3.5. Model checking

After the model is constructed, it is necessary to examine whether the residual sequence conforms to the standard of randomness, that is, to prove that the residual terms are independent from each other, so as to show that the residual sequence of the model is a white noise sequence. It can be seen from the autocorrelation diagram of the residual series in Figure 5 that the autocorrelation function of the residual series basically falls in the random interval; Moreover, the Ljung-Box method is used to perform the white noise test on the residual sequence of the model, and the specific results are shown in Figure 6. This shows that the model passes the model significance test, and the ARIMA (2,1,2) model constructed above is effective and has a good degree of fitting. The expression of the constructed ARIMA (2,1,2) is:

$$P_t - P_{t-1} = 0.8095 * (P_{t-1} - P_{t-2}) - 0.7789 * (P_{t-2} - P_{t-3}) + \varepsilon_t - 0.6821\varepsilon_{t-1} + 0.6375\varepsilon_{t-2}$$

$$P_t = 1.8095P_{t-1} - 1.5884P_{t-2} + 0.7789P_{t-3} + \varepsilon_t - 0.6821\varepsilon_{t-1} + 0.6375\varepsilon_{t-2}$$

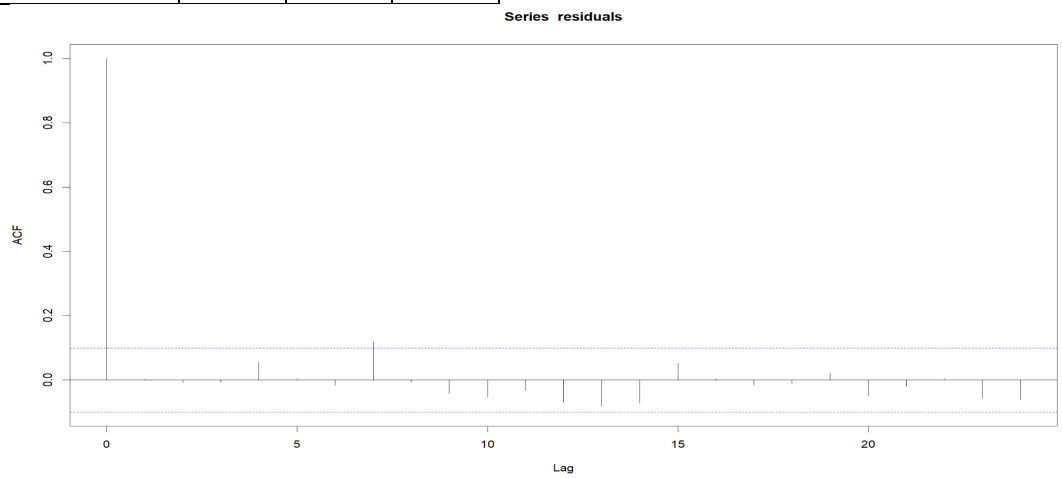

Fig 5 Autocorrelation diagram of residual series

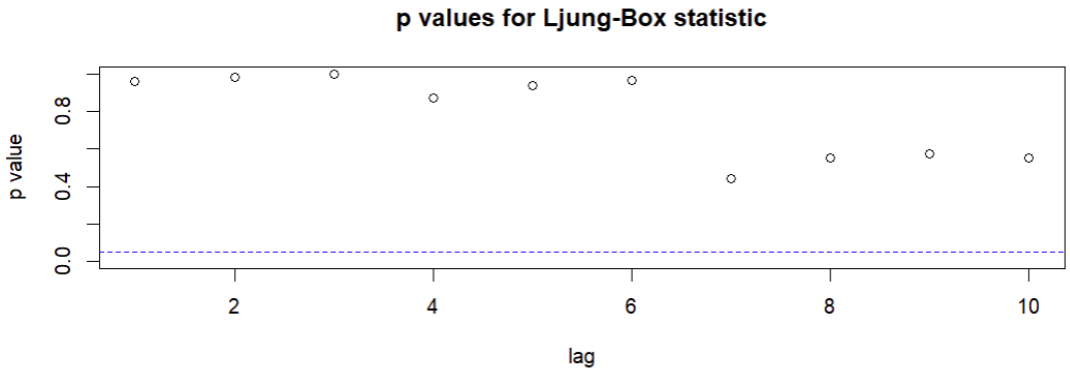

Fig.6 Ljung-Box test results of the residual series

3.6. Model predictive analysis

The constructed ARIMA (2,1,2) is used to predict the price index of the closing price of the Shenzhen Composite Index at the end of the five months from June to October 2023. The specific prediction results, true values and errors are shown in Table 8. After calculation, the mean absolute error (MAPE) of these five months is 2.921%, which is small, indicating that the constructed ARIMA (2,1,2) has a good effect on fitting the price trend of Shenzhen Composite index, and the prediction accuracy is relatively high. However, it can be seen from Table 8 that the farther away the trading day is, the greater the absolute error of the prediction is, which to some extent indicates that the ARIMA model has a better short-term

prediction effect in fitting the stock price index.

Table 8. Price prediction results of Shenzhen Composite Index in the next five periods

Trading day	predicted value	80% confidence lower limit	80% confidence upper limit	95% confidence lower limit	95% confidence upper limit	actual value	absolute error
2023-6-30	2017.772	1899.603	2135.941	1837.047	2198.496	2,049.23	1.535%
2023-7-31	2033.632	1855.555	2211.708	1761.287	2305.977	2,069.51	1.734%
2023-8-31	2035.088	1815.354	2254.822	1699.034	2371.143	1,947.48	4.499%
2023-9-28	2023.899	1776.645	2271.154	1645.757	2402.042	1,910.28	5.948%
2023-10-31	2013.709	1745.307	2282.111	1603.223	2424.195	1,874.51	7.426%

4. GARCH model

4.1. Residual series analysis

From the time series diagram of the residual series in Figure 7, it can be seen that after the influence of deterministic non-stationary factors is eliminated, the fluctuation of the residual series is stable in most periods, but the fluctuation is continuously high in some periods and low in some periods, which indicates that there is variance clustering effect in the residual series. By drawing the Q-Q diagram of the residual sequence, as shown in Figure 8, the upper part and the lower part of the residual sequence are not on the normal curve, showing the characteristics of skewness, so it can be considered that the residual sequence does not conform to the normal distribution, and the residual sequence can be preliminarily determined to have ARCH effect.

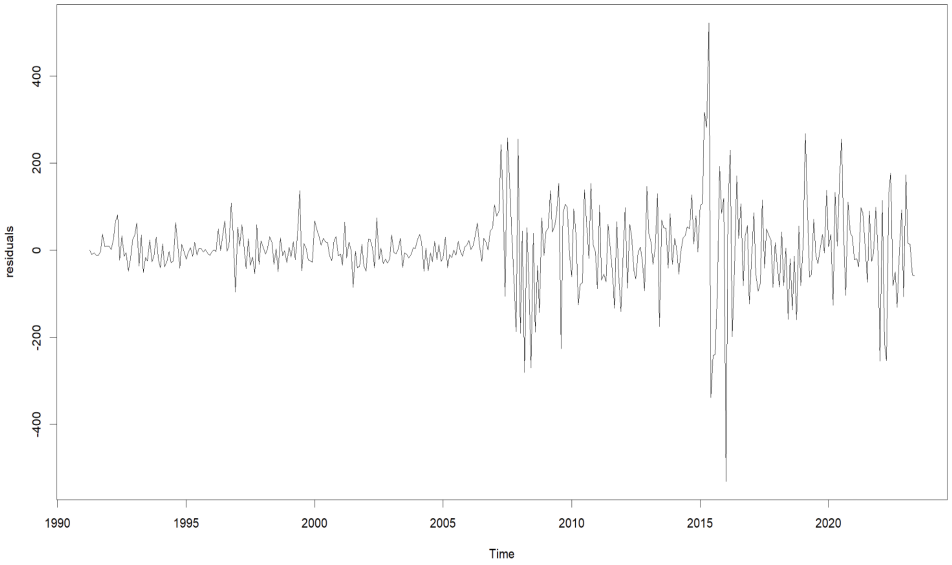

Fig.7 Timing diagram of the residual series

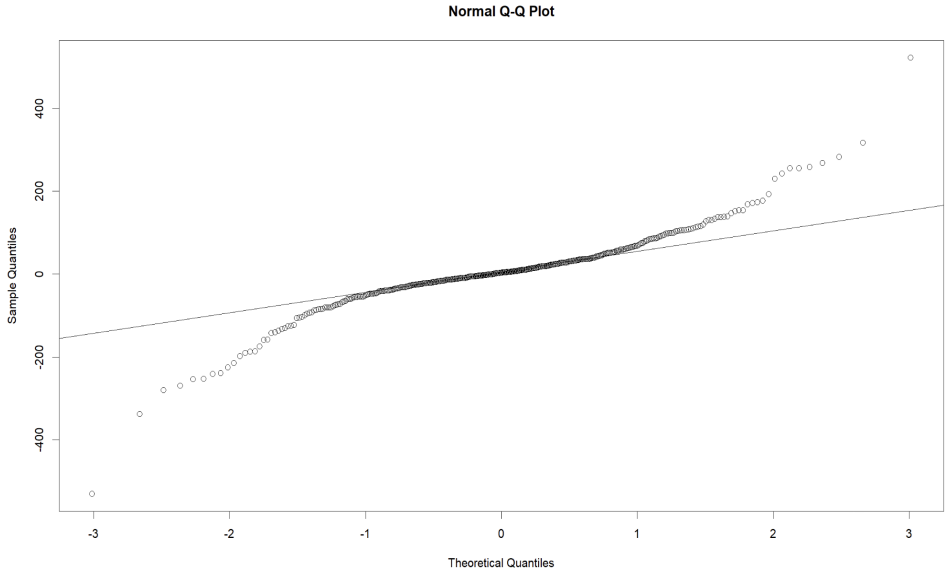

Fig.8 Q-Q plot of the residual series

4.2. ARCH effect test

After extracting the mean information, conditional heteroscedasticity test, also known as ARCH test, is needed for random sequences. The two commonly used ARCH test statistics are Portmanteau Q test statistics and LM (Lagrange multiplier method) test statistics. PQ test (as shown in FIG. 9) and LM test (as shown in FIG. 10) are used to test the serial autocorrelation of the residual series obtained from this series, and it can be seen that there is a P value less than 0.05, which indicates that there is correlation in the squared residual term, that is, the null hypothesis should be rejected, and there is ARCH effect in the residual series of the model. Moreover, it can be seen from the figure that the higher-order P value is still less than 0.05, and the correlation is still significant, which indicates that there is a long-term correlation in the squared residual error series, that is, there is a higher-order ARCH effect in the residual error series. Therefore, GARCH (p,q) model can be considered to extract the correlation contained in the squared residual series.

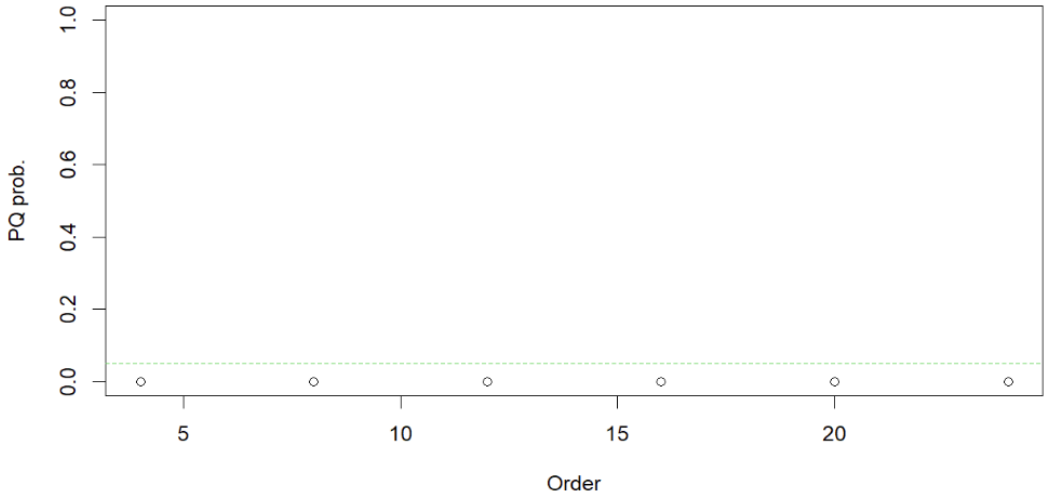

Fig. 9 PQ test diagram

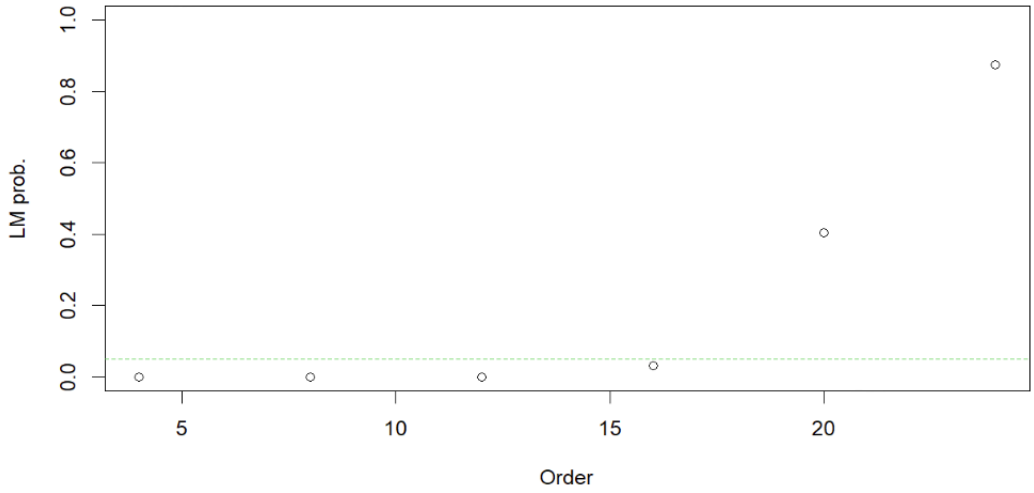

Fig 10 LM test diagram

4.3. GARCH model construction

4.3.1. GARCH model order determination

In order to determine the order of GARCH, the magnitudes of AIC and BIC informativeness of GARCH models with different values of p and q are compared, as shown in Table 9.

Table 9. Comparison of GARCH models of different orders

Model	AIC	BIC
GARCH (1,1)	11.156	11.227
GARCH (1,2)	11.164	11.246
GARCH (2,1)	11.172	11.254
GARCH (2,2)	11.175	11.267

Table 9 shows that when p=1 and q=1, both AIC and BIC are the least informative; therefore, it is most appropriate to choose GARCH (1,1) model to fit the residual series.

4.3.2. Parameter estimation and test of GARCH model

The parameters of GARCH (1,1) model are estimated by the least squares estimation method, and the specific parameter estimates are shown in Table 10. It can be seen from the table that the p-value of each parameter is less than 0.05, indicating that these parameters are significantly non-zero and reasonable.

Table 10. Parameter estimation results of GARMA (1,1) model

	Estimate	Std.Error	t_value	Pr(>|t|)
omega	144.671862	61.107348	2.36750	0.017909
alpha1	0.342641	0.059904	5.71988	0.000000
beta1	0.646353	0.051326	12.78802	0.000000

4.3.3. Model significance test

Whether the GARCH model fits well depends on whether

it fully extracts the heteroscedasticity information contained in the residual series. The Ljung-Box method is used to test the white noise of the residual sequence, and the P value of the residual sequence is greater than 0.05, so the residual sequence can be considered as the white noise sequence, which indicates that the mean model has fully extracted the level information of the sequence. At the same time, the squared residual series has passed the Ljung-Box test, as shown in Table 11, and the P value is also greater than 0.05, which can be considered as the white noise series, indicating that the variance model has fully extracted the volatility related information of the series. The ARCH-LM test is carried out again on the fitted ARIMA-GARCH model, and the results are shown in Table 12. It can be seen that the P value is still significantly greater than 0.05 after the high lag order, which indicates that the null hypothesis cannot be rejected, that is, there is no autocorrelation in the squared residual series after the ARIMA-GARCH model is fitted. That is, there is no ARCH effect in the residual series, and the model has eliminated the conditional heteroscedasticity. The specific expression of the final GARCH (1,1) model is as follows:

$$h_t = 144.672 + 0.343\varepsilon_{t-1}^2 + 0.646h_{t-1}$$

Table 11. Ljung-Box test results of squared residual series

	statistic	p-value
Lag [1]	0.01902	0.8903
Lag [5]	0.77596	0.9079
Lag [9]	1.90365	0.9162

Table 12. ARCH-LM test results of ARIMA-GARCH model residual series

	Statistic	Shape	Scale	P-Value
ARCH Lag [3]	0.000063	0.500	2.000	0.9937
ARCH Lag [5]	0.648600	1.440	1.667	0.8391
ARCH Lag [7]	1.094000	2.315	1.543	0.8978

4.3.4. Forecast of volatility

The volatility forecast is carried out according to the fitted ARIMA-GARCH model, and the volatility forecast results of the next five periods are shown in Table 13. This prediction result shows that the predicted series may face great volatility in the next five periods, and the volatility has a trend of gradually increasing, so investors need to pay attention to risks and do a good job in corresponding risk management. At the same time, the predicted value of yield shows obvious cyclical fluctuations, which may also mean that there is some market uncertainty.

Table 13. Volatility forecast results for the next five periods

Time	Series	Sigma
Jun	-30.55	69.30
Jul	17.18	70.31
Aug	-27.00	71.29
Sep	14.81	72.27
Oct	-23.87	73.22

5. Summary

In this paper, the ARIMA (2,1,2) model is fitted to the stable sequence of Shenzhen Composite index after first difference, and the fitted model is used to forecast the price index of Chinas Shenzhen Composite index in the next five periods. In addition, this paper fits the GARCH (1,1) model for the ARCH effect of volatility cluster in the Shenzhen Composite index, and forecasts the price index volatility of the Shenzhen composite index in the next five periods according to the fitted GARCH model.

The prediction of the price mean and volatility of the Shenzhen Composite Index based on the fitted ARIMA-GARCH model can provide a certain reference for investors. However, as the stock price index is affected by a variety of comprehensive factors, it faces increased uncertainty risk, which will eventually lead to great fluctuations of the stock price index. When investing, investors should not only pay attention to the return rate of stocks, but also pay close attention to the volatility risk of stocks and choose stocks suitable for their risk preferences.

References

[1] Xu, C. & Fang, H. An empirical study on stock price prediction using ARMA model [J]. Economic Research Guide,2019(31):77-82.

[2] XIE Y Y. Empirical research on agricultural stock price based on ARIMA class model: A case study of Longping High-tech Co., LTD. [J]. Hebei Enterprises,2023(2):40-42.

[3] HUANG S M. Stock Price Analysis and prediction based on ARIMA Model: A Case study of China Merchants Bank [J]. Small and medium-sized Enterprise Management and Technology,2022(11):184-187.

[4] Wang, Y. Time Series Analysis: Based on R[M]. Beijing: China Renmin University Press,2015.3:142-143.

[5] GU X. An empirical study on ARIMA model for predicting the closing price of petrochina stocks [J]. Financial Information,2021(12):3-5.