Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 13, No. 3, 2024

Forecasting and Error Factor Analysis of the CSI 300 Index Based on the ARIMA Model

Yumeng Zhang

Department of Finance, Shanghai University, Shanghai, CO 200444, China

Abstract: The CSI 300 stock index futures are futures contracts with the CSI 300 Index as the underlying asset. The CSI 300 Index represents the overall trend of the Chinese stock market, consolidating a representative variety of stocks for analysis. It is generally believed to reflect the highs and lows of the Chinese stock market, with individual stock price anomalies having limited impact on the overall index. Therefore, studying stock price fluctuations through technical means, focusing on the CSI 300 Index as the research subject, is more appropriate. Correspondingly, it also provides guidance for portfolio operations and institutional or fund investments. Specifically, it plays a guiding role in the research of stock index futures based on the CSI 300 Index. Hence, this paper employs the ARIMA model in time series analysis to establish a model for the CSI 300 Index. Utilizing mathematical methods to process the daily CSI 300 data sampled from January 3, 2023, to December 29, 2023, a one-year period, the ARIMA model is developed using the Python programming language based on economic theory and econometric knowledge. By testing, adjusting, and estimating the model, a reasonable economic interpretation of the model is provided to forecast the CSI 300 Index. Through the analysis of the actual results obtained from this model and the reasons for the differences between the actual values of the CSI 300 Index and the model's predicted values, this study aims to offer valuable insights for enterprises and investors in making relevant decisions.

Keywords: ARIMA model, CSI 300 prediction, Error, Influencing factors.

1. Introduction

Stocks first appeared over 400 years ago, emerging alongside the advent of joint-stock companies. The stock market is an inevitable product of a market economy, holding a crucial and immeasurable position in the financial field, and exerting an increasingly profound influence on people's economic status and quality of life. Since its initial trial run in 1989, the Chinese stock market has gradually developed and become an important means to raise funds for key national projects. Moreover, stocks are characterized by their speed, strength, low cost, and adherence to market economic laws when raising construction funds, thereby promoting the rapid development of China's stock market.

Stock investment has become an important means of family financial management and personal wealth management, with the number of Chinese stock investors reaching an unprecedented scale. The mastery of stock investment determines a portion of a family's income. The sharp fluctuations in the stock market can cause significant shocks to the financial market, directly impacting its stability and the healthy development of the economy. However, the prominent feature of the stock market is the coexistence of high risk and high returns. Therefore, the price trend of stocks has always been of widespread concern to the government, entrepreneurs, and investors. Accurately predicting the trend of stock prices and timely intervening in and guiding the stock market rationally will promote the sustained and healthy development of China's economy, minimizing investors' losses and maximizing their profits.

2. Literature References

The time series in business and economics, such as GDP, PPI, stock price indices, unemployment rates, market shares, exchange rates, and interest rates, have long attracted numerous scholars at home and abroad for econometric analysis due to their economic and practical significance. In recent decades, considerable progress has been made in time series modeling and forecasting techniques through the joint efforts of many scholars worldwide. Sargan, Granger and Margenstern [1] provided the random walk model and its various improvements. In 1970, GP.E Box and G.M. Jenkins [2] proposed the Autoregressive Integrated Moving Average (ARIMA) model, which first transforms non-stationary time series into stationary ones through differencing of order d and then identifies the model using autoregressive processes (AR(p) process), moving average processes (MA(q) process), sample autocorrelation coefficients (ACF), and partial autocorrelation coefficients (PACF), among others. This approach has since become the most widely used method for time series forecasting. Mandelbrot [3] observed that many economic series have wide-tailed distributions with varying variances, and large fluctuations tend to concentrate in certain periods while smaller fluctuations concentrate in others. In 1982, Robert Engle proposed the Autoregressive Conditional Heteroskedasticity (ARCH) model [4], which describes the time-varying and clustering properties of financial data variances and has been widely applied, leading to several generalized models. For example, Bollerslev (1986) extended the ARCH model to the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model [5]; Michel and Zakoian in 1994 further extended it to the Threshold Generalized Autoregressive Conditional Heteroskedasticity (TGARCH) model based on the different sensitivity of markets to positive and negative information shocks [6]; Baillie, Bollerslev, and Mikkelsen proposed the FIGARCH model in 1996 based on Engle's ARCH model [7], and so on.

3. Empirical Research

3.1. Data Collection

This study selects the daily closing prices of the CSI 300 from January 3, 2023, to December 29, 2023, excluding holidays, as the daily data (a total of 242 data points). The data source is Netease Finance, and these 242 data points are treated as the data for an entire year, as depicted in Figure 1.

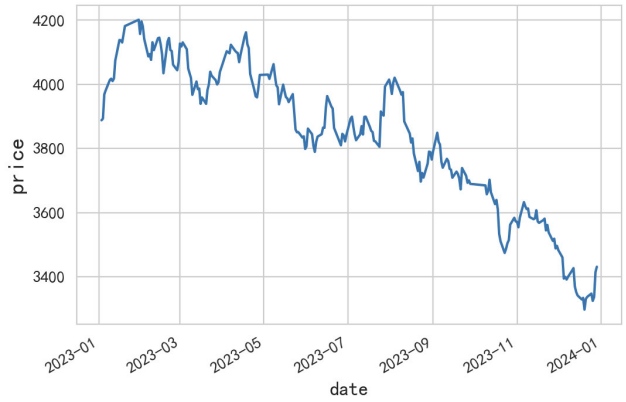

Figure 1. CSI 300 index daily data

3.2. Data Preprocessing

From Figure 1, it can be observed that the approximate trend of the CSI 300 index in 2022 does not exhibit significant cyclic or seasonal trends and has a considerable fluctuation range, indicating a non-stationary time series. White noise tests are conducted on the original data, yielding a p-value close to zero, indicating that the time series is non-white noise. Stationarity tests, commonly performed using strict statistical methods such as autocorrelation, partial autocorrelation tests, and the Augmented Dickey-Fuller (ADF) test, are conducted on the original data. The results of autocorrelation and partial autocorrelation tests are shown in Figure 2.

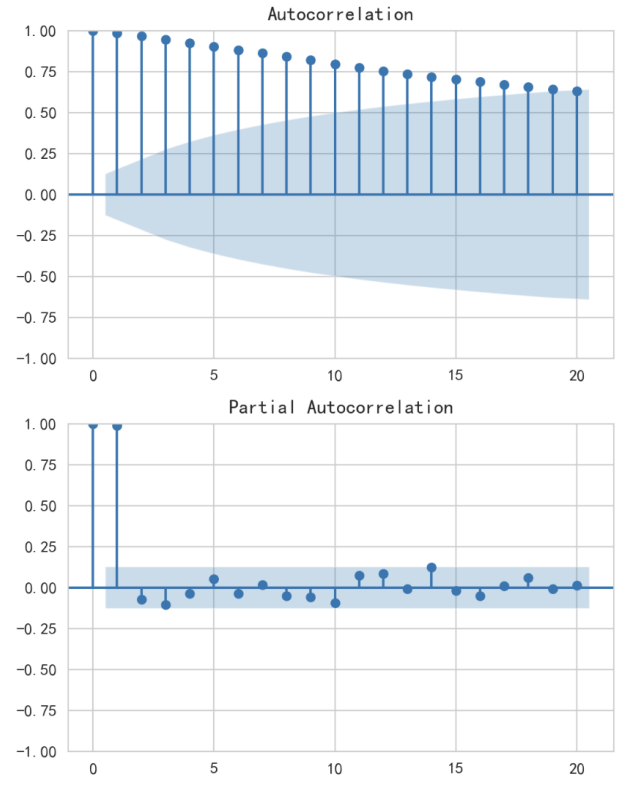

Figure 2. ACF and PACF of the original data

Based on the analysis of autocorrelation and partial autocorrelation results, it is observed that the ACF tails off and the PACF is truncated at the second order, indicating that the original data sequence is non-stationary. A unit root test is conducted on the sequence, with the output results shown in Table 1.

Table 1. ADF test of the original data

	value
Test Statistic Value	-0.411974
p-value	0.90808
Lags Used	0
Number of Observations Used	241

According to the ADF test, the p-value is 0.91, indicating the existence of a unit root, confirming that the sequence is non-stationary. Due to the apparent downward trend followed by an upward trend in the original sequence, in order to eliminate trend factors and satisfy the assumptions of the ARIMA model, differencing is applied to make it a relatively stationary sequence.

After performing first-order differencing on the original sequence, the resulting new time series is depicted in Figure 3, which exhibits no obvious trend or cycle and has stable fluctuations.

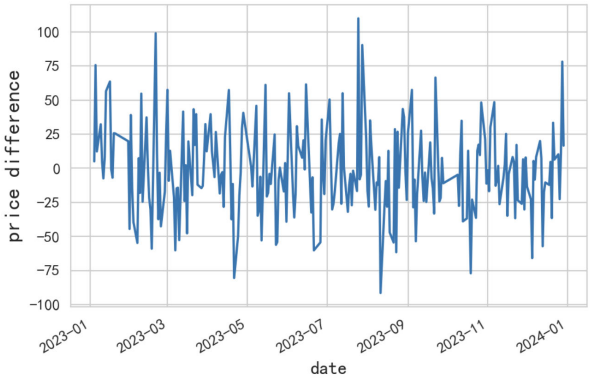

Figure 3. First-order differential sequence

According to Table 2, it can be concluded that after conducting the ADF test on the time series after first-order differencing, the p-value of the test result is zero, indicating that the time series after first-order differencing is stationary, satisfying the prerequisite conditions of the ARIMA model.

Table 2. ADF test of the first-order differential sequence

	value
Test Statistic Value	-14.724756
p-value	0.0
Lags Used	0
Number of Observations Used	240

As shown in Figure 4, based on the results of autocorrelation and partial autocorrelation analysis, it can be observed that the ACF and PACF are truncated. Therefore, it can be inferred that the autocorrelation analysis of the time series after first-order differencing yields a stationary time series.

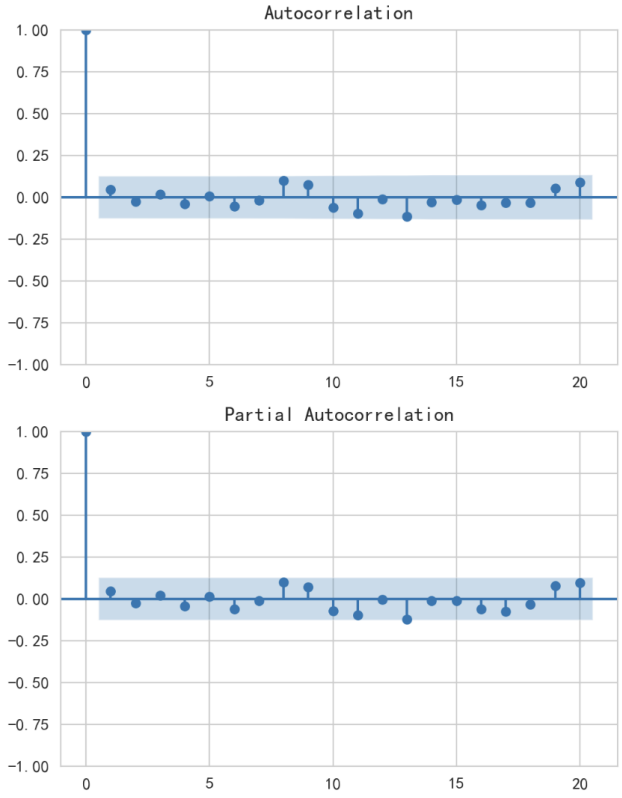

Figure 4. ACF and PACF of the first-order differential sequence

3.3. Model Identification

Based on the above analysis, first-order differencing can transform non-stationary single sequence data into stationary time series. Therefore, this study selects the ARIMA (p, d, q) model to analyze the data. To determine the model order, estimation of the three parameters p, d, and q in the ARIMA model is required. Since first-order differencing is selected for the data, d=1. The Akaike Information Criterion (AIC) can be used to determine the model order, with larger AIC values indicating lower model fit. According to the output results from Python, the optimal p and q values are (3, 3). Therefore, the model with the smallest AIC value is ARIMA (3, 1, 3).

3.4. Model Establishment and Validation

After determining the model parameters, the ARIMA (3, 1, 3) model is established and validated using Python software and the method of least squares. According to the output results, the Durbin-Watson statistic for the model residuals is 1.01, indicating the absence of serial correlation in the residuals. Additionally, a white noise test is conducted on the residual sequence of the model, yielding a p-value of 0.98. Therefore, there is a high probability that the residual sequence is mutually independent, indicating that any factors that could interfere with the prediction results have been eliminated, and the useful information in the residual terms has been fully extracted, rendering the model essentially perfect.

3.5. Application of Model Predictions

After establishing the model, predictions can be made for the prices. Assuming the need to forecast the closing prices of the CSI 300 from January 3, 2023, to December 29, 2024, Python software is used to complete the model predictions. The comparison between the original data and the predicted data is shown in Figure 5. Comparing the predicted CSI 300 index with the actual index, it can be seen from the figure that the predicted values obtained through the model fall within an acceptable range of the actual values. Although there are sometimes significant deviations in the model data, the overall trend predictions are relatively accurate. This indicates that the model predictions can provide relatively reliable investment references for investors.

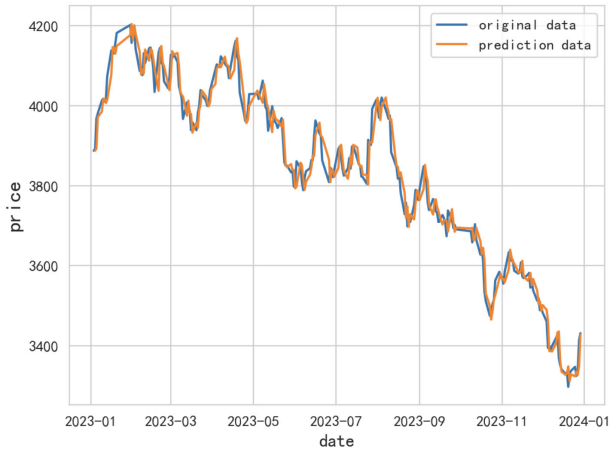

Figure 5. Prediction output results

4. Prediction Deviation Analysis

By comparing the prediction curve with the actual CSI 300 Index, it is observed that significant discrepancies sometimes arise between the actual values and the forecasted values. One aspect of these discrepancies stems from the inherent limitations of the ARIMA model itself and the errors introduced during the model simplification process. Another aspect involves investigating external factors influencing the index. The following focuses on analyzing non-regular factors affecting the index.

4.1. Political Factors

Due to the strong directive capability of the Chinese government's policies on the economy, announcements, cancellations, and modifications of policies and legal documents significantly influence the future direction of the stock market. This influence originates partly from the direct impact of policies on market fluctuations and indirectly from the psychological impact on investors. Investors may develop optimistic or pessimistic sentiments about the future, and these sentiments tend to be relatively amplified. Generally, policies with predictive characteristics do not cause sudden market fluctuations. However, policies with abrupt promulgation processes lead to unnatural and drastic market fluctuations, reflecting as unpredictable fluctuations in the CSI 300 Index.

4.2. Holiday Factors

Due to China's stock trading system, Saturdays, Sundays, and holidays are non-trading days. While market closures themselves do not impact the stock market, empirical observations indicate that significant fluctuations often occur on Fridays, Mondays, and trading days adjacent to holidays, contrary to the overall trend. This phenomenon arises because during non-trading periods, the announcement of new policies may trigger substantial fluctuations on the next trading day. People tend to avoid such risks, opting to sell stocks before holidays, driven largely by psychological influences.

Therefore, this behavior is one of the reasons explaining the differences between actual and predicted values.

4.3. Index Compilation Technical Factors

Compared to the Shanghai Composite Index, the CSI 300 Index has unique selection criteria for its constituent stocks. Firstly, newly listed stocks and ST stocks are not selected as constituent stocks. In other words, once a stock is reclassified from ordinary to ST, it may lose its qualification to enter the constituent stocks and be removed from the original compilation system. During this process, the CSI 300 Index may experience fluctuations due to stochastic factors. However, compared to the aforementioned factors, the contribution of this factor to the deviation of actual index values from predicted values is relatively small. This is because the CSI 300 Index is formulated with the aim of maintaining relative stability and being unaffected by individual stocks. Therefore, fluctuations in index compilation itself are minimal.

5. Conclusion

Due to the non-stationarity, serial correlation, and randomness of the CSI 300 Index, it is suitable for forecasting using ARIMA. However, due to the significant influence of subjective investment psychology and the lack of obvious seasonality in the CSI 300 Index, it is challenging to completely eliminate the non-stationarity of the series during model establishment. In order to simplify the model, this study adopts relatively small values for the p and q parameters. Although the model is simplified, since logarithmic transformation and seasonal differencing are applied directly to the original data before model construction and prediction, important information is retained as much as possible. According to the fitting results and prediction indicators, the predicted results are close to the actual values, indicating a good fit. This model can be utilized for short-term forecasting of the stock market in the near future and serve as an important reference for investors in their investment activities.

Time series, as an effective tool for analyzing financial markets, has significant practical value and development prospects. Exploring the regularities of stock price fluctuations through time series analysis is of great significance for investors in making investment decisions, risk avoidance, and maximizing returns. However, this model only considers the characteristics of the time series itself and conducts some analysis and forecasting of the trend of the CSI 300 Index based on historical data, without considering indicators such as the investment return rate and risk premium of stocks. The study of stock market regularities is complex, with various factors influencing price fluctuations, including not only factors related to the stock market system itself but also factors related to national macroeconomic policies and the direction of economic development. Although these factors are reflected in the model as random terms, they cannot be reflected in the expected values of the forecasts. Further research is needed to refine how to handle raw data and select the optimal model.

References

[1] Sargan, J. D. , C. W. J. Granger, and O. Morgenstern . “Predictability of Stock Market Prices.” Journal of the Royal Statistical Society Series A (General) 135.1(1972):156.

[2] Box, G. E. P. , and G. M. Jenkins . "Time Series Analysis Forecasting And Control." Journal of Time Series Analysis 3.3228(1970).

[3] Mandelbrot, Benoit . “New method in statistical economics.” J of Political Economy 71.5(1963):421-440.

[4] Engle, R. E. . “Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation.” Econometrica 50.50(1982).

[5] Bollerslev, Tim Peter . “Generalized autoregressive conditional heteroskedasticity with applications in finance /.” General Information 31.3(1986):307–327.

[6] [1]Jean-Michel, and Zakoian. “Threshold heteroskedastic models.” Journal of Economic Dynamics & Control(1994).

[7] Baillie, Richard T., T. Bollerslev, and H. O. Mikkelsen. “Fractionally integrated generalized autoregressive conditional heteroskedasticity.” Journal of Econometrics 74(1996).