

Forecasting the NASDAQ Index

– Based on the ARIMA Model

Yuchen Xiang^{1,a,*}

¹Antai College of Economics and Management Shanghai Jiao Tong University, Shanghai, China

a. fx996cnz@sjtu.edu.cn

*corresponding author

Abstract: Forecasting the stock value is an enduring topic for investors. The ARIMA model is an effective model for forecasting proposed by Box and Jenkins in the 1970s. The NASDAQ Composite provides an index for investors to analyze the market of technical companies. Based on the ARIMA model, the author uses the NASDAQ Composite from Jan.2nd, 2020 to Jun.28th, 2024 for forecasting. This paper briefly introduces the NASDAQ Composite, ARIMA model and some judging criteria. The result of the modeling process reflects that the ARIMA (0, 1, 1) can better describe the index. The following actual value of the index also proves the correction of the model. However, as the period of the data is hard to determine, the seasonality of the model is uneasy to find. In future researches, the author plans to figure out a method to determine the days of one period of the index to reveal the index's seasonality, or use models based on ARIMA model for forecasting.

Keywords: ARIMA model, forecast, stationary, NASDAQ Composite.

1. Introduction

Forecasting is always an important part in statistical analysis. People are struggling to identify the properties of certain statistics to make predictions of these datasets. In many fields including finance, medical science and industrial production, the analysis of these datasets to make further predictions is becoming increasingly important. To process these datasets, transforming the data to time series and building models are significant steps. Efficient time series analysis methods can help figuring out the underlying properties of certain time series and benefiting in forecasting.

Founded in 1971, New York City, the NASDAQ (National Association of Securities Dealers Automated Quotations) Stock Market is now the second largest stock market around the world. It focuses more on technology stocks. Almost all stocks listed on the NASDAQ stock exchange are included by the NASDAQ Composite. In some extent, the index of NASDAQ Composite (hereinafter referred as 'the index') can reflect the operating conditions of technology companies, enabling the investors to conduct asset allocation.

Statisticians have been exploring models to analyze the properties of datasets. In 1970s, the ARIMA (Autoregressive Integrated Moving Average) model was proposed by Box and Jenkins, signaling a new era in processing time series datasets. The ARIMA model combines two models (Autoregressive model and Moving Average model) with self-differencing, enabling the model to deal with non-stationary datasets [1].

To better demonstrate the steps of modeling and forecasting, this article would try to apply ARIMA model to the Nasdaq Composite Index. Through the modeling and forecasting process, this paper tries to figure out some drawback of the ARIMA model while dealing with stock value datasets. The model fitting the NASDAQ Composite most is ARIMA (0, 1, 1) according to the paper. For short-time forecasting, the ARIMA model can surely indicate the index's trend. But the preciseness of prediction still has much space to improve.

In this paper, section 2 mainly focuses on the description of NASDAQ Composite and the reasons to choose the index. Section 3 introduces the composition of ARIMA model, methods inspecting the fitness of the model, modeling and forecasting process. Section 4 is some analysis of the results and section 5 is the final conclusions.

2. Data Description

In this paper, the author chooses the NASDAQ Composite as an object of study.

2.1. What is NASDAQ Composite

As a stock market index including almost all stocks listed on the NASDAQ stock exchange, the NASDAQ Composite reflects the overall business situation of companies listed on NASDAQ. Heavily influenced by tech-companies, the index gives investors a broad view to observe the operating situation of tech-companies.

2.2. Why Choose NASDAQ Composite

Information technology companies compose the main body of the NASDAQ Composite, enabling the Index the possession of reflecting whether the technical industry is thriving. Previous studies have demonstrated that as technology firms is the majority on the NASDAQ, moving to it may help the firm to attract the attention of investors in technology field [2], proving that studying NASDAQ Composite's valuable. Helping the investors to make better decisions on asset allocation is also an advantage of observing the NASDAQ Composite.

2.3. Data Resources

This paper imports data from Yahoo Finance. The data used for modeling is the NASDAQ Composite Index from Jan.2nd, 2020 to Jun.28th, 2024.

3. Methods and Analysis

In this section, the author would briefly introduce the components of ARIMA model, fitness judgement methods and the process of modeling.

3.1. Introduction of ARIMA Model

Known as the Box-Jenkins method [3], the ARIMA model was generalized by ARMA (autoregressive moving average) model in 1970s by applying a differencing step to ARMA model. By using differencing, a non-stationary dataset can be stationed for modeling and forecasting.

3.1.1. AR Processing

The AR (autoregressive) processing indicates the predicted values relies on its own lagged values. The AR model needs one positive integer p to determine the model.

The formula (1) shows an example of AR(p) model.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t \quad (1)$$

In formula (1), y_t is one predicted value, $y_{t-1} \dots y_{t-p}$ are lagged values of y_t . ε_t is white noise. Other parameters are constants.

3.1.2. MA Processing

The MA (moving average) processing of ARIMA model indicates the predicted regression error is past errors' linear combination. One positive integer constant q is required.

Formula (2) is an example of MA(q) model.

$$\varepsilon_t = c + \varepsilon_{t-1} \theta_1 + \varepsilon_{t-2} \theta_2 + \dots + \varepsilon_{t-q} \theta_q \quad (2)$$

In formula (2), ε_t is the regression error of predicted values, $\varepsilon_{t-1} \dots \varepsilon_{t-q}$ are errors in the past, $c, \phi_1, \phi_2 \dots \phi_p$ are constants, ε_t is white noise.

3.1.3. ARMA Model

Combining the AR processing and MA processing, it's natural to build ARMA (p, q) model, like formula (3) shows.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (3)$$

A dataset's stationarity can be checked by Autocorrelation function (ACF). For stationary series, the ACF value would decay quickly with the lag increasing. In contrast, the ACF value of non-stationary series will decrease slowly [4].

There are many techniques for ARMA order selection, including the well-known Akaike information criterion (AIC), Bayesian information criterion (BIC) [5]. Both the methods will be introduced in the following text.

ACF and PACF (partial autocorrelation function) can be used to select the suitable model from AR, MA and ARMA model. Table 1 demonstrates simple criteria to determine the model selected and its parameter using the ACF and PACF values adapted from [6].

Table 1: Select the model and determine the value of p and q using ACF and PACF.

Model	ACF	PACF
AR(p)	Decrease slowly	Cuts off after lag p
MA(q)	Cuts off after lag q	Decrease slowly
ARMA(p, q)	Decrease slowly	Decrease slowly

3.1.4. Differencing Processing (I in ARIMA)

The propose of the differencing (integrated) processing of ARIMA model is stationing the non-stationary original dataset for further processing. Previous studies have shown that the dataset's mean and variance can be stationed using differencing, reducing the non-stationarity of the dataset [7].

3.1.5. ARIMA Model

The above three models can only be applied to stationary datasets. A method is needed to eliminate the non-stationary of original datasets. For non-seasonal datasets, use the d^{th} order differencing for stationing is enough. For seasonal datasets, seasonal differencing have to be conducted before d^{th} order differencing.

Three non-negative integer constants (p, d, q) are needed to determine the exact ARIMA model [8]. ARIMA (p, d, q) (without seasonal differencing) is like formula (4) demonstrates.

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) (1 - L)^d y_t = \left(1 + \sum_{j=1}^q \theta_j L^j\right) \varepsilon_t \quad (4)$$

In this example, the operator L means lag, like $Ly_t = y_{t-1}$.

For seasonal ARIMA, it needs four more parameters (P, D, Q, m), m is the number of observing times during one season of dataset.

ARIMA (p, d, q)(P, D, Q)_m is like formula (5) indicates.

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) \left(1 - \sum_{i=1}^P \Phi_i L^{m*i}\right) (1 - L)^d (1 - L^m)^D y_t = \left(1 + \sum_{j=1}^q \theta_j L^j\right) \left(1 + \sum_{j=1}^Q \Theta_j L^{m*j}\right) \varepsilon_t \quad (5)$$

3.2. Measurements for Goodness of Fit

This part introduces some measurements to examine the fitness of the ARIMA model.

3.2.1. Akaike Information Criterion

Formulated by statistician Hirotugu Akaike, AIC (Akaike information criterion) is a method to judge the quality of given models. Considering the versatility of AIC, it's practicable to use AIC for the quality check of a given ARIMA model [9].

For a certain model, formula (6) presents its AIC value.

$$AIC = 2k - 2 \ln(\hat{L}) \quad (6)$$

In the criterion, $\hat{L}$ is the maximized value of the likelihood function for the model. In ARIMA models, parameter k is often set as $p + q + 1$.

3.2.2. Bayesian information criterion

Developed by Gideon E. Schwarz, BIC (Bayesian information criterion) is a criterion for model selecting. The BIC is similar to AIC but it gives a greater punishment to the number of parameters used in the model.

Formula (7) is the BIC value of one certain model.

$$BIC = k \ln(\hat{n}) - 2 \ln(\hat{L}) \quad (7)$$

In the criterion, k and $\hat{L}$ is equal to k and $\hat{L}$ in AIC. n is the total number of data points used in the model.

3.2.3. Root Mean Square Error

RMSE (Root Mean square error) is the square root of MSE (mean squared error), like formula (8) shows.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

y_i is the value of actual data and $\hat{y}_i$ is the value of forecasted data.

Taking the root of MSE can produce a result with the same units as the original data, which naturally represents the standard error for the normal distribution error of the original data [10].

RMSE can be used in model selection. A given data set can be divided into two datasets, training set and test set. The training set is used for modeling and the test set is used to calculating the RMSE value. Models having the smallest RMSE are often the best fitting models.

3.3. Main Procedures of Modeling

In this section the author explains the main steps of modeling and forecasting using ARIMA model.

3.3.1. Forehead Preparation

The author imports the data and transforms the data to a time series. The author set 21 days as a period of the data for the reason that the NASDAQ Composite is not continuous and a month has only 21 trading days in average. Here is the data used for modeling (Figure 1).

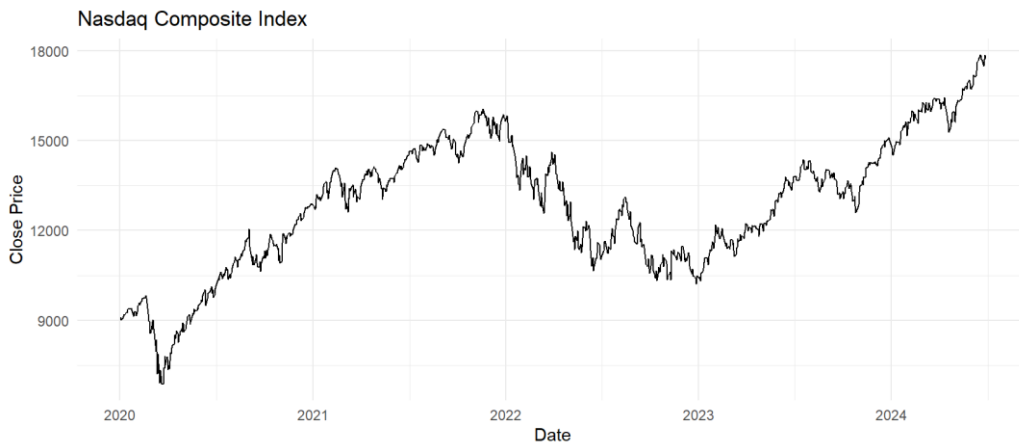


Figure 1: NASDAQ Composite from Jan.2nd, 2020 to Jun.28th, 2024 (Picture Credit: Original)

3.3.2. Stationary Check

Before modeling, the stationary of the index has to be checked. Only stationary data can proceed to the next step of modeling. Figure 2 shows the ACF value and the residuals of the index.

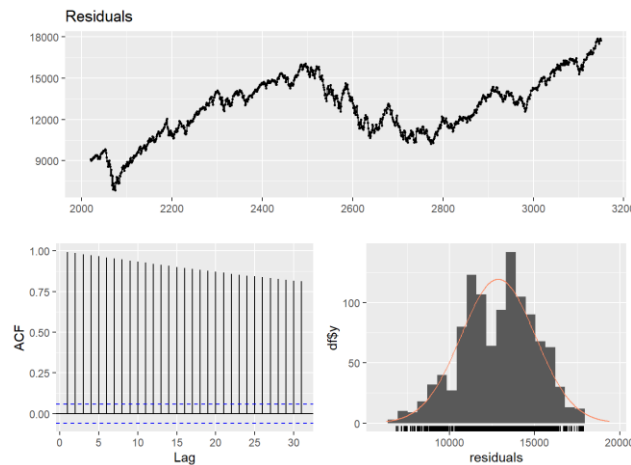


Figure 2: ACF value and the residuals of original data (Picture Credit: Original)

Most residuals do not fall within the normal distribution, indicating the original data's non-stationary. The Ljung-Box test shows that the p-value of the index is significantly smaller than 0.05, indicating there is no clear evidence to prove the stationary of the data. Differencing steps on the time series needs to be done further modeling.

3.3.3. Model Selection

There are some measures to determine the values of p , d , q , P , D , Q .

As the plot of the time series shows no obvious seasonality, non-seasonal ARIMA model can be applied. The R package *forecast* provides a function *ndiffs()* to determine the differencing times of the time series. The d value is equal to the result of *ndiffs()*, $d = 1$. Figure 3 displays the ACF and PACF values of the time series used to determine the p and q values.

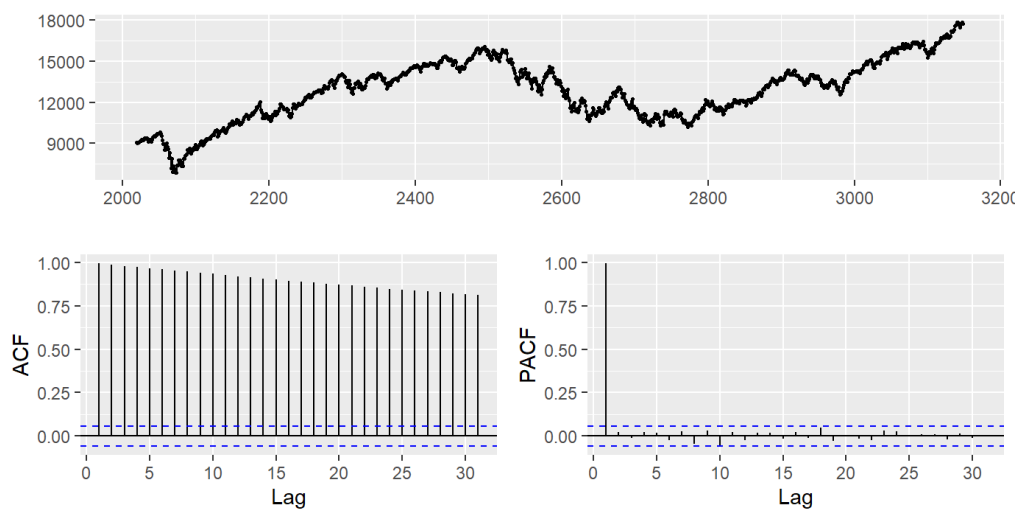


Figure 3: ACF value and PACF value of the index (Picture Credit: Original)

The ACF value of the index decreases slowly as the lag increasing, indicating the parameter p is zero. While after lag 1, the PACF value drops dramatically, giving that the q value equals to one.

In another way, the R package *forecast* provides a simple function *auto.arima()* to help build ARIMA model. The function will compare the AIC, BIC, RMSE and other judging methods to select an ARIMA model fitting the index best. The *auto.arima()* function also chooses ARIMA(0,1,1) with drift as fitting model. With an AIC value of 15011.44, a BIC value of 15026.53 and RMSE is 186.0609, the ARIMA (0, 1, 1) has the smallest AIC, BIC and RMSE compared to other models.

3.3.4. Forecast

The author uses the model selected to predict the NASDAQ Composite in next 40 days. Then the author use function in R package *ggplot2* to output the result of forecasting (Figure 4).

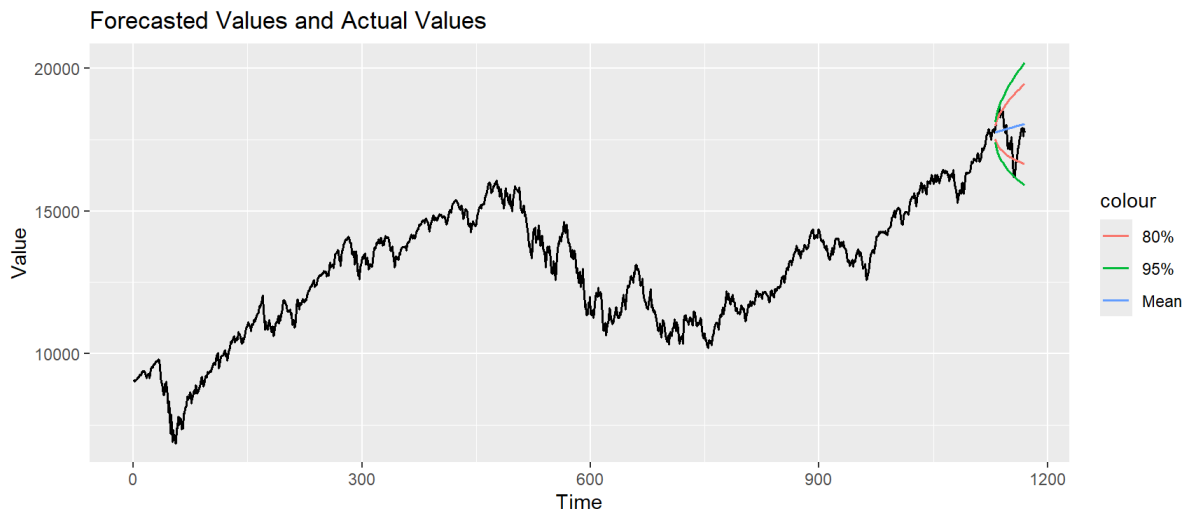


Figure 4: The forecasted values and actual index (Picture Credit: Original)

In Figure 4, different lines have different meanings. The black line shows the actual value of NASDAQ Composite, the blue line is the average forecasted values of the Index using ARIMA (0, 1, 1). 80% confidence interval are signaled by red lines, and the green lines point out the limits of 95% confidence interval.

4. Result Analysis

Both the data selection methods choose the ARIMA (0, 1, 1) model, indicating that this model is probably the most effective model fitting the time series.

In Figure 4, almost all of the actual data of NASDAQ Composite falls within the 95% confidence interval, most of the actual data falls within the 80% confidence interval. The result reveals that the model is essentially possible for forecasting future data. The mean value of forecasted values, however, can only point out the vague trend of the index, showing no peaks and troughs. The forecasted values are difficult to use as a reference for investors.

5. Conclusion

Forecasting the stock value is a goal pursued by every investor. The NASDAQ Composite is set as a weathercock of the market of technology companies. The analysis demonstrates that the ARIMA (0, 1, 1) model fits the NASDAQ Composite from Jan.2nd, 2020 to Jun.28th, 2024 extracted from yahoo finance best. The model does reflect some properties of the index and revealing the trend of the index, suggesting ARIMA model's effectiveness and simplicity. However, there are some limitations of the model. First, the number of days in a period is hard to determine. This causes the seasonality of the index may be hidden during the process of modeling. The index probably has an underlying seasonality hard to find through the figures. Second, the uncertainty of the stock values. The stock values may be affected by numerous reasons, adding the uncertainty of the index. In further researches, the author may try to determine the period of the index and apply seasonal ARIMA model to the NASDAQ Composite, or apply other models based on the ARIMA model for further forecasting.

References

- [1] Khan, S., Alghulaiakh, H. (2020). ARIMA Model for Accurate Time Series Stocks Forecasting. *International Journal of Advanced Computer Science and Applications*, 11.

- [2] Dang, V.A., Michayluk, D., Pham, T.P. (2018). *The curious case of changes in trading dynamics: When firms switch from NYSE to NASDAQ*. *Journal of Financial Markets*, 41. 17-35
- [3] Khan, F.M., Gupta, R. (2020). *ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India*. *Journal of Safety Science and Resilience*, 1. 12-18.
- [4] Schaffer, A. L., Dobbins, T. A., & Pearson, S. A. (2021). *Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions*. *BMC medical research methodology*, 21(1), 58.
- [5] Moon, J., Hossain, M.B., Chon, K.H. (2021). *AR and ARMA model order selection for time-series modeling with ImageNet classification*. *Signal Processing*, 183.
- [6] D. Banerjee. (2014). *Forecasting of Indian stock market using time-series ARIMA model*. 2014 2nd International Conference on Business and Information Management (ICBIM), Durgapur. 131-135
- [7] Huang, C., Petukhina, A. (2022). *Stationary Time Series Models*. In: *Applied Time Series Analysis and Forecasting with Python*. Statistics and Computing. Springer, Cham. 71-106
- [8] Fattah, J., Ezzine, L., Aman, Z., El Moussami H., Lachhab A. (2018). *Forecasting of demand using ARIMA model*. *International Journal of Engineering Business Management*, 10.
- [9] Zhao, L.X., Li, Z.Y., Qu, L.L. (2022). *Forecasting of Beijing PM_{2.5} with a hybrid ARIMA model based on integrated AIC and improved GS fixed-order methods and seasonal decomposition*. Elsevier, 8.
- [10] Hodson, T.O. (2022). *Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not*. *Geoscientific Model Development*, 15. 5481-5487