Modelling and Forecasting Stock Returns

Financial Econometrics

Lappeenranta University of Technology

Master's Degree in Business Analytics

0000098821 Prashant Shrestha

2022

Table of contents

1 Introduction

This section gives the information about topic of the research, motivation for research, research questions and research objectives.

Forecasting and modelling the financial market has always been a matter of interest among the market researchers and investors. Several successful investors and investment firms use various linear and nonlinear analytical models to forecast the financial markets. The appropriate modelling of the financial market can help investors to optimize their trading, investing and hedging decisions. This can help the investors to maximize their profit. Moreover, the model can provide the insight to decision maker and guide them in making suitable decisions.

The problem statement of the research is "Is it possible to predict the NASDAQ composite using ARIMA model?". The research helps to check the serial correlation in the time series stock price of NASDAQ composite. The prediction of the NASDAQ index will help the investors and market participants interested in technology companies like Apple, Amazon, Meta etc. The objective of the research is listed as below:

1. Testing the stationarity of the univariate time series data of NASDAQ composite.
2. Finding the differencing order of the data at which it would be stationary.
3. To estimate and check the ARIMA model.
4. To predict the future values of the NADAQ using the estimated model.

2 Literature review

This section will give the insights how other authors approach the problems and the empirical framework used by them in their research.

Most of the literature on modelling and forecasting univariate time series model are focused on finding the best fit ARIMA model and forecasting for gold prices, oil prices, inflation rates, electricity consumptions, stock price predictions and commodity derivative markets.

In research paper “Modelling and forecasting of NCDEX AGRIDEX”, the author has focused on identification of the best fit stochastic ARIMA model and forecasting of National commodity and Derivatives Exchange Ltd index. The author has used Augmented dicky fuller test to check for the stationarity. To identify the order of ARIMA, correlogram and the information criteria like AIC, BIC, adjusted r squared, Root Mean Squared Error and Mean Absolute Percentage Error (MAPE) are used. The author has used the historical data of the National Commodity and Derivatives Exchange Ltd to forecast the future values. The author has used out of sample forecasting. The author suggests that the time series data is essentially the random walk process and cannot be forecasted using historical values. (Reddy 2020)

In research paper “ARIMA Model to Forecast the ISX60 Indicator: An Applied Study an Iraqi Financial Market” the author has used ARIMA model to forecast the Iraqi Financial Market which is comprised of 102 companies listed in the Iraq Stock Exchange. The author did not use correlogram and any test to check for stationarity. The author uses line graph of the original time series to justify the stationarity in the time series. The author uses correlogram to determine the order of the ARIMA model. The author uses in sample method for forecasting. (Chyad 2021)

In research paper “Comparison of Forecasting Energy Consumption in Shandong, China Using the ARIMA Model, GM Model, and ARIMA-GM Model” the author has used ARIMA model and GM model to predict the energy consumption in Shandong, China. The author has use augmented dicky fuller test for testing the stationarity. Moreover, the correlogram is used to find the order of the ARIMA model. The author has used in-sample method for forecasting. The author has used grey model (GM model) with ARIMA model to obtain better accuracy of the prediction. Moreover, the author argues that GM-ARIMA model has higher precision than each single models. (Li 2017)

3 Empirical Framework

In this section of the research the important concepts are discussed that are used for forecasting the NASDAQ stock prices. In the research, the stock return of the NASDAQ composite is modelled using ARIMA model. The dataset of NASDAQ composite is a univariate time series with no explanatory variable. Unlike regression model and VAR model, ARIMA model do not need any explanatory variable to forecast and predict the data

so, ARIMA model is used to forecast future stock return using the historical stock return. Furthermore, the research is focused to find if the NASDAQ composite can be forecasted using the ARIMA model. To estimate and use ARIMA model there are various important concepts we need to define.

3.1 Stationarity Condition

Stationarity is a desirable property of an estimated ARIMA model. A given time series is supposed to be stationarity when the observations are not dependent on the time. In other words, if the distribution of the values remains the same as the time progress, then the time series is stationary. Stationarity can be categorized into two categories i.e., weakly stationary process and strictly stationary process. Strictly stationary process has constant mean, constant variance and constant autocovariance structure whereas weakly stationary process has constant mean, constant autocovariance but do not have a constant variance. (Brooks 332)

Stationarity of the time series can be tested by Dicky-Fuller Tests for unit root. The null hypothesis of the test is the presence of the unit root, and the alternative hypothesis of the test is stationarity in the time series. (Brooks 448)

H0 - Time series contains a unit root

H1 - Time series is Stationary

3.2 White noise, Moving average and Autoregressive processes

White noise process is one with no noticeable structure. The white noise process has constant mean and variance and zero autocovariances except at lag zero. In other words, each observation is uncorrelated with all other values in the sequence. (Brooks 333)

Moving average model is the one where the current values of the variable depend on the current and previous values of the white noise disturbance term. The moving average model of order q is denoted by MA(q). The model can be expressed in equation as below. (Brooks 336)

$$y_t = u + \sum_{i=1}^{q} Q_i \, u_{t-i} + \, u_t$$

An autoregressive model is one where the current value of variable depends upon the values that the variable took in previous periods plus the error terms. The autoregressive model of order p is denoted by AR(p). The model can be expressed in the equation as below. (Brooks 341)

$$y_t = u + \sum_{i=1}^{p} \phi_i \, y_{t-i} + \, u_t$$

3.3 Autocorrelation and Partial Autocorrelation function

The autocorrelation functions measure the correlation between the current values and the set of past values. Similarly, the partial autocorrelation function measures the correlation between an observation k period ago and the current observation, after controlling the observations at intermediate lags. For example, the partial autocorrelation for lag 3 would measure the correlation between y_t and y_{t-3} after controlling for the effect of y_{t-1} and y_{t-2}. (Brooks, 349)

3.4 ARIMA processes

ARMA processes is the combination of AR(p) and MA(q) models. The model states that the current values of some series y depend linearly on its own previous values and the combination of current and previous values of a white noise term. Moreover, the term I in the ARIMA model refers to integrated, it is the differencing order for the time series to be stationary. So, ARIMA model can be denoted as ARIMA (p, d, q) where p, d, and q are the auto regressive order, differencing order and moving average order. An ARMA (p, q) model in the variable differenced d times is equivalent to an ARIMA (p, d, q) model on the original data. The model can be written as the equation below. (Brooks, 361)

$$y_t = u + \, \phi_1 y_{t-1} + \, \phi_2 y_{t-2} + \ldots + \phi_t y_{t-p} + \, \theta_1 u_{t-1} + \, \theta_2 u_{t-2} + \cdots + \, \theta_q u_{t-q} + u_t$$

The characteristics of the AR, MA and ARIMA processes can be summarized as below:

1 AR has geometrically decaying ACF and non-zero points of PACF as AR order.

2 MA has geometrically decaying PACF and non-zero points of ACF as MA order.

3 ARIMA has geometrically decaying ACF and PACF.

3.5 Building ARMA Models: Box-Jenkins Approach

Box – Jenkins Approach of model building consists of 3 steps:

3.5.1 Identification

The first step for building the ARIMA model is to identify the order of the model required to capture the feature of the data. The plots of ACF and PACF i.e., correlogram can be used to determine the order of AR and MA. (Brooks 358) The plots cannot be useful in all the situations so the information criteria can be used to identify the order of AR and MA. There are three popular information criteria that can be used to determine the order i.e., Akaike's information criterion (AIC) and Bayesian information criteria (BIC) and Hannan-Quinn criterion (HQIC). Furthermore, adjusted r squared can also be used as the information criterion. (Brooks 360)

3.5.2 Estimation

In second step, the model identified in first step is used estimate the parameters. This can be done using least squares or any other technique like maximum likelihood, depending on the model. (Brooks 358)

3.5.3 Diagnostic checking

Lastly, the model is checked for overfitting. Deliberately, using the larger model can help to check if the suggested model i.e., step one is appropriate. If extra terms in the larger model are insignificant the model in step 1 is adequate. Similarly, the residual errors are also diagnosed for linear dependency. The linear dependency reveals the inability of the model to capture the features of the data. For this purpose, ACF, PACF and Ljung-Box test can be used. (Brooks 358)

3.6 Forecasting

3.6.1 In-sample forecasting and Out-of-forecasting

The forecasting can be done by using two different methods i.e., In-sample forecasts and Out-of-sample forecasts. In-sample forecasts implies the predictions generated for the same set of data that was used to estimate the model's parameters. On the other hand, the forecasts that are predictions generated from the holdout samples are known as Out-of-sample forecasts. (Brooks 368) In this research the in-sample method is used for forecasting.

3.7 Accuracy of the forecast

The mean squared error (MSE) and Mean Absolute Error (MSE) is used to evaluate the accuracy of the model. The MSE and MAE is compared with those of other models for the same data and forecast period and the model with the lowest value of the error measure would be argued to be the most accurate.

4 Data

The research is related with the Nasdaq composite stock index. Nasdaq composite stock index is one of the important stock indexes alongside the S&P 500 and Dow Jones industrial average. The index primarily focuses on the technology stock likes Apple, Alphabet, Amazon, Meta etc. The data used in the research is a univariate time series data of daily stock prices of Nasdaq composite. The data is gathered from the yahoo finance. The time series data starts from 1st November 2020 to 4th November 2022. There are altogether 506 data points in the time series data. The adjusted closing price is used in the research for modelling the stock price. The out-of-sample forecast is used for the forecasting. The research forecast the future stock prices from 10th October 2022 to 4th November 2022 for each day. There will be altogether 19 predictions. Moreover, the predictions will be based on the past data from 1st November 2020 to 7th October 2022 using the best fit ARIMA model.

The mean adjusted closing price of the NASDAQ composite is 13397 USD. The data is divided into high and low stock price at median of 13539 USD. Furthermore, on average the stock price deviate by 1437 USD from the average. The skewness value of the distribution is -0.1975 which suggest that the distribution is slightly negatively skewed. The histogram below also shows the slightly negatively skewed distribution. The h value of the Jarque-Bera test is 1 and the p value is 0.0013, so the null hypothesis can be rejected at 5% significance level. So, the distribution does not come from the normal distribution.

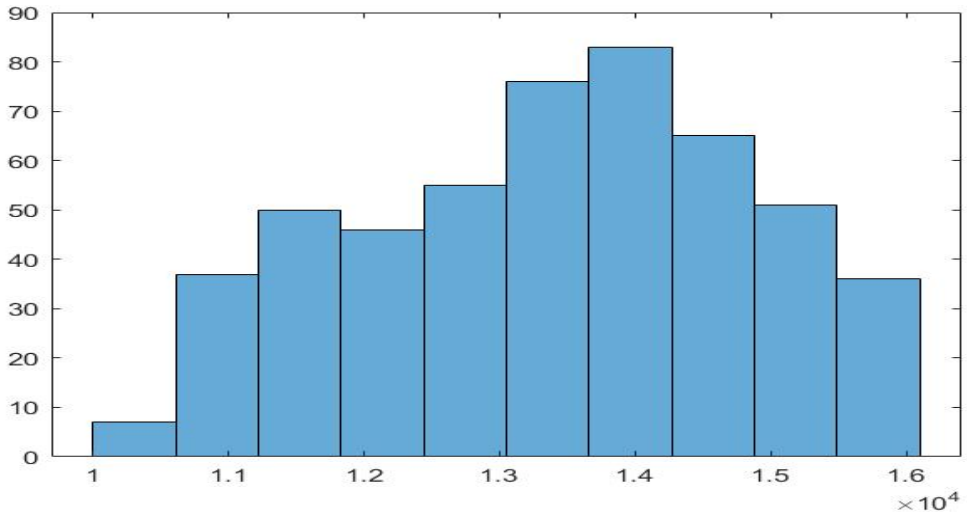

Figure 1 Histogram NASDAS composite

5 Results and Findings

5.1 Stationarity in Time Series Data

The first step for forecasting the time series data using ARIMA model is checking for stationarity. The time series must be stationary for the results to be valid. We can check the stationarity of the data using Augmented Dicky fuller test. The h and p value of the Augmented Dicky fuller test is 0 and 0.5715 so we cannot reject the null hypothesis of time series containing unit root. Therefore, the time series is non-stationary. The figure below shows the line graph of the stock price of NASDAQ composite from 1st November 2020 to 7th October 2022.

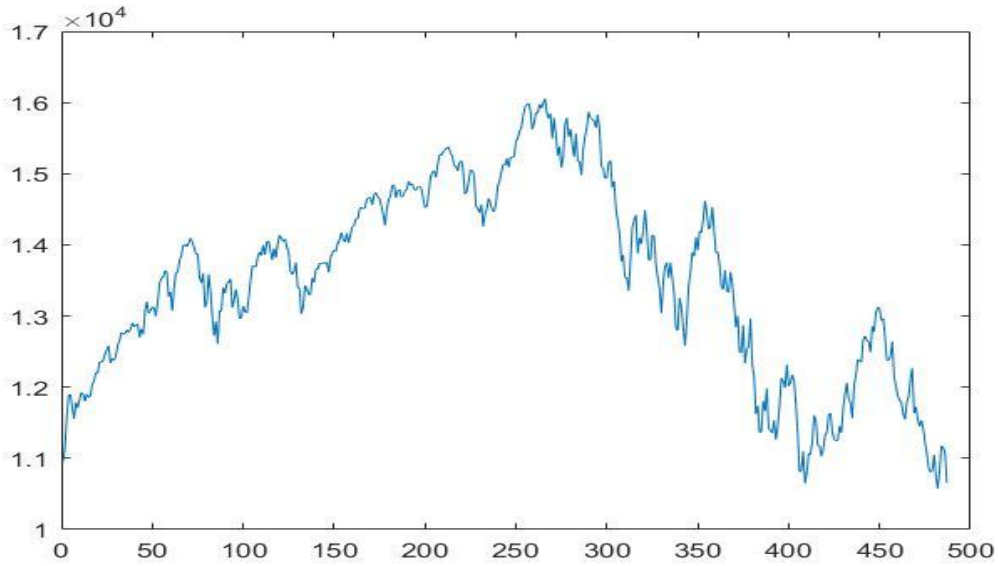

Figure 2 Original time series for training dataset 1st Nov 2022 - 7th Oct 2022

We can make the time series stationary by differencing the values in the time series. We can generate new series in first degree form. We can again use Augmented Dicky fuller test in the differentiated training dataset. The and p value of the test is 1 and 1.0000e-03. We can reject the null hypothesis at 5% significance level. So, the time series is stationary and can be used for the estimation of ARIMA model. The figure below shows the line graph of the differentiated stock price of NASDAQ composite from 1st November 2020 to 7th October 2022.

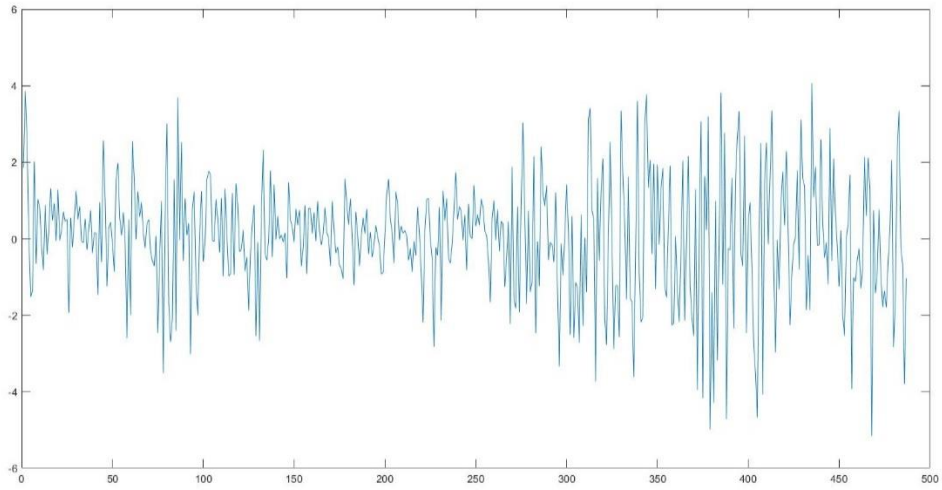

Figure 3 Differentiated time series for training dataset 1st Nov 2020 - 7th Oct 2022

5.2 Building ARIMA Models

5.2.1 Selection of ARIMA (p, d, q) model and Model Estimation

Figure 4 shows the correlogram of NASDAQ stock prices at first degree differences. From the figure we can see that the ACF and PACF values are random from lag 1 to 20. Moreover, there is similar pattern between ACF and PACF. The figure does not fit characteristics of any model i.e., AR, MA and ARMA. This figure suggests that the time series is a random walk process, and we cannot forecast the stock price of the NASDAQ by examining the historical stock prices.

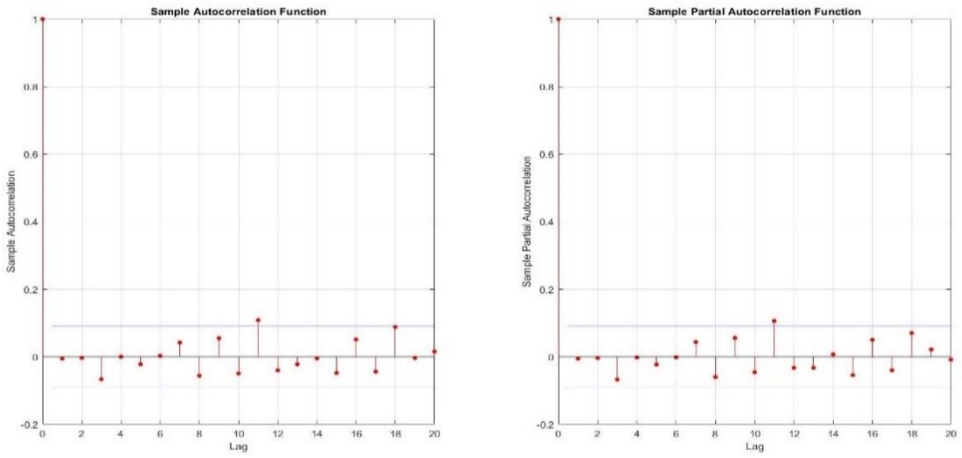

Figure 4 Correlogram of NASDAQ at First Degree Difference 1st Nov 2020 - 7th Oct 2022

However, we can use the information criteria i.e., low AIC and low BIC to choose the best ARIMA model among different combination of the ARIMA model with maximum order being 5 for both AR and MA. Furthermore, RMSE, MSE, MAE and variance of the model is used to compare the accuracy of the model. The ARIMA model with lowest AIC is ARIMA (4,1,5) at 6.5494e+03 and the ARIMA model with lowest BIC is ARIMA (0,1,0) at 6.5649e+03. Moreover, the ARIMA (4, 1, 5) has the lowest RMSE, MSE, MAE and variance than ARIMA (0,1,0). Therefore, the ARIMA (4, 1, 5) is the best fit model than ARIMA (4, 1, 5).

Table 1 information criteria and accuracy statistics on testing data

ARIMA	AIC	BIC	RMSE	MSE	MAE	Variance
(0, 1, 0)	6556.5	6549.40	262.9019	109910.00	76.0572	40992
(4, 1, 5)	6549.40	6595.4	256.4055	58817.00	55.6382	38767

5.2.2 Diagnostics of ARIMA model

The information criteria i.e., minimum BIC is used to select the ARIMA (4, 1, 5) model for the time series of NASDAQ. Figure 5 shows that all the information has been captured in the model. The ACF and PACF in the figure shows that there is no autocorrelation in the lags. We can also use the Ljung-Box Q-test to show if there is autocorrelation in the residual. The h and p value of the Ljung-Box Q-test is 0 and 0.9318. We cannot reject the null hypothesis of no residual autocorrelation. Therefore, there is no autocorrelation in the residuals.

The p value for the constant, AR (2), MA (2) and MA (5) are statistically insignificant at 5% significance level. All other coefficients are significant at 5% significance level. However, the ARIMA (4, 1, 4) increases the number of insignificant coefficients from 3 to 5. Based on the estimation results of ARIMA (4, 1, 5) the model can be expressed as equation below:

$$\Delta Y_t = 1.73 - 0.38y_{t-1} + 0.091\, y_{t-2} - 0.31\, y_{t-3} - 0.77\, y_{t-4} + 0.38\, e_{t-1} - 0.10\, e_{t-2} + 0.29e_{t-3} + 0.87e_{t-4} - 0.021\, e_{t-5} + e_t$$

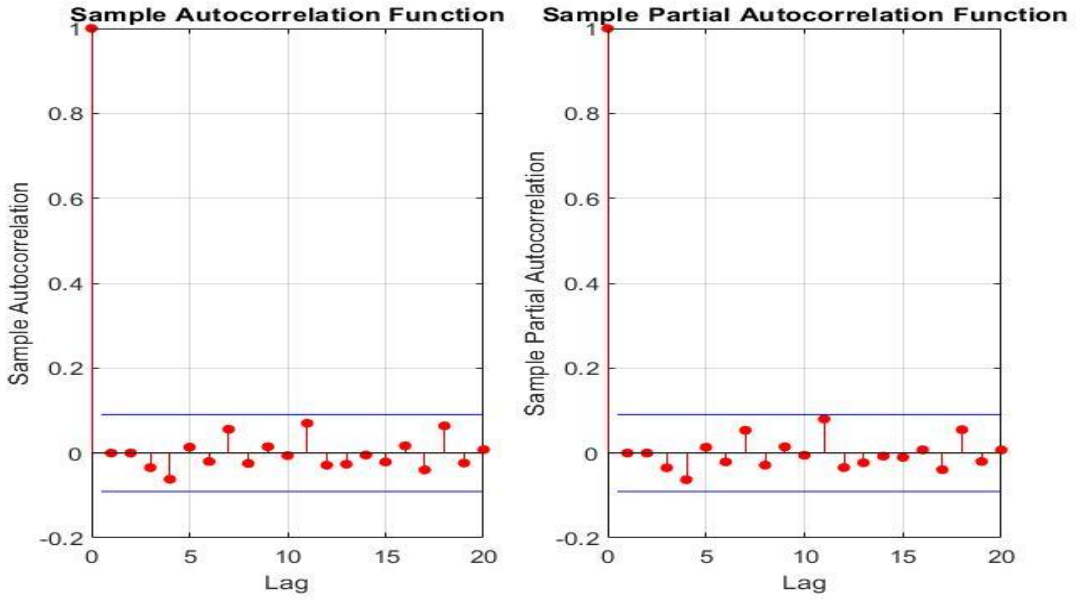

Figure 5 Correlogram of Residuals from ARIMA (4, 1, 5)

5.3 Forecast of NASDAQ composite

For forecasting the NASDAQ composite, in-sample forecast method is used. The historical data from 1st November 2020 to 7th October 2022 is used as the training data for the ARIMA model. The data is predicted for 19 days from 10th October 2022 to 3rd November 2022. The actual data is used to validate the direction and the size of the prediction.

Figure 6 shows the forecasting result of ARIMA (4, 1, 5) which is the best model for forecasting the future closing price of the NASDAQ composite. The figure depicts the daily predicted values from 10th October 2022 to 3rd November 2022 along with the 95% confidence boundary.

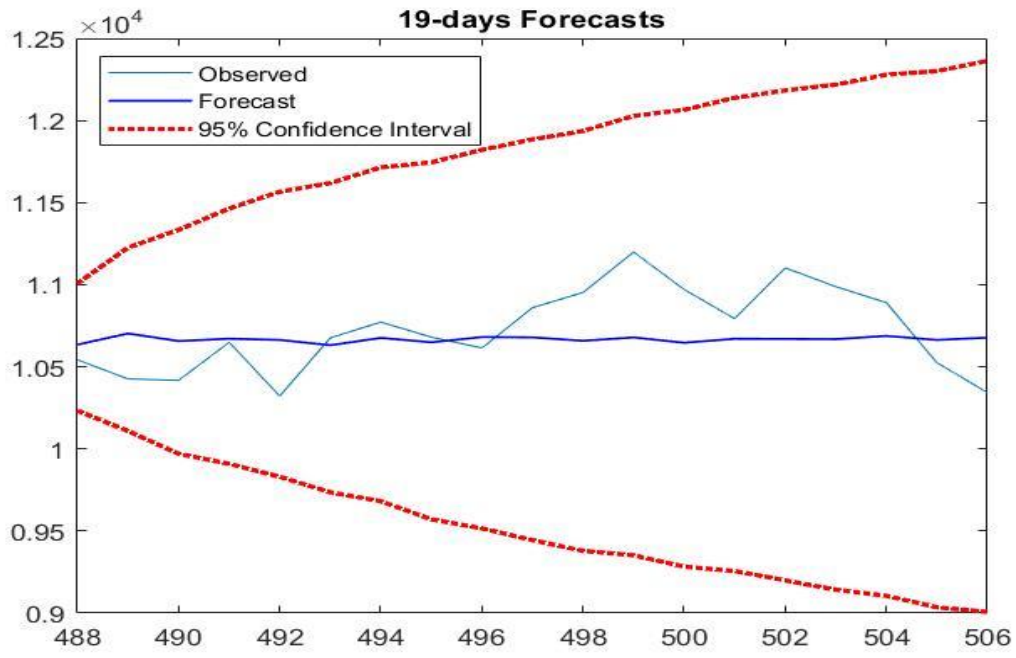

Figure 6 Graphical presentation of forecasted data

Table 2 shows the predicted values, actual values along with high and low values at 95% confidence level. The actual values increase from 10542.099 to 10988.150.150 however, the model forecasted the decrease in the stock price from 10620.9 to 10684.98. The model predicted the adjusted closing stock price to remain almost constant with slight increase and decrease. The model is unable to predict direction nor size of the stock price of NASDAQ for the next day accurately.

Table 2 Forecasted results of NASDAQ closing price 10-Oct 2022 to 3-Nov 2022

Date	Actual values	Forecast	High	Low
10/10/2022	10542.09961	10620.09	11007.49	10232.69
11/10/2022	10426.19043	10667.28	11225.81	10108.76
12/10/2022	10417.09961	10652.21	11334.16	9970.252
13/10/2022	10649.15039	10686.31	11463.58	9909.048
14/10/2022	10321.38965	10697.82	11565.89	9829.763
17/10/2022	10675.79981	10677.09	11618.39	9735.799
18/10/2022	10772.40039	10697.96	11714.67	9681.256
19/10/2022	10680.50977	10658.13	11744.85	9571.402
20/10/2022	10614.83984	10668.67	11821.26	9516.082
21/10/2022	10859.71973	10664.57	11884.78	9444.365
24/10/2022	10952.61035	10657.51	11936.04	9378.993
25/10/2022	11199.12012	10689.49	12027.14	9351.826
26/10/2022	10970.99023	10673.92	12064.81	9283.033
27/10/2022	10792.66992	10697.27	12138.28	9256.267
28/10/2022	11102.4502	10691.15	12182.95	9199.356
31/10/2022	10988.15039	10680.45	12218.33	9142.568
01/11/2022	10890.84961	10692.05	12278.67	9105.435
02/11/2022	10524.79981	10668.19	12300.71	9035.678
03/11/2022	10342.94043	10684.98	12362.67	9007.283

6 Conclusions

The main aim of the research paper was to study if it is possible to forecast the stock prices of NASDAQ composite using ARIMA model. For this purpose, the data related to closing stock prices of NASDAQ from 1St November 2020 to 3rd November 2022 was collected from yahoo finance. ARIMA (4, 1, 5) was chosen among the 36 models because of minimum BIC. The model was used to predict the stock prices from 10th October to 3rd November 2022. The model failed to predict size and the direction the stock prices correctly for the next day. Moreover, the correlogram of the data also shows that the time series represent the random walk process, and it is not possible to predict the stock price of NASDAQ with historical prices. Therefore, it is not possible to forecast the stock prices of NASDAQ

composite using ARIMA model with historical prices from 1St November 2020 to 7th October 2022

This paper can be used by other researcher to extend the research to multivariate time series forecasting by using various explanatory variables like inflation, unemployment rate, economic growth rate, dividends amount etc. This research paper can help to describe the process for testing the stationarity in the time series using ADF test. Moreover, this paper also has the information regarding the process for selecting the best fit ARIMA model. Furthermore, the model can be used by other researcher to create the ARIMA model to predict the stock price for different periods or steps ahead. Therefore, this paper is useful for researchers, investors, and other market participants not to get insights on future values of the stock prices but as a resource to understand the limitation of ARIMA model.

The data that is used is collected for the year 2020 to 2022. There was COVID and Russia-Ukraine War that caused economic instability in the major economies which might have impacted the stock price of the NASDAQ. This might have resulted in the unpredictability of the time series of NASDAQ. This also proves that the ARIMA model is unable to predict and forecast result when there are large variations in data due various external factors.

List of References

Printed Sources

Brooks. C. 2019. Introductory Econometrics for Finance. United Kingdom: Cambridge University Press.

Chyad. K. A. 2021. ARIMA Model to Forecast the ISX60 Indicator: An Applied Study an Iraqi Financial Market. Iraq: University of Baghdad.

Li. S. 2017. Comparison of Forecasting Energy Consumption in Shandong, China Using the ARIMA Model, GM Model, and ARIMA-GM Model. East China: China University of Petroleum.

Reddy. 2020. Modelling and Forecasting: NCDEX AGRIDEX. SCMS. India: Journal of Indian Management.