Time-Series Analysis and Forecasting of S&P500 Index Based on ARIMA and ETS Model

Hairong Zhang
College of Art and Science, New York University, New York, US
hz2773@nyu.edu

Abstract. The study embarks on an insightful journey into the world of stock indices through the deeper understanding of time series analysis application, especially during the special period of oil price volatility. The goal of this paper is to utilize sophisticated statistical models, including Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing (ETS), to extract meaningful information from historical stock index data of S&P500 by Yahoo Finance, enhancing predictive accuracy and operating forecasts to inform strategic decision-making. In addition to the prediction results, the study also compares the forecasts of the two models through some values, and concludes that the prediction effect of the ARIMA model is better than that of the ETS model. The reason behind this result has a lot to do with the processing of the data itself and the fit of the models. For S&P500 during the period of the study, the ARIMA model's prediction result is better. The insights derived from this analysis are expected to empower investors, researchers, and market analysts with a deeper understanding of the stock index's past behavior and its implications for future performance.

Keywords: Stock Index, S&P500, Time-Series Forecasting Analysis, ARIMA Model, ETS Model.

1. Introduction

1.1. Research Background and Significance

The global financial markets have evolved significantly in recent decades, becoming increasingly interconnected and complex [1]. One of the world's most influential economy, the United States, has been at the forefront of this transformation. Its representative stock index, S&P500, has a profound impact on both domestic and global market dynamics. Understanding the development behavior of S&P500 is crucial for investors, policymakers, and researchers to do further research and analysis, since the stock index often serves as a sign of the broader economic health of the US and even the whole market, especially during some special periods, such as oil price volatility, war, epidemic period, and so on [2-4]. Traditional analysis methods can be valuable, but they often fail to predict future trends effectively due to the inherently dynamic and non-linear nature of the stock index. This needs the adoption of more sophisticated analytical methods. One such method is time series analysis, which takes into account temporal dependencies and fluctuations in the data over time, providing a deeper understanding of patterns and predictive insights.

1.2. Literature Review

As a representative stock index that has been developed for a long time, S&P500 has been analyzed by many scholars and researchers as a research object, and they use different methods to do the analysis and forecasting of it. For example, two researchers, Zahid Iqbal and Shaikh A. Hamid, use artificial neural network method to do the analysis and forecasting of S&P500 [5]. Han Lin Shang uses time series analysis of functions and some statistical methods to do the forecast of 1-day-ahead intraday stock returns in her research, and the same 5-minute intraday forecasting is also done for the S&P500 to verify this forecasting method [6]. And Eric Glenn Chan also uses time series analysis and simulation methods to do the prediction of S&P500, and he also do the comparison between these methods to get the results that the Double Exponential Smoothing is better than ARIMA for his research [7].

1.3. Research Contents and Framework

The study mainly conducts the time series analysis and forecasting of S&P500 stock index. The data used covers the 20 years from 1999 to 2019, and the time series forecasts are made under the background of oil price volatility events from 2012 to 2017. The methods and models used in this study mainly include two time series models, Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing (ETS). First of all, the data are processed accordingly, such as differencing operation and logrithm transformation, to achieve stationarity of the series, so as to facilitate subsequent research. Then, predictions are operated by ARIMA model and ETS model respectively. After obtaining specific prediction results and some numerical values, it is also necessary to compare the two models. This allows for more analysis of the prediction results and the selection of the optimal model. Even though the whole research is centered on the S&P500 stock index, the research is still pretty significant because it has a strong representation and importance for the entire the US stock market and even the whole financial market.

2. Methods

The research is focused on using time series analysis to examine the fluctuations and trends of S&P500 of the United States for 20 years, from 1999 to 2019. And from 2012 to 2017, the volatility of oil price, especially the oil price plunge between 2014 and 2016, had a major influence on the economy and the capital markets of the US, including the performance of S&P500 [8]. Fluctuations of oil prices, driven by factors such as geopolitical tensions, supply-demand imbalances, and global economic conditions, influenced various sectors and the whole market sentiment. So to do the forecasting through ARIMA model and ETS model, the 5-years data getting from Yahoo Finance are used, from 2012 to 2017 [9]. The study centers on stock market of the United States, represented by one main stock index, S&P500.

The Autoregressive Integrated Moving Average (ARIMA) model is a technique that is widely used for time series forecasting. It combines the three main components to the model to predict time series data, including autoregressive (AR), moving average (MA), and the differencing part. It is particularly useful for capturing complex temporal patterns, such as trends, seasonality, and autocorrelation, in various types of data, including stock indices. The formula of the model is written as ARIMA (p, d, q). In the formula, the order of autoregressive component is represented by “p”, the order of differencing is shown by “d”, and “q” indicates the moving average order. To choose the most suitable combination of “p”, “d”, and “q”, analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the data is needed, since the performance of the lags in these plots can help to give some instructions about deciding the values of these three components. These plots help identify the presence of autocorrelation and guide the selection of model orders.

The Exponential Smoothing (ETS) model is another time series forecasting method that models the underlying components of a time series: error, trend, and seasonality. It is particularly useful for capturing and forecasting patterns in data that exhibit seasonality and trends, and it is also widely used in forecasting tasks involving stock indices. In more specific terms, the component of error is the result of noise or random fluctuations in the time series, the trend component represents the overall direction and tendency of the data during a period of time, and the component of seasonality indicates the repetitive parts that appear by a fixed frequency, including daily, monthly, or yearly. So choosing appropriate formulation of ETS model should depend on the characteristics of the data, such as the presence of trends and seasonality.

To conduct this analysis, the data need to be stationary, which is the assumed premise of the models, especially the ARIMA model. To be more specific, the mean, variance, and auto-correlation structure should all be zero for the time series. If there is no stable expectation, no constant variance, and it is very highly correlated with residuals, the logarithm transformation should be operated on the data. After taking the natural logarithm of asset returns, the resulting distribution will follow a normal distribution. The one-time differencing can also be operated on the data to make it stationary, which is equal to the value of the closing price minus the opening price of each day. The purpose of this is to try to eliminate any trends in the data so that the new series can have a constant mean and variance.

3. Results

From 1999 to 2019, the S&P500 index has experienced many unstable periods, and the forecasting of it can be operated by a 5-years period from 2012 to 2017. Through the analysis of past data, predicting the future trend and market conditions of the S&P500 index will provide an important judgment for the future development direction of the whole US stock market.

Firstly, the raw data need to be processed efficiently. If testing the original data directly, obviously it is not a good object for analysis. To be more specific, as shown in Figure 1, since the ACF does not show a gradual decline and the PACF also does not have a sharp cutoff after lag, differencing may be necessary, and the data might not be stationary here. In addition, the residuals show a clear trend, so it indicates some problems. To be more specific, the underlying structure of the series may have not be captured by the model at this time, and differencing might be necessary. In the Normal Q-Q plot, the residuals also depart significantly from the reference line, which indicates the non-normality in the residuals. So differencing can help to improve the modeling process here.

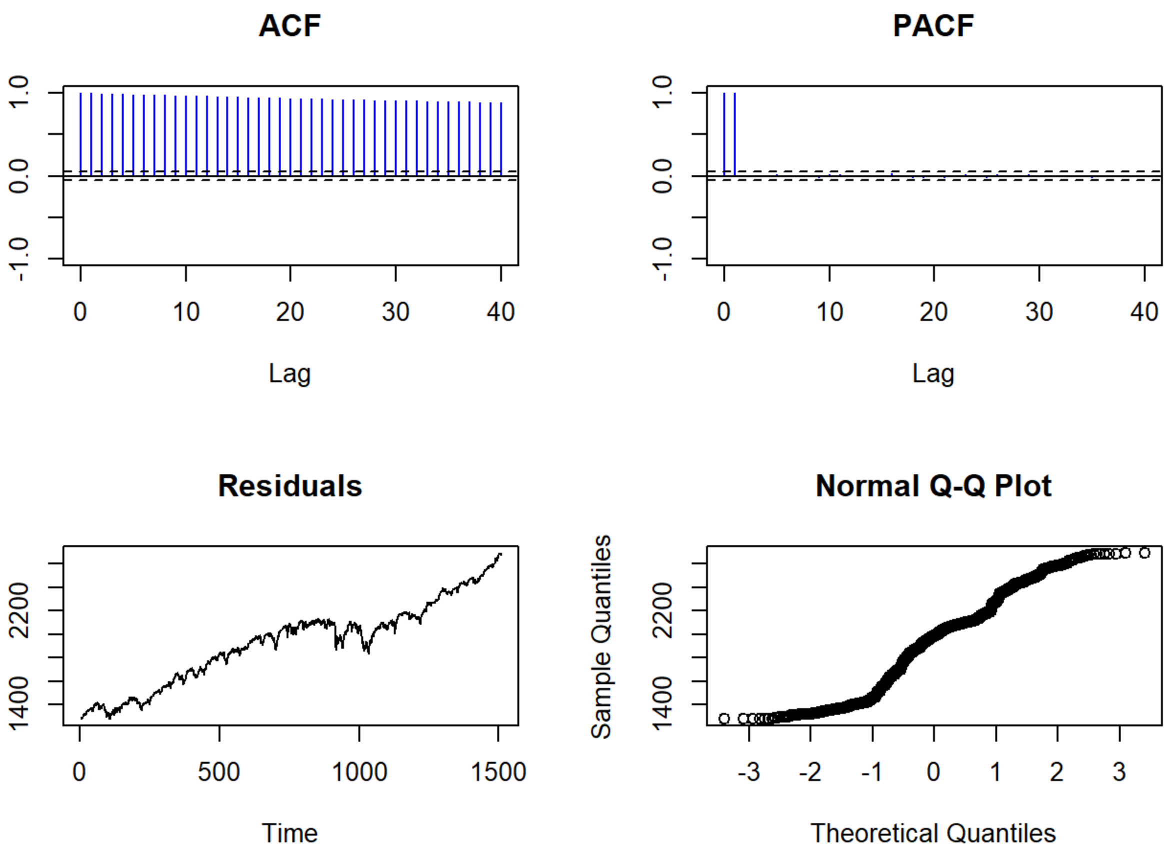

Fig. 1 Results of testing raw S&P500 data

After applying one-time differencing operation here, based on Figure 2, the autocorrelation values in ACF drop quickly and the partial autocorrelation in PACF shows a sharp cutoff after lag 1, it suggests that one-time differencing has successfully removed any linear trends in the data. In the Normal Q-Q plot, the residuals follow a relatively straight line, indicating approximate normality. However, the residual plot still shows some unstable patterns, although it looks much more random than before.

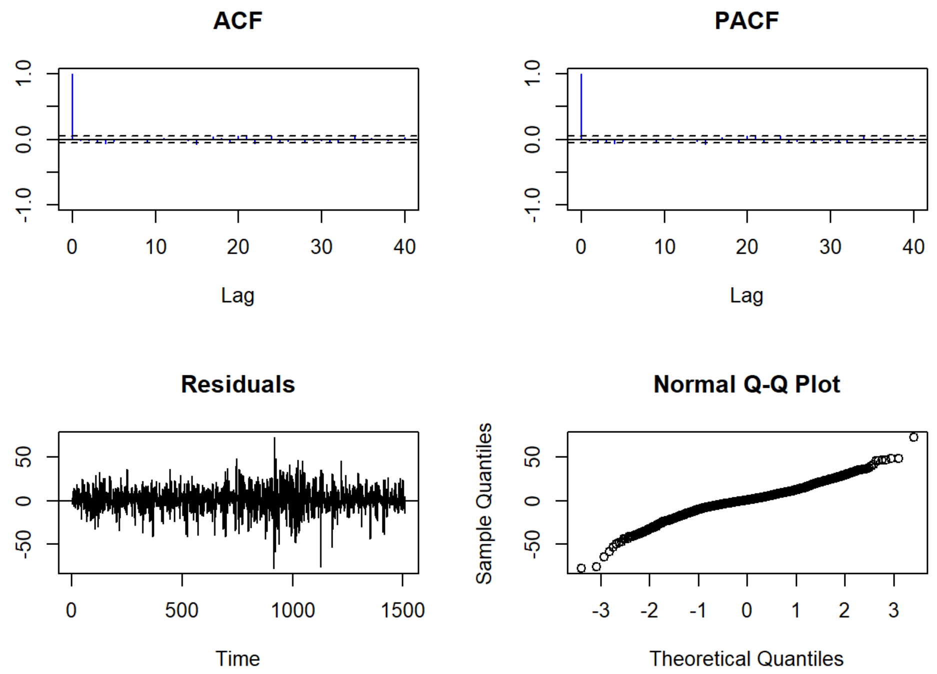

Fig. 2 Results of testing S&P500 data with one-time differencing operation

The logrithm transformation of the original data can also make the data to be more stable and feasible. However, only operating log transformation of the original data is still not enough, because after the test, the same problems as the original data will appear, including ACF, PACF pattern does not meet the prediction assumptions requirements, and residuals problems exist. The test results are shown in Figure 3.

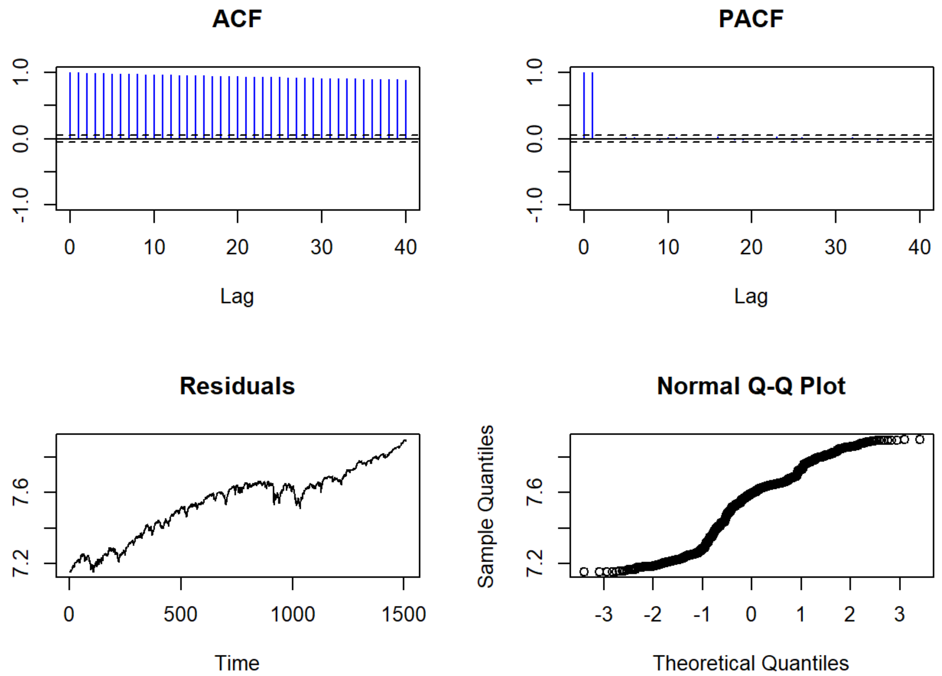

Fig. 3 Results of testing S&P500 data with logrithm transformation

Therefore, one-time differencing operation needs to be applied again on the data with logrithm transformation. The results in Figure 4 show the feasibility of the processed data. To be more specific, both ACF and PACF show a sharp cutoff after lag 1, suggesting that the one-time differencing has successfully removed any linear trends in the data. Also, the residuals look almost random and relatively stable. And the residuals now follow a relatively straight line in the Normal Q-Q plot, indicating approximate normality of it. Based on these analysis, the differenced data has achieved stationarity, so the data has been already prepared to make subsequent time series predictions.

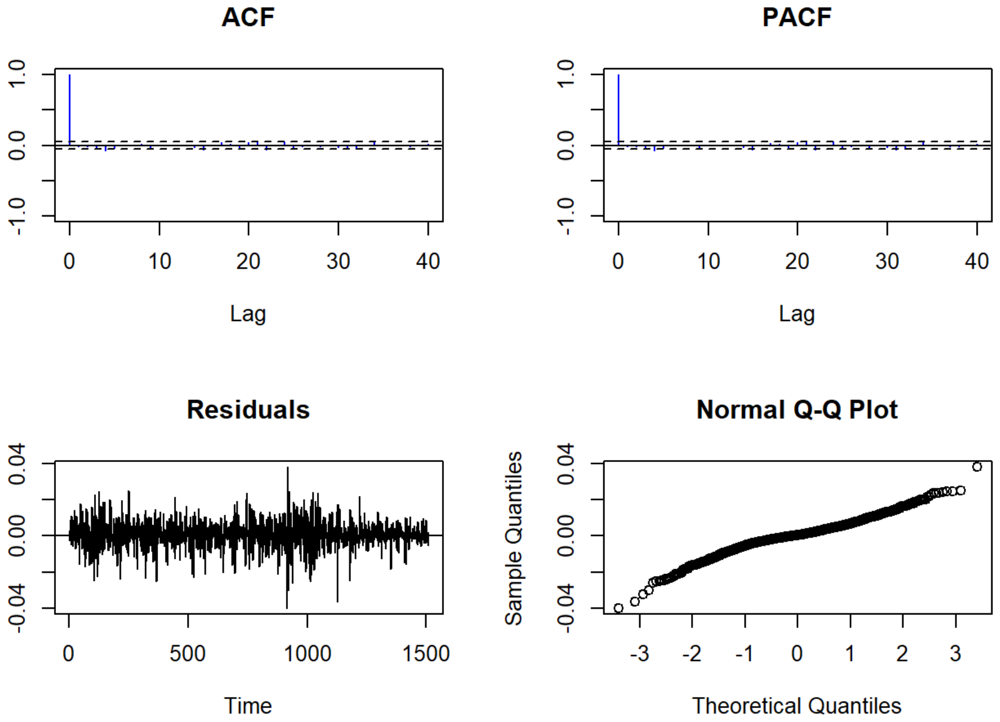

Fig. 4 Results of testing S&P500 data with one-time differencing and logrithm transformation

The first time series model is ARIMA model. The first step of operating the model is dividing the data into training and test sets so that the effect of the prediction can be tested by this way. Here, for the data, 80% are assigned as the training set, and similarly, the rest 20% are assigned to be the test set, as shown in Figure 5.

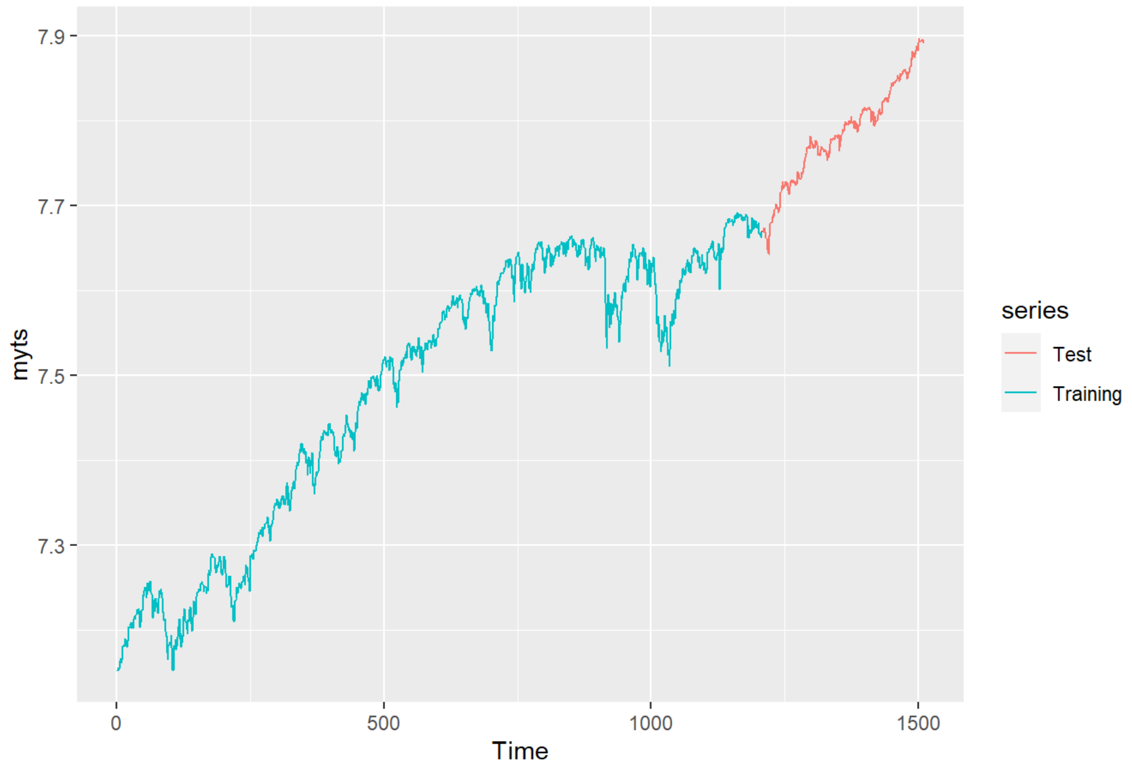

Fig. 5 Test and training sets of S&P500 data for ARIMA model

For the ARIMA model, the auto.arima function is used here. As shown in Figure 6, here is the ARIMA(2,1,1) model, which means that the parameters (p,d,q) of the ARIMA model is (2,1,1) at this time. To be more specific, the potential autoregressive orders (p) is 2, the potential moving average orders (q) is 1, and the one-time differencing operation before means d is 1.

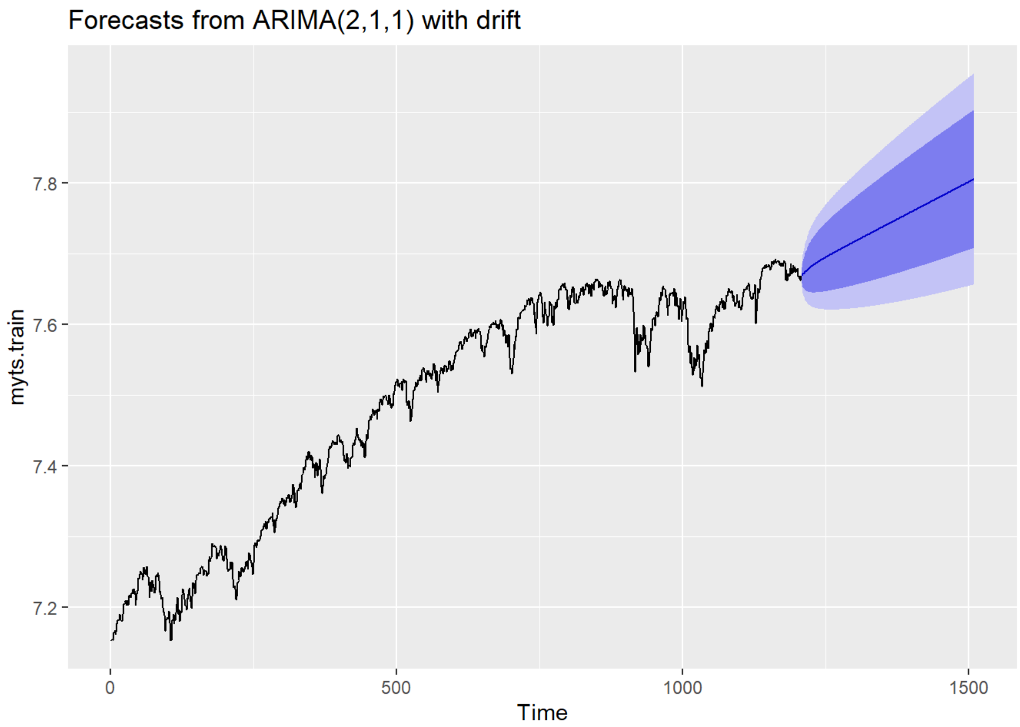

Fig. 6 Forecasts of S&P500 by ARIMA(2,1,1) model

Figure 7 indicates that the forecast of the data by the ARIMA model is good with the test set, because the test set lies within the 80% confidence interval here. So the forecast of S&P500 by the data through ARIMA model accurately covers the expected trends and performs well.

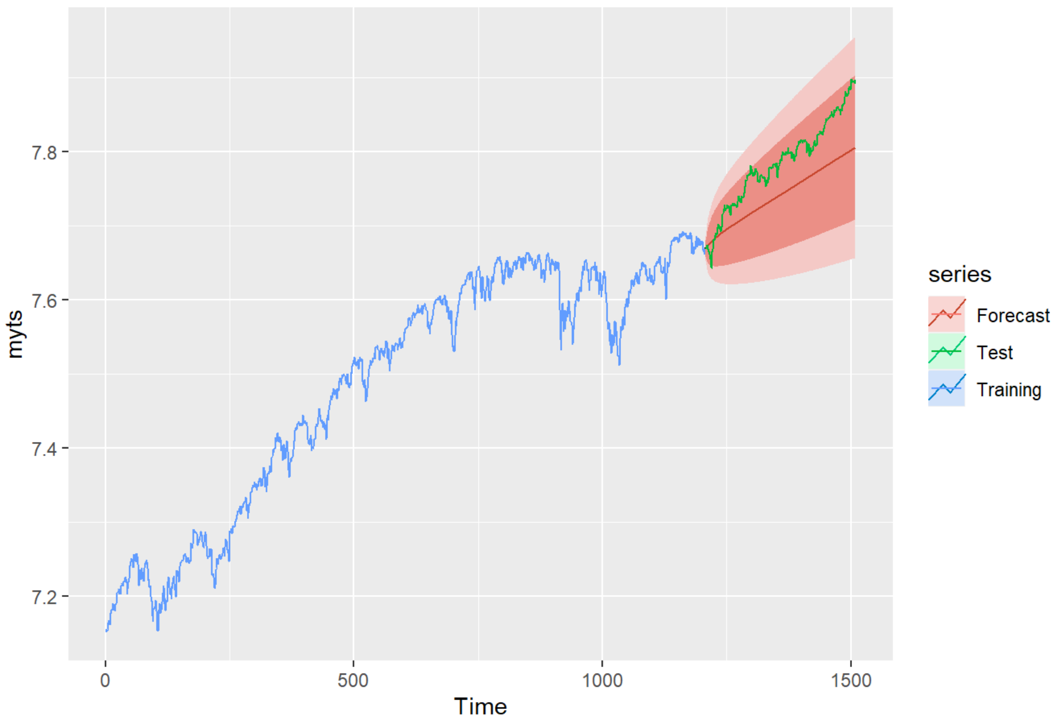

Fig. 7 Results of the forecasts of S&P500 by ARIMA(2,1,1) model

To test the extent and effect of this prediction through ARIMA model, checking residuals can be done here. As shown in Figure 8, although there are some obvious fluctuations in the residuals, it is relatively average and stable on the whole. In ACF, although there is one autocorrelation lag out of the critical line, it is still like a white noise here. And the residuals also show approximately normal pattern. So for such a result, it can be seen as a relatively good prediction.

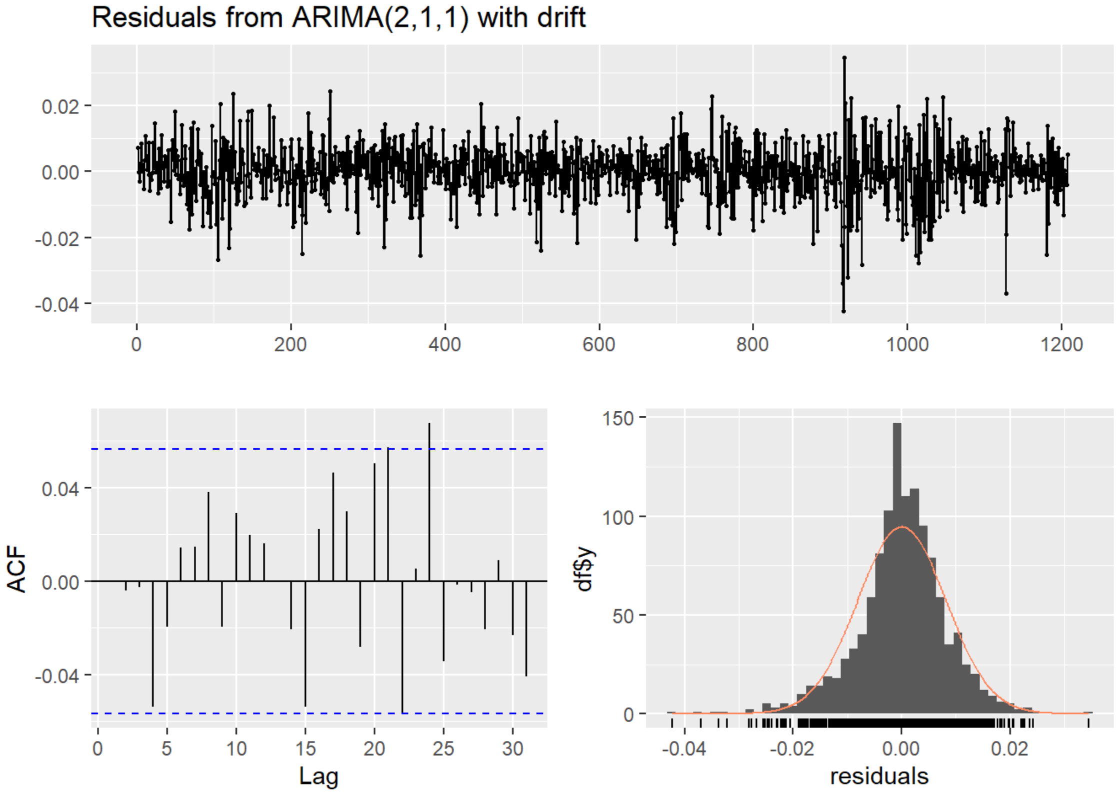

Fig. 8 Residuals of the forecasts of S&P500 by ARIMA(2,1,1) model

Then, checking its accuracy can also be processed, that is, to evaluate it by some specific values. Calculating several important values in evaluation metrics is the main part of checking accuracy here, in particular Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Since they all represent errors, the lower, the better. As shown in Table 1, for RMSE, MAE, and MAPE, the values are all very low, regardless of the training set and the test set, meaning that the model is effectively minimizing the impact of large prediction errors. This shows that the prediction of this ARIMA model also still has certain feasibility and reliability from the analysis of RMSE and other values. Therefore, in summary, the fitting degree of this ARIMA model to the prediction of the future trend of S&P500 is relatively high and appropriate.

Table 1. Evaluation Metrics of the forecasts of S&P500 by ARIMA(2,1,1) model

	ME	RMSE	MAE	MPE
Training Set	2.505617e-05	0.008151126	0.005958462	0.0003575575
Test Set	4.282435e-02	0.048625278	0.044155729	0.5481007908
	MAPE	MASE	ACF1	Theil’s U
Training Set	0.07964047	0.9961046	3.288009e-05	NA
Test Set	0.56549402	7.3817245	9.702540e-01	10.75785

In addition to the ARIMA model, ETS model can also be used to make a time series forecasting. Firstly, like ARIMA, the data are also be divided by 80% and 20% into training and test sets showing in Figure 9.

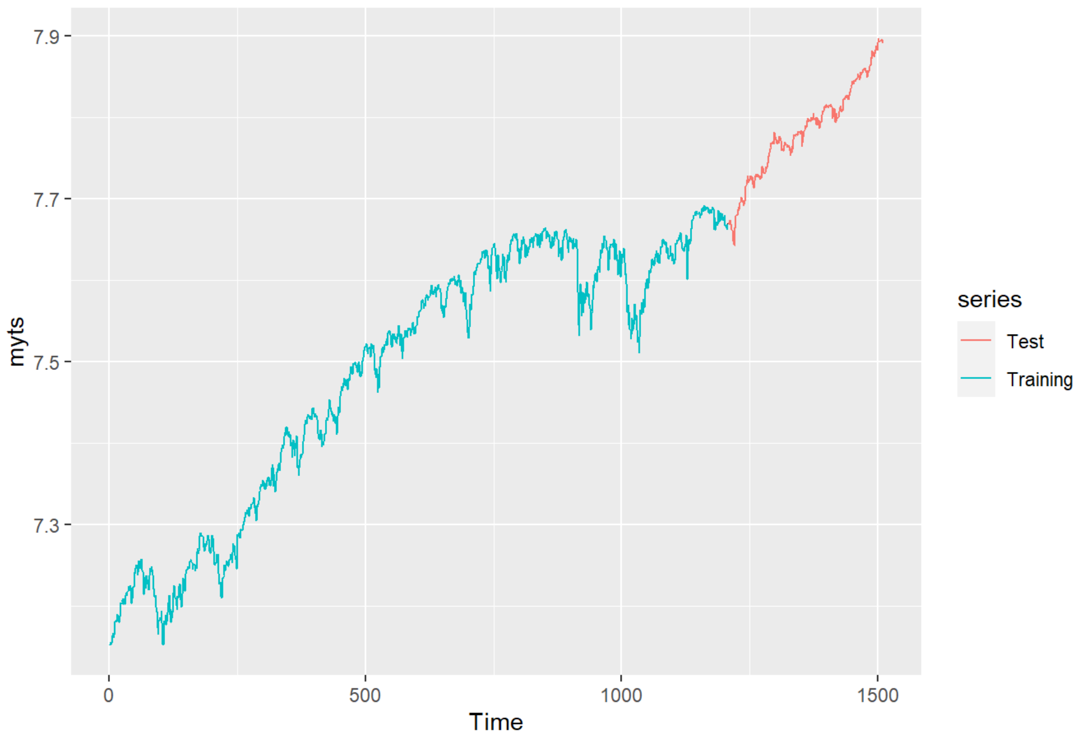

Fig. 9 Test and training sets of S&P500 data for ETS model

Then ETS model is used to analyze and predict the training set. From Figure 10 and Figure 11, the forecast of the data by the ETS model may not so good, because some of the test set data are out of the main range (80% confidence interval) of the prediction, especially the end of the period.

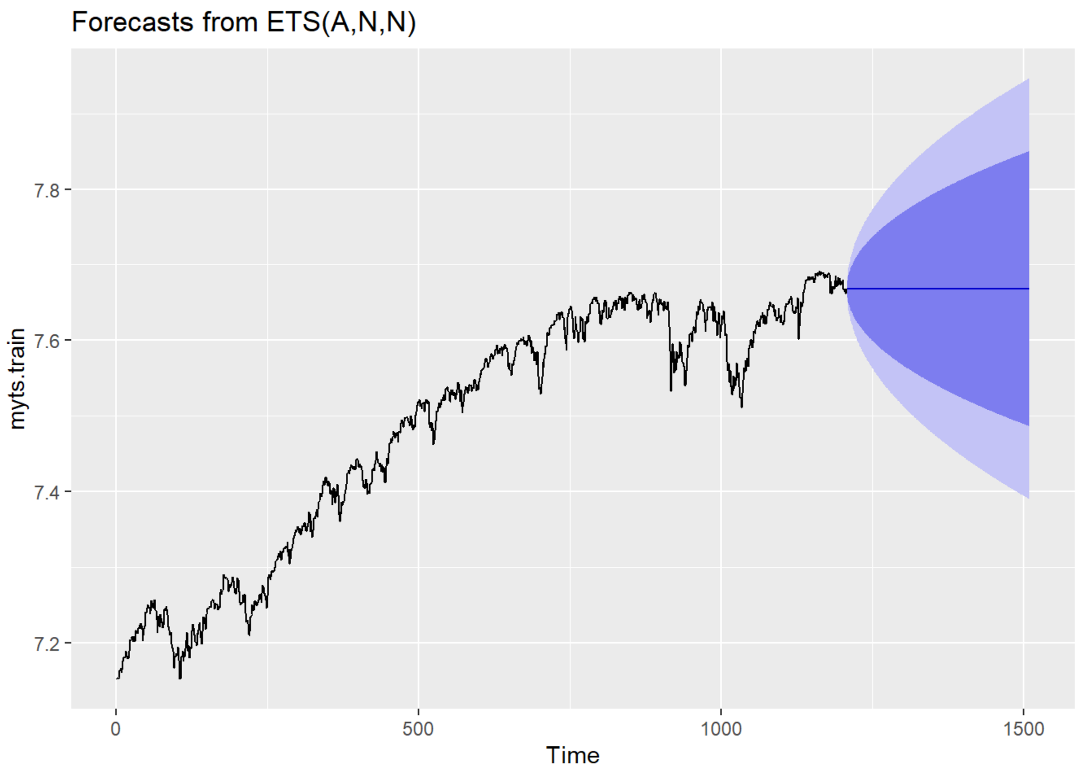

Fig. 10 Forecasts of S&P500 by ETS(A,N,N) model

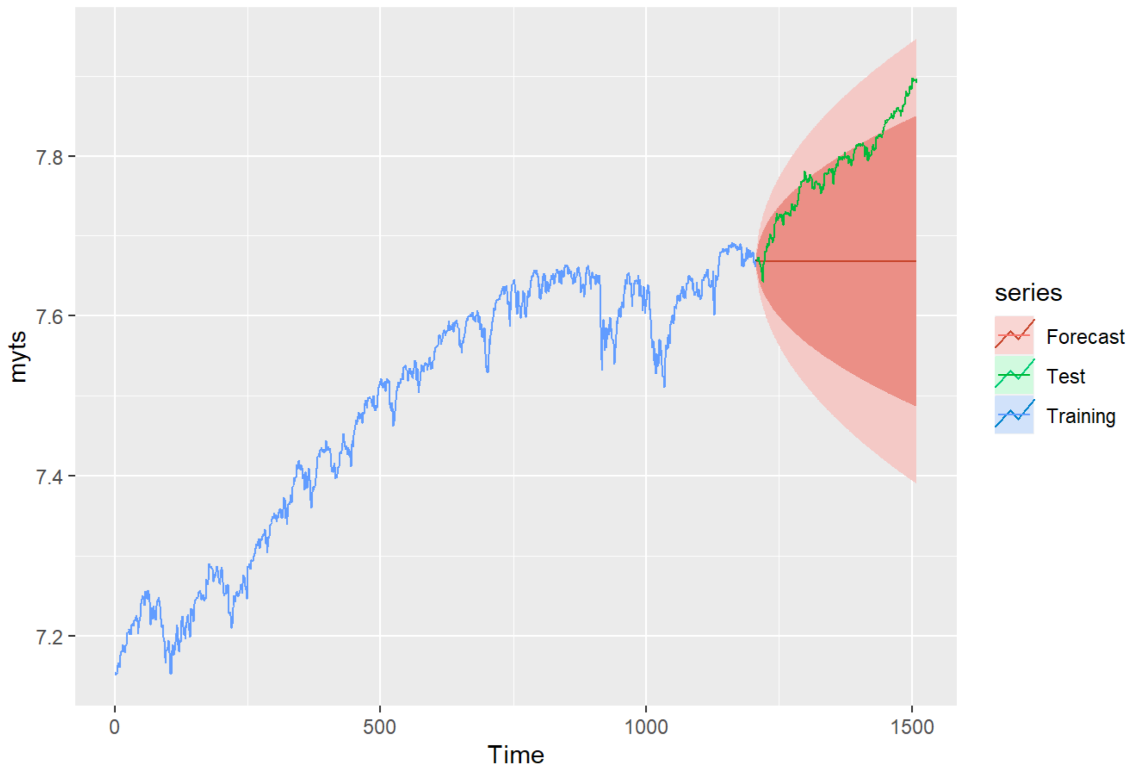

Fig. 11 Results of the forecasts of S&P500 by ETS(A,N,N) model

Similarly, the methods of check residuals and accuracy are also adopted here to demonstrate the effectiveness of this model. In Figure 12, the whole residuals basically show a uniform trend, with several prominent fluctuations. And residuals seem relatively normal. But ACF is not a white noise, since there are several lags out the critical line. So ETS model's forecast of the data is feasible but not so good here.

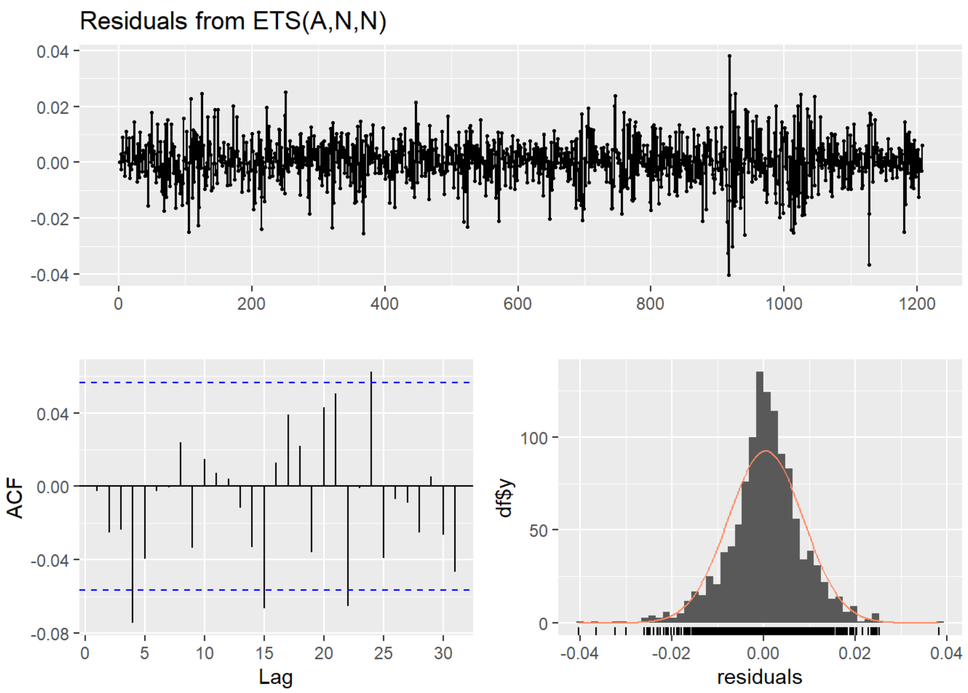

Fig. 12 Residuals of the forecasts of S&P500 by ETS(A,N,N) model

Moreover, in evaluation metrics by Table 2, some of the RMSE, MAE, and MAPE values are low, but some are not low enough. For example, some are lager than 1. So the effectiveness of the ETS model is not high enough here.

Table 2. Evaluation Metrics of the forecasts of S&P500 by ETS(A,N,N) model

	ME	RMSE	MAE	MPE
Training Set	0.0004295247	0.008188356	0.005975395	0.005738502
Test Set	0.1156707316	0.129583717	0.116344950	1.480446406
	MAPE	MASE	ACF1	Theil’s U
Training Set	0.07985916	0.9989354	-0.002532314	NA
Test Set	1.48926008	19.4499422	0.985232541	28.65392

4. Discussion

Although both the ARIMA model and the ETS model can make reasonable and effective predictions on S&P500 data, the comparison between them can also help to find the best model for forecasting S&P500 data. In order to compare the two models, the most efficient way is to compare the evaluation metrics, that is comparing the magnitude of RMSE, MAE, and MAPE between ARIMA and ETS models. Lower values of RMSE, MAE, and MAPE indicate better accuracy of the forecasting. From the two evaluation metrics based on Table 1 and Table 2, all of RMSE, MAE, and MAPE, and all other values of the ARIMA model are all lower than that of the ETS model. And although the differences between some of them are very small and almost identical, if have to choose an optimal one, based on most values, the ARIMA model is probably superior than the ETS model here.

The comparison result is the same as the conclusion by another author about this topic. According to Zhanao Sun’s research paper, the forecasting of S&P500 also has a higher goodness of fit of the ARIMA model, and the main reason is the stationarity of the data, that is, the ARIMA model needs stationary differenced data while the ETS model does not require [10]. So based on this analysis, when comparing the models, different processing requirements for the raw data may lead to different forecasting effects, especially for volatile data such as the stock indices.

5. Conclusion

In the financial market, the stock index is a very representative indicator of the market direction. Analyzing the previous performance of the stock index and predicting the development in the future can provide effective tips and judgments for the whole market to a certain extent. This research focuses on the US stock market, selects and processes S&P500 data during the special period of oil price volatility, and makes time series forecasting for S&P500 through ARIMA model and ETS model respectively. After the prediction results are published, the comparison results of the two models are also obtained. Through the comparison and analysis of some values, ARIMA model is more suitable and more accurate than ETS model for this prediction. This kind of comparison can select the best prediction model, so that some regularities, such as directly choosing the more suitable model to make predictions without making comparisons again, can be found by multiple comparisons, and prepare for the subsequent forecasts. For researchers, investors, and market analysts who are interested in financial markets and future economic developments, this study can help to provide a more specific and deeper understanding of the stock market’s behavior and its future development. However, this study only uses one stock index, the S&P500, as the research object, which cannot cover all situations. And only two models are used for this study, which is just a very small part of all analysis and forecasting methods. Therefore, combined with the research of more scholars with more methods, it is more helpful to analyze the situation and future trend of the entire market, so as to achieve more effective analysis and more accurate forecasts.

References

[1] Raddant M, Kenett D Y. Interconnectedness in the global financial market. Journal of International Money and Finance, 2021, 110: 102280.

[2] Hayo B, Kutan A M. The impact of news, oil prices, and global market developments on Russian financial markets. Economics of Transition, 2005, 13(2): 373-393.

[3] Wójcik D, Ioannou S. COVID-19 and finance: market developments so far and potential impacts on the financial sector and centres. Tijdschrift voor economische en sociale geografie, 2020, 111(3): 387-400.

[4] Sun M, Zhang C. Comprehensive analysis of global stock market reactions to the Russia-Ukraine war. Applied economics letters, 2022: 1-8.

[5] Hamid S A, Iqbal Z. Using neural networks for forecasting volatility of S&P 500 Index futures prices. Journal of Business Research, 2004, 57(10): 1116-1125.

[6] Shang H L. Forecasting intraday S&P 500 index returns: A functional time series approach. Journal of forecasting, 2017, 36(7): 741-755.

[7] Chan E G. Forecasting the S&P 500 index using time series analysis and simulation methods. Massachusetts Institute of Technology, 2009.

[8] Nguyen H, Nguyen H, Pham A. Oil price declines could hurt US financial markets: the role of oil price level. The energy journal, 2020, 41(5).

[9] Yahoo Finance, URL: https://finance.yahoo.com/quote/%5EGSPC?p=%5EGSPC, last accessed 2023/8/7.

[10] Sun Z. Comparison of trend forecast using ARIMA and ETS Models for S&P500 close price. Proceedings of the 2020 4th International Conference on E-Business and Internet. 2020: 57-60.