S&P500 Time Series Analysis

Yeyunfei Jiang (1004170354)

2021/4/10

Abstract:

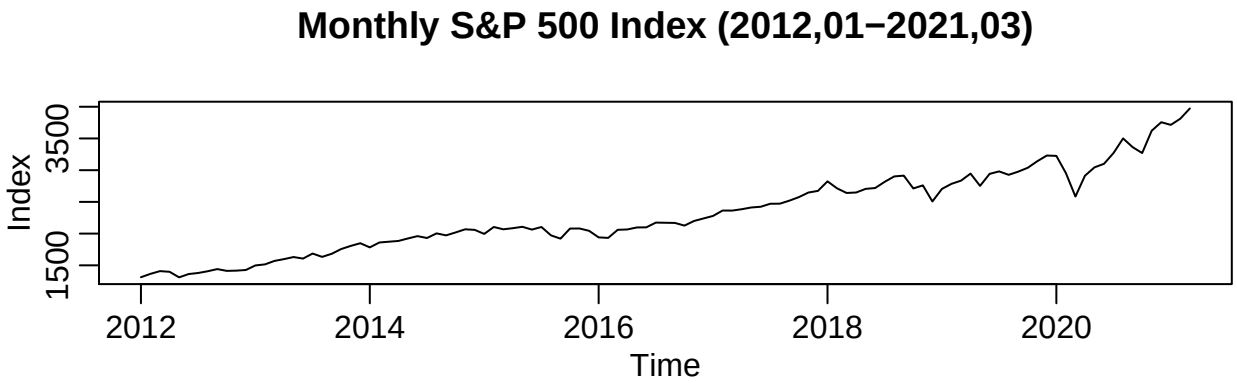
In this report, I perform adequate exploratory analysis to S&P500, models the trend and forecast 10 periods into the future, shows the predicted values and 95% confidence intervals, and draw a conclusion on the trend. Additional to that, the first first three predominant periods were found by periodogram analysis.

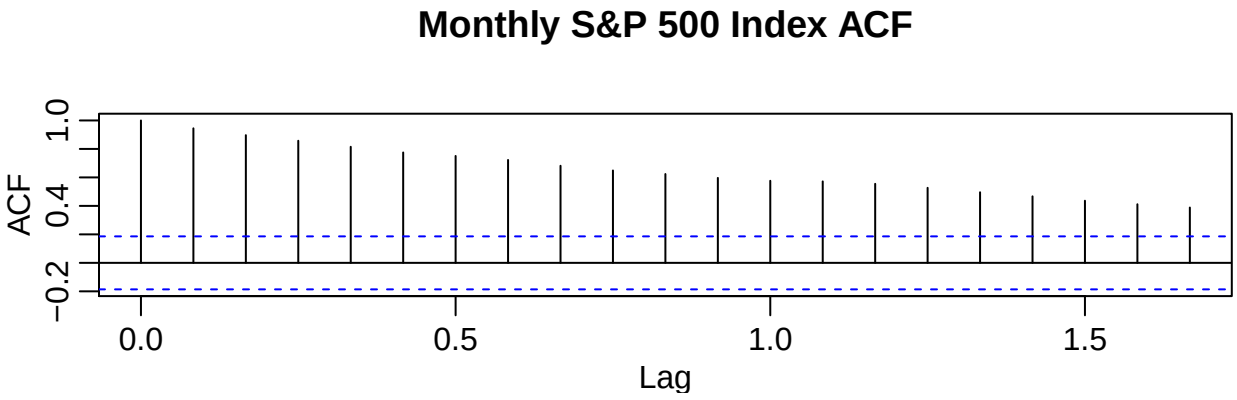
Introduction:

The S&P500 includes 500 large companies listed on stock exchanges in the United States, which people believes is a well diversified portfolio that indices how the stock market moves and reflects economic situation for investors. This report mainly focus on analyzing this time series, constructing appropriate models which captures trend the most, using this model to forecasting 10 future periods, and use spectral analysis to find the first three predominant periods.

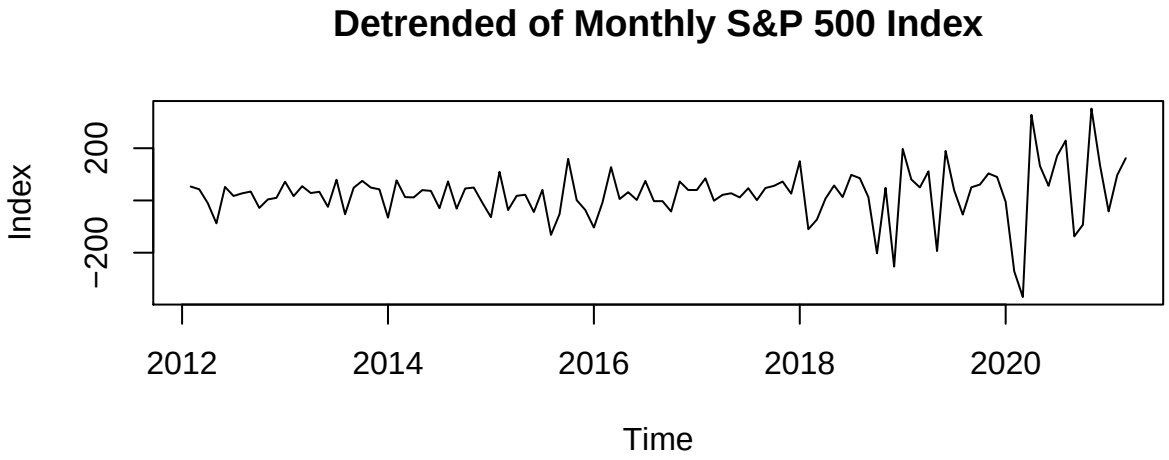
The data set used in this report is from Federal Reserve Economic Data (FRED-https://fred.stlouisfed.org/series/SP500). It contains 111 monthly end of period prices from Jan 2012 to Mar 2021.

Statistical Methods:



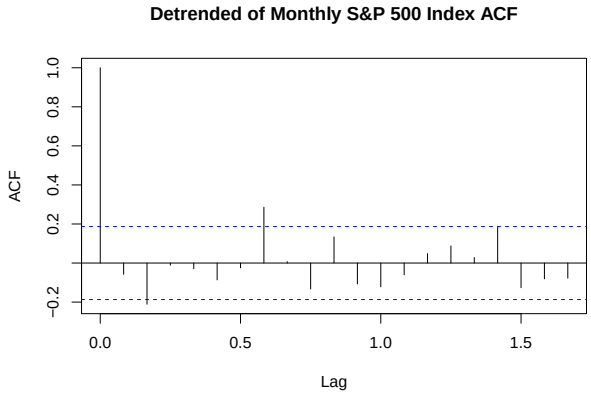


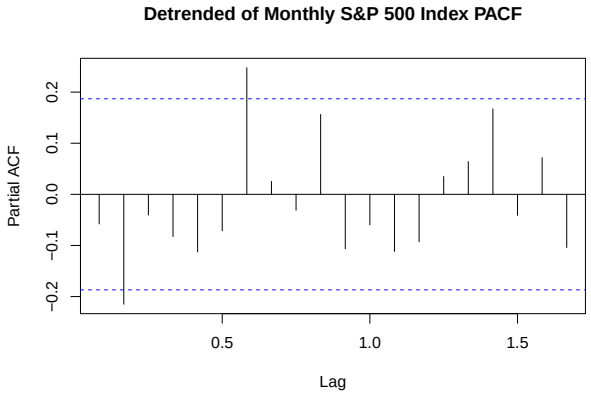
From the ACF of this time series, we an see that the autocorrelation does not decay to zero fast as h increases, it indicates the time series is not stationary. Hence, need to take the first difference to transform this time series to a stationary process.



```
## Augmented Dickey-Fuller Test
## alternative: stationary
##
## Type 1: no drift no trend
##      lag    ADF p.value
## [1,]   0 -10.33    0.01
## [2,]   1  -8.40    0.01
## [3,]   2  -6.15    0.01
## [4,]   3  -5.12    0.01
## [5,]   4  -4.51    0.01
## Type 2: with drift no trend
##      lag    ADF p.value
## [1,]   0 -10.88    0.01
## [2,]   1  -9.21    0.01
## [3,]   2  -7.05    0.01
## [4,]   3  -6.23    0.01
## [5,]   4  -5.85    0.01
## Type 3: with drift and trend
##      lag    ADF p.value
## [1,]   0 -10.96    0.01
## [2,]   1  -9.36    0.01
## [3,]   2  -7.23    0.01
## [4,]   3  -6.41    0.01
## [5,]   4  -6.08    0.01
## ----
## Note: in fact, p.value = 0.01 means p.value <= 0.01
```

The first difference of S&P 500 Index is plotted in the above figure, and using adf test, it is a stationary process under 0.05 level. Next observe ACF and PACF to identify the dependence order of model.





From the graphs, we can see that the PACF is cutting off at lag 7, and ACF is cutting off at lag 2 too. No seasonality was found in the acf/pacf. So, considering two candidate models, ARIMA = (7, 1, 2), and ARIMA(6,1,2).

Result:

Output for ARIMA(7,1,2)

```
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##     Q), period = S), xreg = constant, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ar5      ar6     ar7     ma1
##       -0.6239  -0.9362  -0.2739  -0.3243  -0.2468  -0.2290  0.1047  0.5458
## s.e.   0.1716   0.1885   0.1521   0.1493   0.1505   0.1362  0.1275  0.1476
##          ma2  constant
##       0.7178   22.9127
## s.e.  0.1647    5.5534
##
## sigma^2 estimated as 8084:  log likelihood = -652.05,  aic = 1326.09
##
## $degrees_of_freedom
## [1] 100
##
## $ttable
##          Estimate     SE t.value p.value
## ar1       -0.6239 0.1716 -3.6362  0.0004
## ar2       -0.9362 0.1885 -4.9679  0.0000
## ar3       -0.2739 0.1521 -1.8014  0.0746
## ar4       -0.3243 0.1493 -2.1723  0.0322
## ar5       -0.2468 0.1505 -1.6399  0.1042
## ar6       -0.2290 0.1362 -1.6805  0.0960
## ar7        0.1047 0.1275  0.8212  0.4135
## ma1        0.5458 0.1476  3.6970  0.0004
## ma2        0.7178 0.1647  4.3579  0.0000
## constant  22.9127 5.5534  4.1259  0.0001
##
```

```
## $AIC
## [1] 12.05539
##
## $AICc
## [1] 12.07559
##
## $BIC
## [1] 12.32544
```

Output for ARIMA(6,1,2)

```
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##     Q), period = S), xreg = constant, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ar5      ar6     ma1     ma2
##       -0.6989  -1.0275  -0.3262  -0.3721  -0.3285  -0.3007  0.6030  0.7834
## s.e.   0.1415   0.1397   0.1442   0.1429   0.1185   0.1094  0.1194  0.1194
##       constant
##        22.7799
## s.e.    5.1218
##
## sigma^2 estimated as 8130:  log likelihood = -652.37,  aic = 1324.74
##
## $degrees_of_freedom
## [1] 101
##
## $ttable
##          Estimate     SE t.value p.value
## ar1       -0.6989 0.1415 -4.9379  0.0000
## ar2       -1.0275 0.1397 -7.3571  0.0000
## ar3       -0.3262 0.1442 -2.2625  0.0258
## ar4       -0.3721 0.1429 -2.6036  0.0106
## ar5       -0.3285 0.1185 -2.7728  0.0066
## ar6       -0.3007 0.1094 -2.7482  0.0071
## ma1        0.6030 0.1194  5.0505  0.0000
## ma2        0.7834 0.1194  6.5632  0.0000
## constant  22.7799 5.1218  4.4477  0.0000
##
```

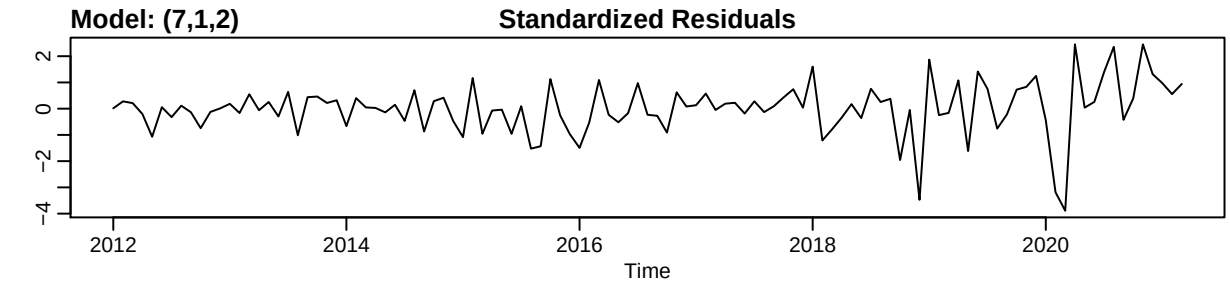
```
## $AIC
## [1] 12.04313
##
## $AICc
## [1] 12.05949
##
## $BIC
## [1] 12.28863
```

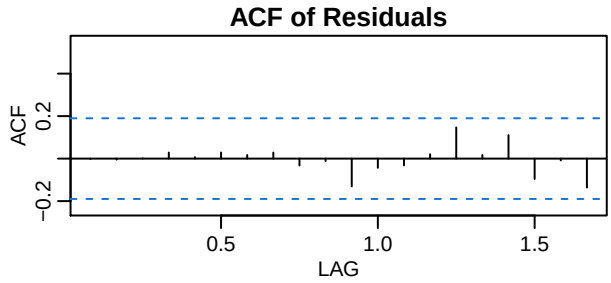
We can see from the above output for ARIMA(7, 1, 2)model, the p-values for all estimates are smaller than$\alpha = 0.05$, except ar5, ar6, ar7. Showing other than ar5, ar6, ar7, all parameters are statistically significant.

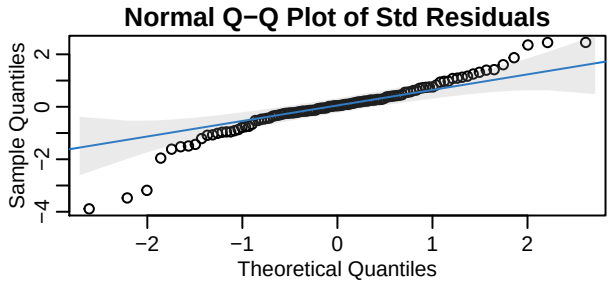
As well as the output for ARIMA(6, 1, 2)model, the p-values for all estimates are smaller than$\alpha = 0.05$ with no exception. Showing that all parameters are statistically significant.

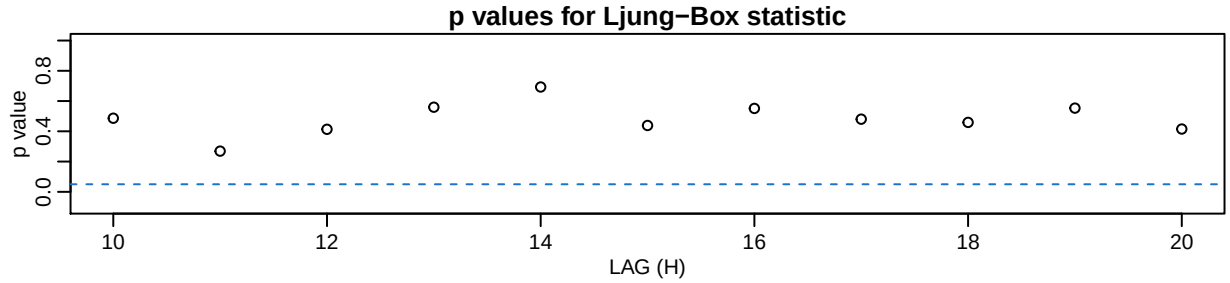
The parameters for ar indicate how would today's first difference of S&P500 move when the value of ith period ago increased by one, keeping other variables constant. In the ARIMA(6,1,2) model, all ar parameters are negative. A negative parameter for ar indicate a increase in the value ith period ago would decrease today's value.

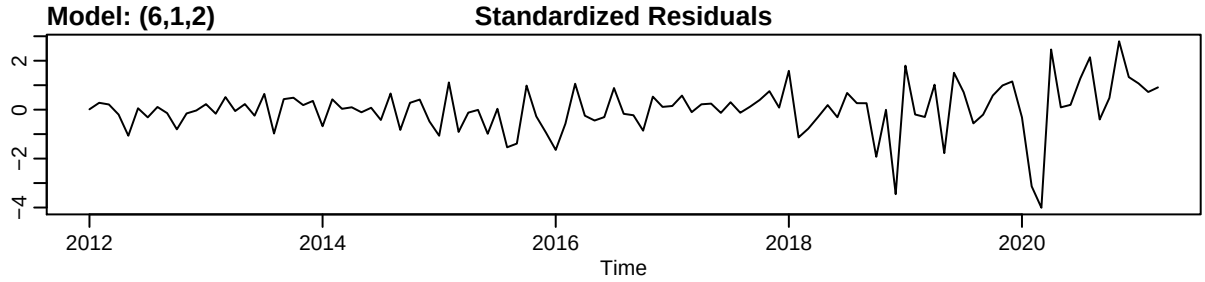
The parameters for ma indicate how would today's first difference of S&P500 move when the moving average of ith period ago increased by one, keeping other variables constant. In the ARIMA(6,1,2) model, all ma parameters are positive. A positive parameter for ma indicate a increase in the moving average ith period ago would increase today's value.

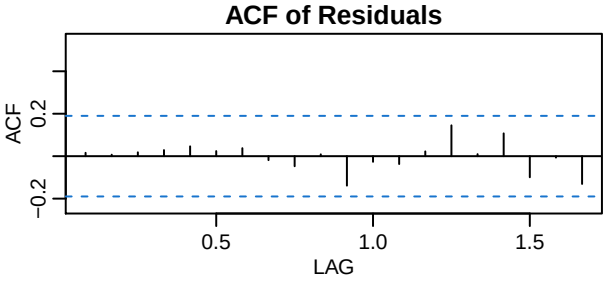

Model: (7,1,2)
Standardized Residuals
2
0
−2
−4
2012
2014
2016
2018
2020
Time

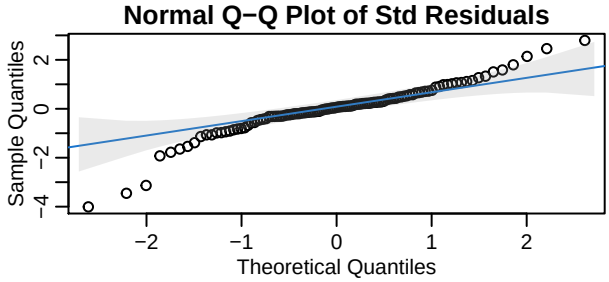

ACF of Residuals
ACF
0.2
−0.2
0.5
1.0
1.5
LAG

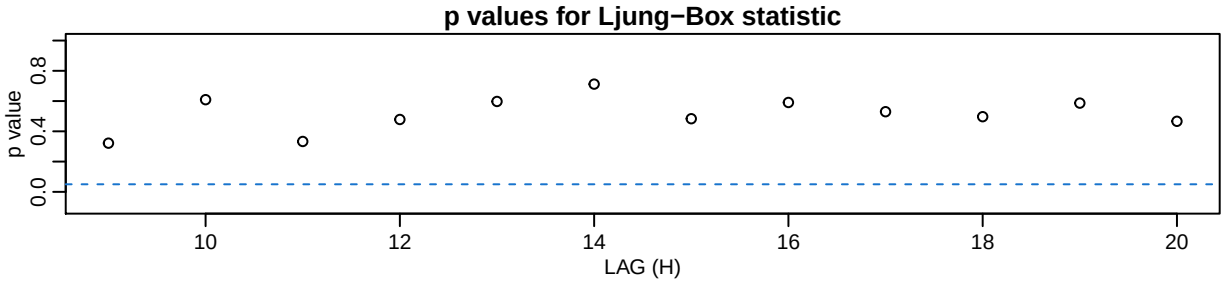

Normal Q−Q Plot of Std Residuals
Sample Quantiles
2
0
−2
−4
−2
−1
0
1
2
Theoretical Quantiles


p values for Ljung−Box statistic
p value
0.0
0.4
0.8
10
12
14
16
18
20
LAG (H)


Model: (6,1,2)
Standardized Residuals
2
0
−2
−4
2012
2014
2016
2018
2020
Time


ACF of Residuals
ACF
0.2
−0.2
0.5
1.0
1.5
LAG


Normal Q−Q Plot of Std Residuals
Sample Quantiles
2
0
−2
−4
−2
−1
0
1
2
Theoretical Quantiles


p values for Ljung−Box statistic
p value
0.0
0.4
0.8
10
12
14
16
18
20
LAG (H)

The time plot of the standardized residuals shows no obvious patterns, with few outliers exceeding 3 standard deviations in magnitude. The ACF of the standardized residuals shows all value lies in the confidence bands, suggests no apparent departure from randomness for both models. The normal Q-Q plot of the residuals shows the data fits the line pretty well and is reasonable follows normal distribution, with the some possible outliers at the tails show a deviation from normality. The Q-statistic is not significant, and shows we can not reject the data are independently distributed.

Overall,the ARIMA(7, 1, 2) and ARIMA(6, 1, 2) models' residuals seem iid and normal with mean zero and constant variance, which suggest the models fit well. Based on the AIC, AICc, and BIC of the two models, ARIMA(6,1,2) model has a smaller value for all of them, hence, we choose ARIMA(6,1,2).

Forecasting

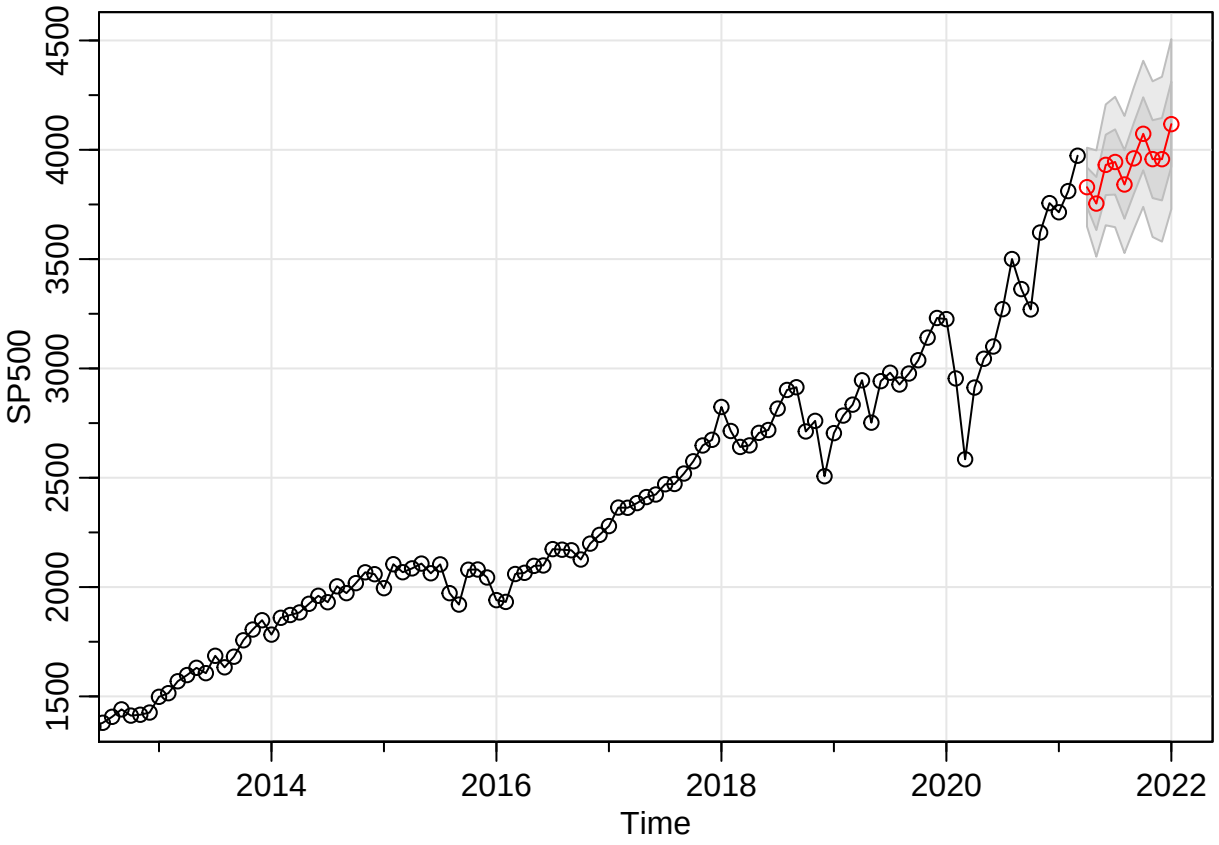
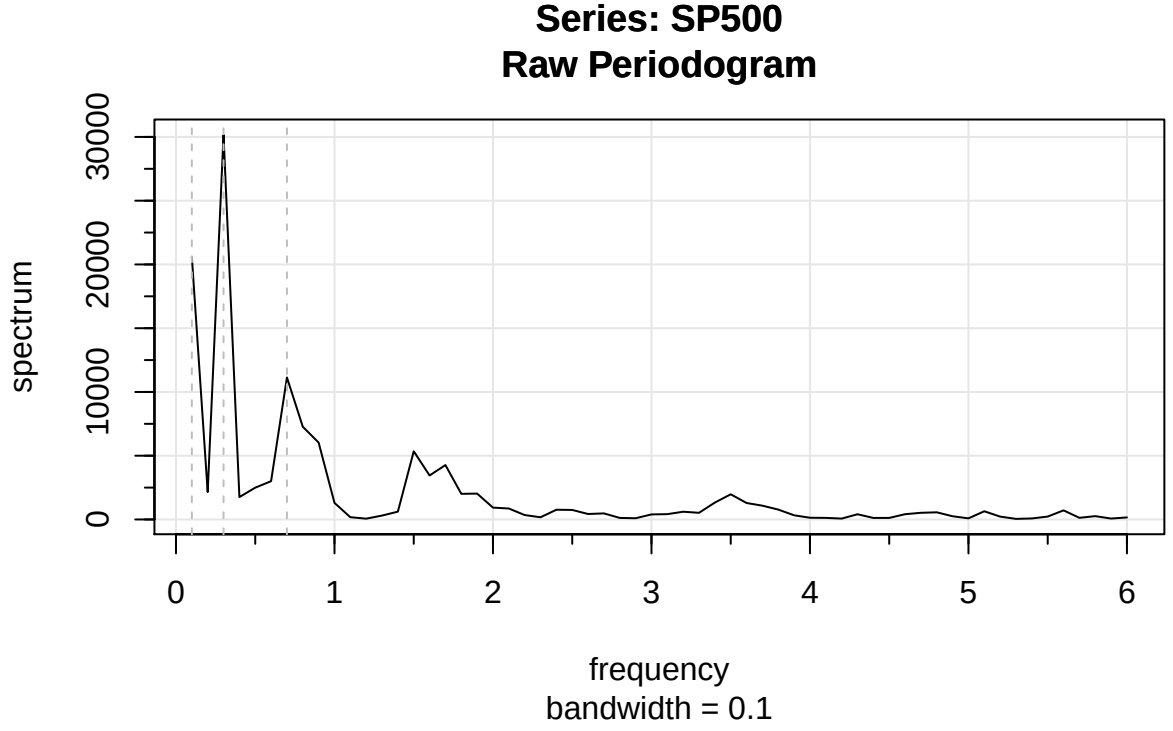


Table 1: Forecasting value and 95% CI

Period	Prediction	PI.95..Lower.Bound	PI.95..Upper.Bound
1	3829.156	3652.431	4005.880
2	3754.143	3515.897	3992.389
3	3931.095	3660.406	4201.785
4	3944.281	3652.048	4236.515
5	3841.291	3534.044	4148.539
6	3960.846	3642.619	4279.073
7	4073.157	3745.748	4400.567
8	3957.288	3608.369	4306.207
9	3957.014	3587.645	4326.382
10	4117.342	3736.473	4498.212

The forecasting values for the S&P500 shows a positive trend, but is fluctuating around. The width of the 95% confidence interval is around 400 for the first period and getting wider over time. We can not conclude the value would increase in 10 periods according to the 95% confidence interval, since the lower bound of period 10 is less than the value of last period.

Spectral Analysis



Order	Dominant.Freq	Period	Spec	Lower.Bond	Upper.Bond
1	0.3	3.3333	30161.51	8176.335	1191316.2
2	0.1	10.0000	20582.14	5579.510	812951.1
3	0.7	1.4286	11135.98	3018.797	439847.7

The 95% confidence interval for the first three dominant frequencies of S&P500 series seem extremely wide. Hence, we cannot establish the significance of the peak.

Discussion:

The study modeled the trend of S&P500 from Jan 2012 to Mar 2020 using time series analysis. Two models were proposed, it appears that both ARIMA(7,1,2) and ARIMA(6,1,2) model fit the data well and satisfied all the model's assumption, based on model selection criteria, ARIMA(6,1,2) was selected.

As a result, this model predicts for the future 10 periods, S&P500 would have a positive trend that seems to lead the index upwards, but by considering the 95% confidence interval, it is not promised to be increase in value in the next 10 periods.

The study also uses periodogram analysis to identify the first three predominant periods, as it shown in the periodogram, the three predominant periods are (3.33, 10, 1.4286). However, the 95% confidence interval are too wide to draw the conclusion of the significance of the peak.

Yet, there are still some limitations of this study. The model only count the stochastic trend, but not the deterministic trend. Some outliers are not fixed, for example, the period of pandemic has cause S&P500 a sudden drop around 34%. These unexpected events can not be predicted, however they can change the trend by a lot.