The final report should comprise the following sections:

1. *Introduction:* The data collected should be described in detail, with specifics on what was recorded and how, along with the motivation behind the analysis of the time series.

2. *Analysis:* This section should start with a suitable display of the time series observed, and a summary of any key features. Methodology should then be applied to the data to address the question(s) of interest motivating the study.

3. *Conclusion:* A short synopsis of the findings from the analysis should be presented, along with any other pertinent comments of interest.

4. *Appendix:* The raw data should appear in an appendix, together with the R commands used in the analysis.

S&P 500 Index Analysis and Prediction

(Jan 4th, 2016 ~ March 31th, 2016)

- Introduction
- Analysis (ARIMA, Holt Winter)
- Prediction
- Conclusion
- Appendix

Section 1: Introduction

Introduction

The S&P 500 is an American stock market index composed of the leading 500 companies trading on the NYSE and the NASDAQ. It is different from other major U.S. stock market indices as it uses float-weighted model, which gives companies that have the most public shares a larger impact on the price of the index. For this reason it is considered to be a great representation of the U.S. stock market. Fitting a time series model to the S&P 500 would be of benefit, as it allows one to produce quantitative values such as trends, averages, and even forecasted values for the index, and ultimately allows quantitative inferences to be made about the American economy, in an industry where qualitative opinions dominate. Also, If one could create a time series model that accurately follows the S&P 500, they could use it to make predictions for future prices and ultimately profit financially.

The goal of this report is to decide upon a suitable time series model for the closing prices of the S&P 500 and, once determined, use it to make predictions for future closing prices. To determine the best model, multiple models will be created using various methods taught in STAT 443, then their statistical properties will be compared amongst each other to decide which most accurately captures characteristics of the index. The data will consist of 61 data points, which are the closing prices of the S&P 500 from Jan 04 2016 to Mar 31 2016 (excluding holidays, and weekends where market is closed). Once the appropriate model has been selected, the predicted values will be compared to the actual closing values, to determine if the final model is a good fit for the index.

The “S&P 500” is an American stock index based on the market capitalizations of 500 large companies in the world. We collected the index data from December 17th, 2015 to March 1st, 2016.

We are interested in building suitable models using the dataset we have and the knowledge we have learned so far from STAT 443 Time Series course. We would like to build several models using different statistical methods and **making prediction of the indexes during week after**

March 1st. Eventually, we would also like to **compare the prediction models using the prediction results.**

Section 2: Analysis

I. Summary and Key Features of the Data

From the original dataset, we can see that a few data points are NA, because we have some missing data points. After checking the dataset carefully, we could not determine a pattern in the missing points. They occurred on Jan. 25th, Feb. 1st, and Mar. 25th. Since these points are few in number, we imputed the missing points by taking an average of the previous and following day's values. The plot looks as follows after removing missing data, where time measures number of observations from Jan. 4st:

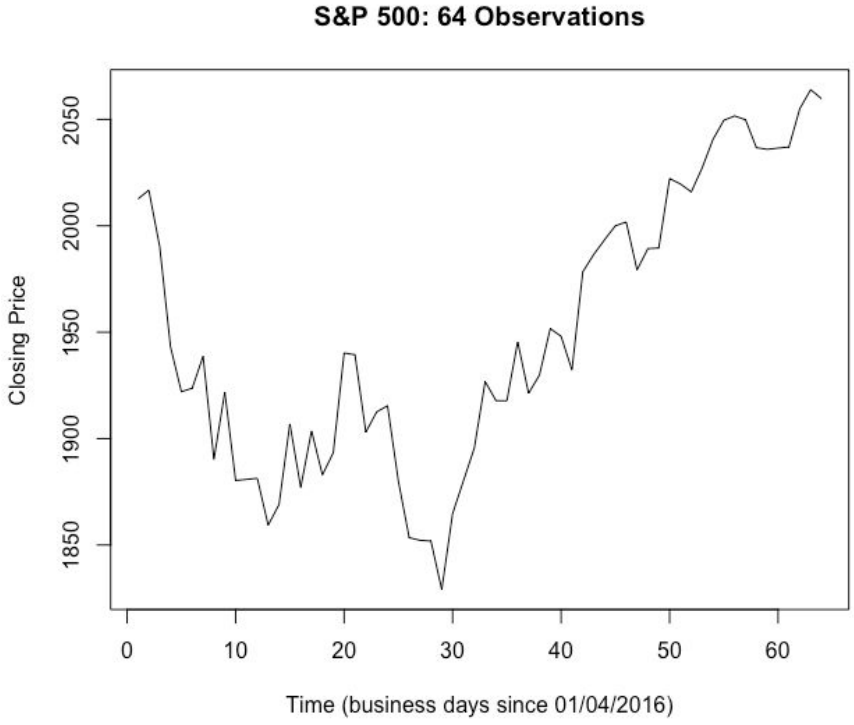

Fig.1: Initial time series plot.

II. Model Fitting

II.I First Model Fitting: ARIMA Model

In the following discussion and plots, we refer to the S&P 500 time series as X(t). First, we note that there appears to be no periodic component to the time series over the range of values observed. Although a peak of some longer-scale periodicity in the time series may occur around time 20 (corresponding to Jan. 21st), too few periods are captured in the 60 observations to

warrant a frequency-domain analysis of the data. Consideration of the raw periodogram supports this hypothesis:

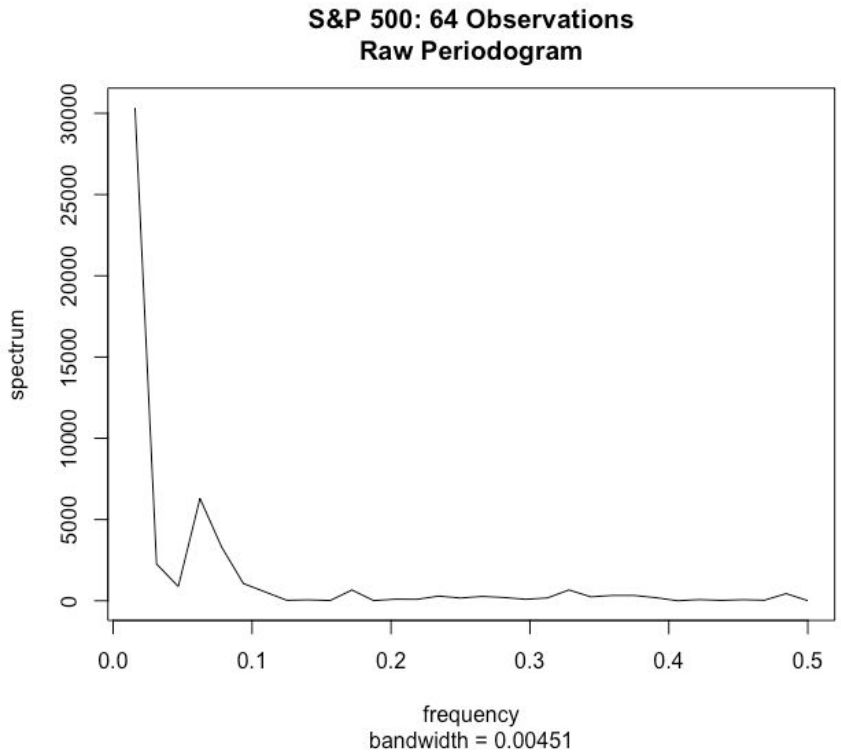

Fig. 2: Raw periodogram for time series.

The peak occurs at the first frequency, 0.015625, which corresponds to wavelength of 64 days. This is the same length as the entire time series, suggesting that the majority of the variance occurs over the entire span of the days rather than at particular wavelengths. A secondary peak occurs at frequency 0.0625, corresponding to every 16’th day, but the variance accounted for by this frequency is not as large as the 64-day cycle. We conclude that a restriction to time-domain analysis is appropriate for this dataset.

From Fig. 1, we can see that the original time series plot has a trend, so we should take the first difference to remove the trend. After taking the first difference, the plot looks as follows:

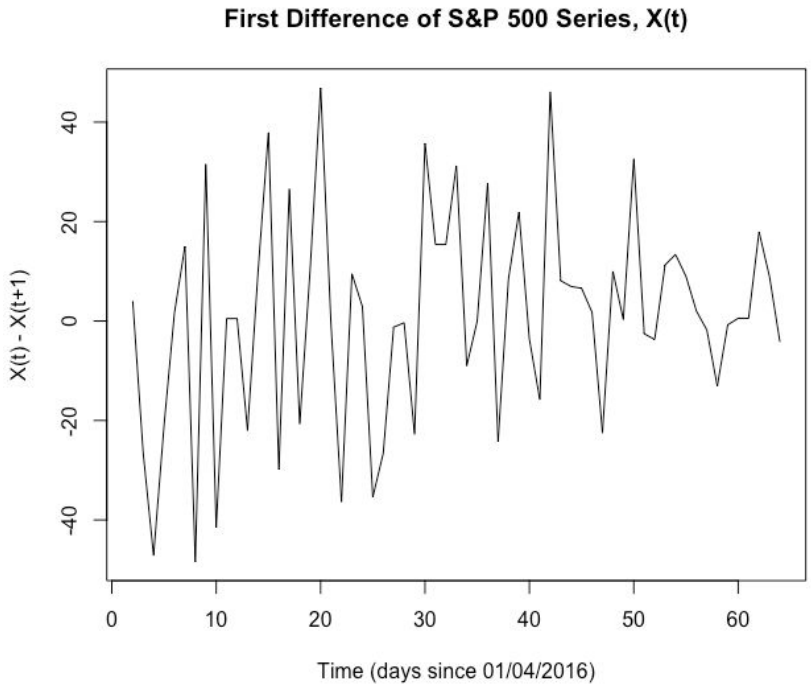

Fig. 3: First difference of X(t).

We can see that the plot still has a trend through the first 30 observations. We decided to take a first difference again. Y(t) is the name of the resulting time series, with the following plot:

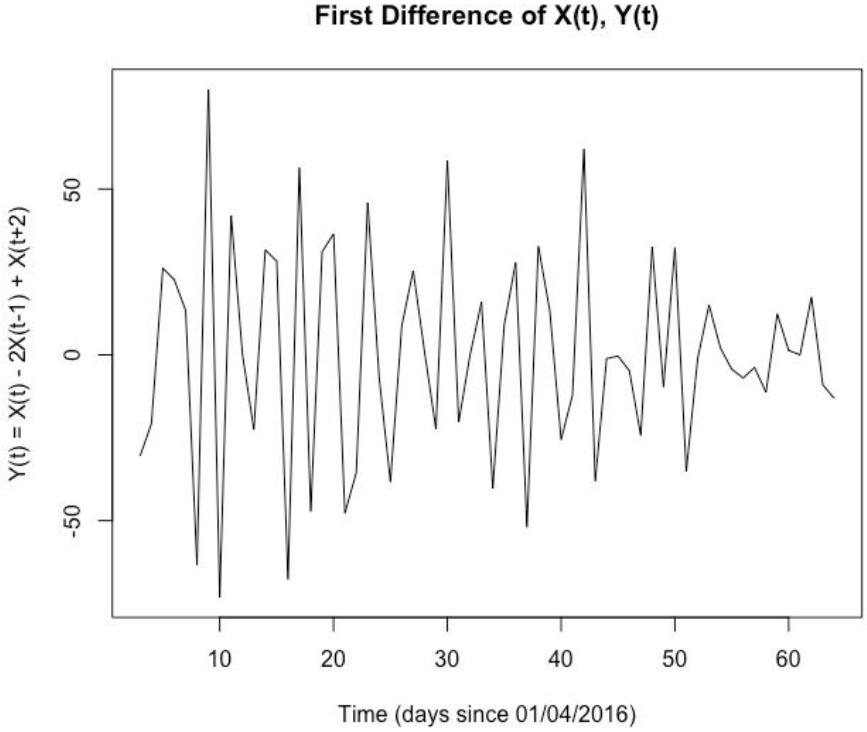

Fig. 4: Y(t), the first difference of X(t).

Note that the plot looks stationary now after taking the first difference of the first differences of the original time series. The ACF and PACF plots for Y(t) are:

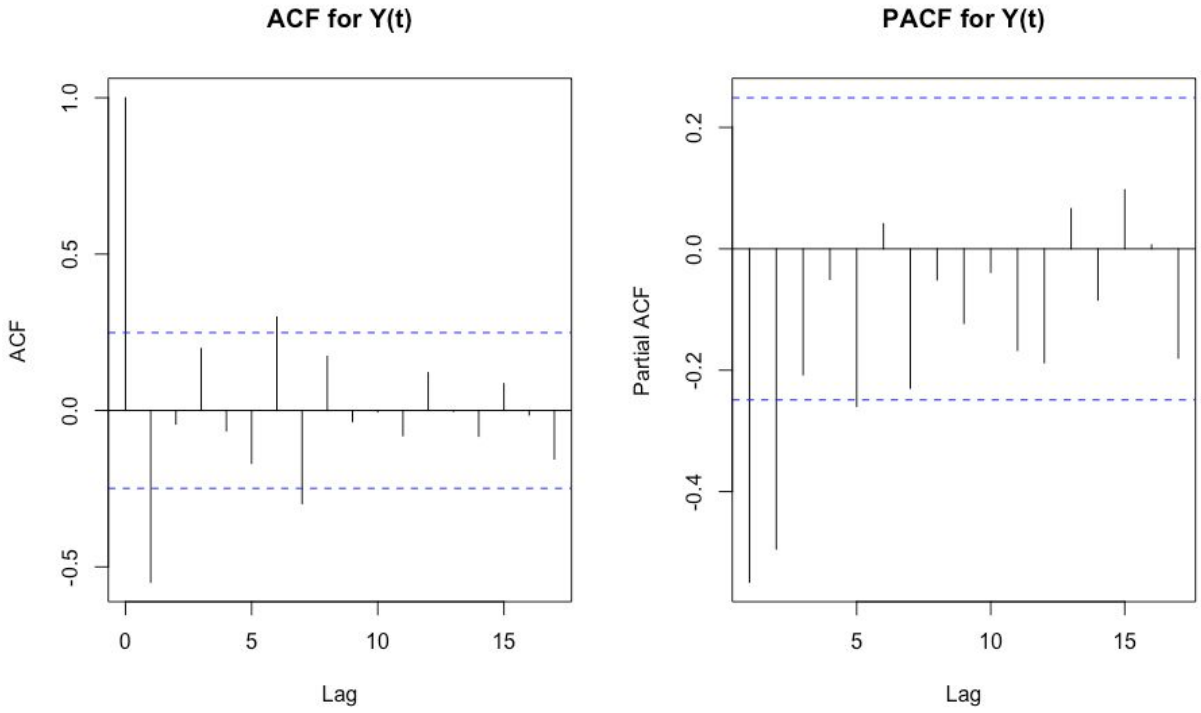

Fig. 5: PACF and ACF of Y(t).

The ACF of Y(t) shows a very significant lag at 1, and there is some evidence of alternation from lags 6 and 7. An AR component with a negative alpha coefficient appears to be appropriate. So, based on all the information we've collected, we can get the ARIMA(2,2,1)x(0,0,0)_0
p=2 is because in the PACF plot, the values is significant at lag 2
d=2 is because we took two differences in trend.
q=1 because in ACF plot, the value is significant in lag 1.

Since there is no seasonal effect, the seasonal component are all 0.
ARIMA model validation:

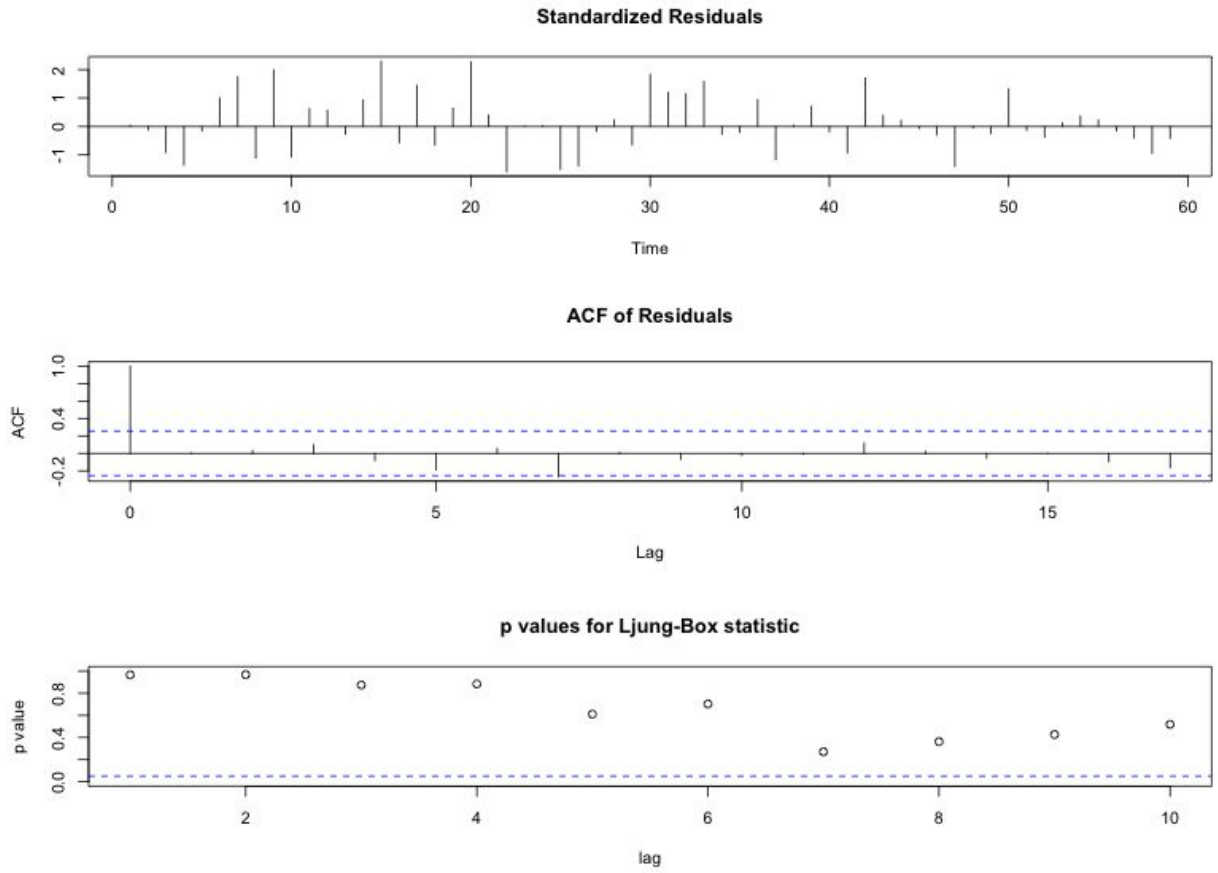

Fig. 6: Diagnostic Plot of ARIMA(2,2,0) x (0,0,0) for X(t)

From all of the three plots above, the Standardized Residuals plot and ACF of Residuals both show random pattern. P values for Ljung-Box statistics are all above the line. It concludes that ARIMA (2,2,1)x(0,0,0)_0 is a suitable model for the dataset.

II.II Second Model Fitting: ARIMA Model for Data after Taking a Forward Ratio, Applying a Log Function, then Adding 100

Next, we try a common and very popular statistic approach to analyse data in stock market. We make a transform our X(t), take a forward ratio, apply a log function, and then add 100 to the resulting value.

Applying a forward ratio enable us to test if there is an associations between consecutive days, and taking a log function transforms the ratio data to a full range range from negative infinity to positive infinity. W(t) is the name of the resulting time series, with the following plot:

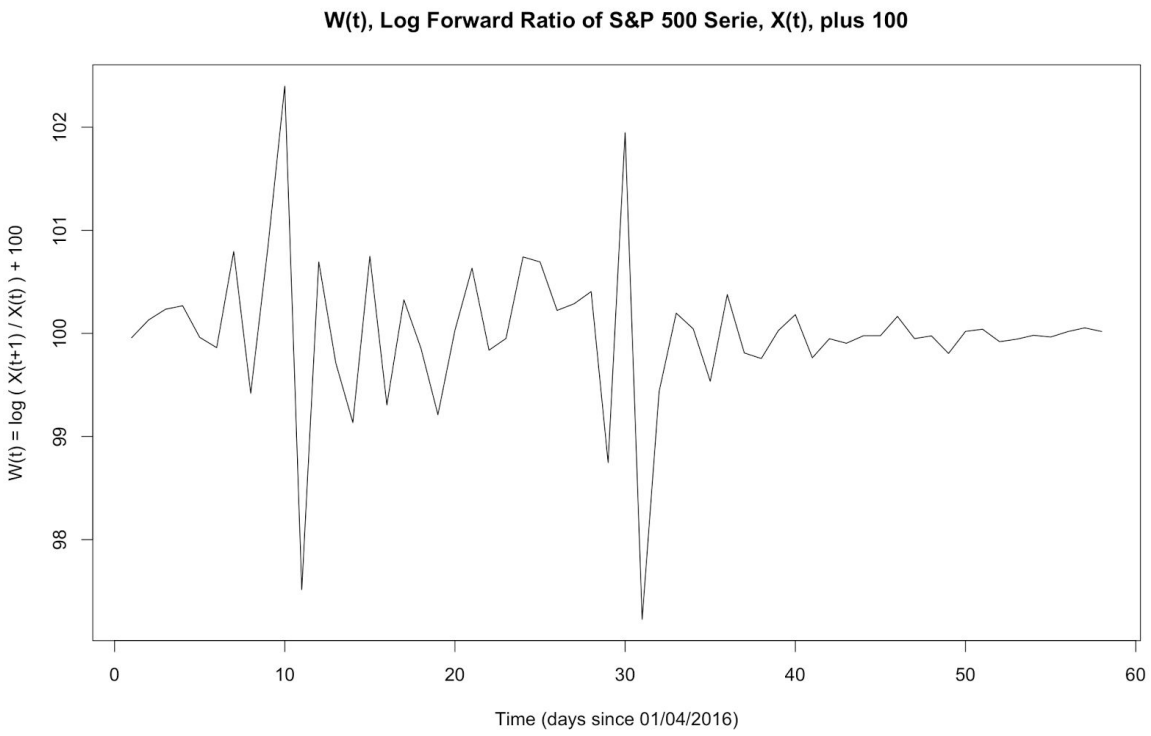

Fig. 7: W(t), Log Forward Ratio of S&P 500 Series of X(t).

It seems that the above transform removes the trend of the original time series, X(t), and smoothes X(t), since the new time series seems less variate.

The ACF and PACF plot of W(t) are:

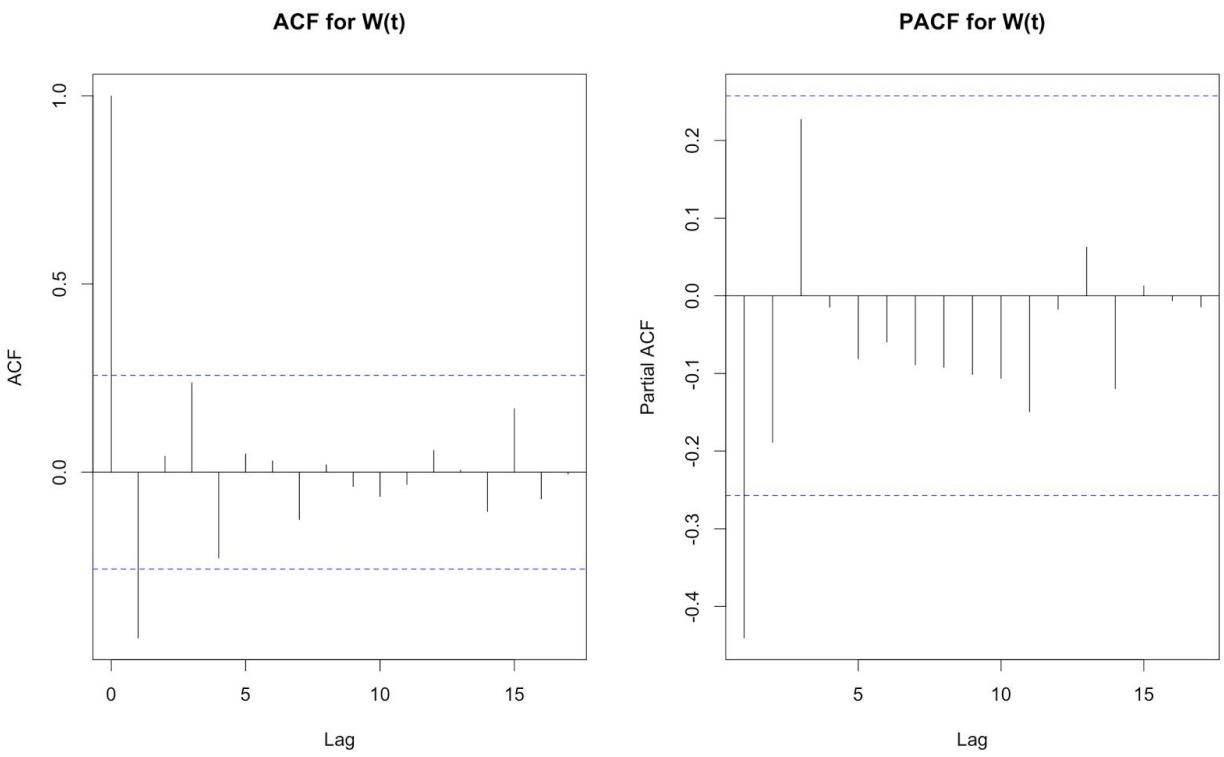

Fig. 8: ACF and PACF of W(t).

The ACF has clear cut off at lag 1 and the PACF plot has a sharp cut off at lag 1. Based on this information, we try to fit W(t) with a ARIMA(1,0,1)x(0,0,0) model

p=1 because there is clear cut off at lag 1 on the pacf plot
d=0 because it seems that the new series do not have any trend
q=1 because the acf seems to have a sharp cut off at lag 1

ARIMA(1,0,1)x(0,0,0) model with estimated parameters:

W(t) - 99.9977 = - 0.2772 * (W(t-1) - 99.9977) + Z(t) - 0.2067 * Z(t-1)

, where Z(t) denotes a white noise process with estimated variance 0.4453, and distribution Z(t) ~ N (0, 0.4453)

ARIMA model validation:

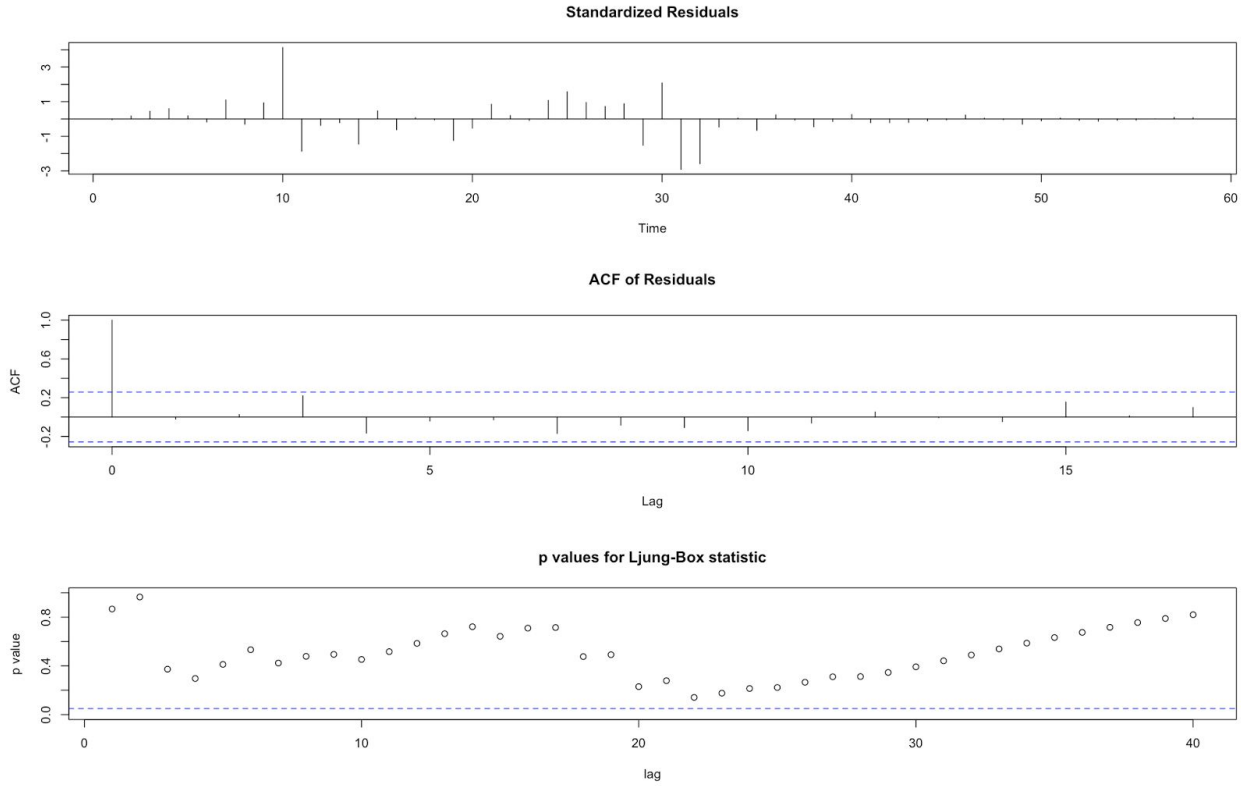

Fig. 6: Diagnostic Plot of ARIMA(1,0,1) x (0,0,0) for W(t)

From the above diagnostic plots, we could see that the model we fitted seems appropriate for W(t). The Standardized Residuals plot does not show any pattern. The ACF of Residuals seems to have a clear cut off at lag 0. P values for Ljung-Box statistics are all above the line. It concludes that ARIMA (1,0,1)x(0,0,0) is a suitable model for W(t).

II.III Third Model Fitting: Holt-Winters Double-Exponential Smoothing

Geoff's TODO: want to talk to TA first about HW and models, see if there's a better option!

II.IV Model Comparison in Model Fitting

With two very different approaches, we obtained two optimized models based on our analysis on the ts plots, acf plots, and pacf plots.

While all the diagnostic plots seem to suggest that these two models are both appropriate, we try to compare the how the two models are fitted with our data in two different approaches: using their AIC values or MSE of the models.

II.IV.I Comparison Based on AIC

The AIC value for the first model, ARIMA(2,2,1)x(0,0,0) for X(t), is 524.6198, and the AIC value for the second model, ARIMA (1,0,1)x(0,0,0) for W(t), is 125.9138.

According to the AIC criteria that the smaller AIC value, the better the model fit, the second model seems to be a better fitted with our data.

II.IV.II Comparison Based on MSE

The MSE value for the first model, ARIMA(2,2,1)x(0,0,0) for X(t), is 2912.369, and the MSE value for the second model, ARIMA (1,0,1)x(0,0,0) for W(t), is 25.82928.

According to the MSE criteria that the smaller MSE value, the better the model fit, the second model seems to be a better fitted with our data, as well.

III. Prediction Using Our Fitted Models

III.I Prediction Values Comparison

	ARIMA(2,2,1)x(0,0,0) for X(t)	ARIMA (1,0,1)x(0,0,0) for W(t)	HoltWinters	Actual result
March 28th, 2016	2041.391	2074.350		2037.05
March 29th, 2016	2043.607	2079.631		2055.01

March 30th, 2016	2045.816	2084.252		2063.95
March 31th, 2016	2048.459	2089.070		2059.74
April 1th, 2016	2050.993	2093.847		2072.78
MSE at $l = 5$	275.1147	742.7455		

III.II Model Comparison in Model Prediction

Unlike what we conclude from the comparison in model fitting, it seems that there is a opposite result when we compare the two models in their abilities to predict.

Base on the MSE value of the next five prediction, the first model, ARIMA(2,2,1)x(0,0,0) for X(t) seems to do a better job in prediction, compared to the second model, ARIMA (1,0,1)x(0,0,0) for W(t).

Section 3: Conclusion

After analysis, the ARIMA (2,2,1)x(0,0,0) and the ARIMA (1,0,1)X(0,0,0) models created are somewhat suitable models for the S&P 500 index over the given 61 business days. When comparing their diagnostics, the models produced acceptable results indicating both were appropriate fits for the data as mentioned above. One could argue that the latter of the two models was a better fit for the data as it resulted in smaller AIC and MSE values than the first model. However one could also argue the first model was a better model as it produced predicted values that were closer to the actual values than the first model for the first five future values. It seems as though choosing the better mode would be up to the reader as neither seems to be a significantly better fit than the other for the given data.

In conclusion, one could say that although both models aren't perfect, they seem to capture the general characteristics of the data. What could be taken away from this report is that creating a model that more accurately fits the S&P 500 over 61 days is a

complicated task where a level of knowledge for time series beyond that of STAT 443 is perhaps needed.

When comparing the predicted values of both models to the actual closing prices for the first five time points in the future, both models fared reasonably. The ARIMA(2,2,1)x(0,0,0) model did a great job at capturing the rising prices of the index, where one would observe the predicted values rose with the actual closing prices over the five day interval. Unfortunately, it appears as though the first model doesn't capture the volatility of the index. Although the predicted and actual prices showed an upward trend over the five future data points, the values generated from the model increase in small, steady increments relative to the actual price which had much larger swings in both the negative and positive direction.

The second model also provides somewhat acceptable predicted values. Like the first model, it captures the upward trend of the five future values, and once more, seems to capture the volatility unlike the first model. However this extra volatility in the model seems to make the predicted values worse than the first model for the first five future data points

In summary, the ARIMA (2,2,1)x(0,0,0) model does a better job at predicting the data as the average difference between predicted value and actual value is smaller .

Section 4:
Appendix A: Raw Data

Series ID: SP500
Source: S&P Dow Jones Indices LLC
Release: Standard & Poors (Not a Press Release)
Seasonal Adjustment:,Not Seasonally Adjusted
Frequency: Daily
Units: Index
Date Range: 2015-12-17 to 2016-03-01
Last Updated: 2016-03-03 8:36 PM CST

DATE, VALUE

2015-12-17,2041.89
2015-12-18,2005.55
2015-12-21,2021.15
2015-12-22,2038.97
2015-12-23,2064.29
2015-12-24,2060.99
2015-12-25,#N/A
2015-12-28,2056.50
2015-12-29,2078.36
2015-12-30,2063.36
2015-12-31,2043.94
2016-01-01,#N/A
2016-01-04,2012.66
2016-01-05,2016.71
2016-01-06,1990.26
2016-01-07,1943.09
2016-01-08,1922.03
2016-01-11,1923.67
2016-01-12,1938.68
2016-01-13,1890.28
2016-01-14,1921.84
2016-01-15,1880.33
2016-01-18,#N/A
2016-01-19,1881.33
2016-01-20,1859.33
2016-01-21,1868.99
2016-01-22,1906.90
2016-01-25,1877.08
2016-01-26,1903.63
2016-01-27,1882.95
2016-01-28,1893.36
2016-01-29,1940.24
2016-02-01,1939.38
2016-02-02,1903.03
2016-02-03,1912.53
2016-02-04,1915.45
2016-02-05,1880.05
2016-02-08,1853.44
2016-02-09,1852.21
2016-02-10,1851.86
2016-02-11,1829.08
2016-02-12,1864.78
2016-02-15,#N/A

2016-02-16,1895.58
2016-02-17,1926.82
2016-02-18,1917.83
2016-02-19,1917.78
2016-02-22,1945.50
2016-02-23,1921.27
2016-02-24,1929.80
2016-02-25,1951.70
2016-02-26,1948.05
2016-02-29,1932.23
2016-03-01,1978.35

Appendix B: R Commands