Colombia Coffee Milds prices Time-Series analysis

Robin Hornak

1. Background

International Coffee organization classifies coffee exporting countries into four groups: Colombian Milds, Other Milds, Brazilian Naturals and Robustas. This report analyzes the times series of Colombian Coffee Milds prices between 2005-2016. Colombian Coffee Milds refer to green coffee produced in countries of Colombia, Tanzania and Kenya. Green coffee represents the stage when coffee was picked, depulped (fleshly layer removed and leaving the beans itself) and dried. The next stage of the processing would be roasting of the coffee beans. Colombian Milds are of Arabica variety and they are dried using the wet method. This processing method requires large quantities of water and so only more developed farms have the means to adopt this method. However, these coffees are regarded as of higher quality and of better flavor resulting in their large popularity amongst coffee professionals and their subsequent higher price.
Coffee prices are dependent on weather, demand from consumers, seasonality, exchange rate, economic outlook and many other factors. Specifically the price of Colombian Milds might be dependent on changes in the sentiment and attitude amongst coffee drinkers. For example, in recent years customers care more about the natural flavor of coffee and tend to be more interested in good coffee in general and they are willing to pay higher prices for a better product.
For example, as shown in Figure 1, there seems to be a large increase in the price since January 2009 and peaking in June 2011. This could be partially explained by the heavy rains and low temperatures damaging the crops and resulting in poor harvests. Another reason could be the weak dollar and the increased demand from China, India and Brazil who have been increasingly adopting the 'Western coffee culture' and drinking habits in the recent years. Following that, the price fell quite steeply, reaching around 120 US cents/lb in December 2013. This drop could be potentially a result of a strong crop in Brazil, a mass replanting programme in Colombia and/or by an excessive increase in supply as a response to the high prices in 2011. In 2014 the price of Colombian Milds again rapidly increased reaching some 200 cents/lb in October 2014. Most common explanation for this rise was the severe and protracted draught in the producing regions leaving damaged crops and an outbreak of a fly-like insect known as 'broca' which harms the crops as well.

One of the main motives for this paper was to explore whether it is possible to model the Colombian Coffee Milds prices using concepts of Time-Series analysis and to potentially build an accurate predictive model.

2. Data

The data set is a monthly price in US cents/lb of green coffee of the Colombian Milds and covers a period from January 2005 until December 2016. It was published by the International Coffee Organization and is not adjusted for inflation. Therefore, before conducting any analysis, the prices were first adjusted for inflation with a base year 2008. It allows for a more precise and useful analysis. January 2006 until December 2015 was used as training set for the time series modeling and the 12-month period in 2016 was used as a test set to compare the predictive power of different models that performed well on the training set.

As displayed in Figure 1, the data does not resemble a stationary time-series since both variance and mean seem to vary throughout the series. Taking a first difference of the series led to a much more suitable data set for time series analysis. The mean seems to be constant and the variance appears to also be somewhat constant as well, with a few large values at the end of 2011 and in the beginning of 2014 as shown in Figure 2.
Following the differencing, it was important to investigate whether there was any significant trend in the data and if so then remove it. Fitting a simple linear regression did not prove any significant trend in the series as also graphically displayed in Figure 3 using a fitted line. This result means that the expected differenced value does not increase/decrease over the time period.

For the remainder of the analysis, the first differenced series was used, as it closely resembled a stationary time-series.

3. Frequency Domain Analysis

Both the ACF and the PACF shown in Figure 4 of the series seemed to be tailing off after lag 1 with potentially significant seasonal components at lags 8 and 26 suggesting that a seasonal ARIMA might be a good fit.
All the models' AIC's and RMSE's fitted in the following section can be found in Figure 7.
Fitting ARIMA(1,1,1) seemed to fit the data quite well with an AIC 5.7278. The standardized residuals in Figure 5 seem to be oscillating around zero with a constant variance between 2005-2009 followed by a more inconsistent level of variability in 2010-2015, there also seems to be almost no significant correlation between the residuals and the Q-Q plot suggests residuals follow a normal distribution. However, the MA(1) parameter had a p-value 0.8188 suggesting that this term should be left out from the model. This was proved to be true by fitting an ARIMA(1,1,0) with an AIC 5.713 suggesting that this is a more accurate model than ARIMA(1,1,1). By inspecting Figure 6, it also seems that the residuals are normally distributed and are not significantly correlated. The variance of the residuals seems to be constant between 2005-2009, but it appears to increase between 2010-2012 with another spike in mid-2014.
Since adding an MA component does not improve the model, the next model to assess was an ARIMA(2,1,0) to see whether adding an additional AR component improves the model. With an AIC 5.7275 it seems that addition of an AR component does not improve the model and the ARIMA(1,1,0) fits the data more accurately. Analysis of residuals does not seem to provide any improvement over the previous models with variance still increasing at certain time periods. ARIMA(1,1,0) seem to be therefore the best fit to this time-series whilst still excluding any seasonal components.

As mentioned at the beginning of this paper, coffee is a seasonal commodity and adding a seasonal component to the current ARIMA(1,1,0) model by applying seasonal ARIMA models could result in more accurate models. For interpretability reasons, there aren't any models fitted using backshifts of the seasonal periods of the ARIMA model (i.e. D=0). The following analysis considers adding seasonal AR and seasonal MA components to the ARIMA(1,1,0).
Firstly, ARIMA(1,1,0)(0,0,1) models with various numbers of periods (between 1:30) per season were fitted where an ARIMA(1,1,0)(0,0,1)[26] had the lowest AIC 5.632 and ARIMA(1,1,0)(0,0,1)[8] had the second lowest AIC 5.658 as somehow expected from the ACF and PACF figures. Inspecting Figure 8, the residuals of ARIMA(1,1,0)(0,0,1)[26] appear to be randomly scattered around zero with a constant variance throughout most of the series. There are, nonetheless, a few values that have higher variance than the rest of the series, primarily in the beginning of 2009 and 2014. Inspecting the ACF of the standardized residuals, there seems to be almost no significant correlation between the residuals except one value at lag 8. Moreover, the Q-Q plot suggests normality of the residuals with a very slight heavier tails on both ends of the distribution. Adding an additional seasonal MA component resulting in ARIMA(1,1,0)(0,0,2)[26] did not lead to a better model with AIC 6.644 and with many of the residuals being significantly correlated violating the assumption of ARIMA models. Next model to investigate was an ARIMA(1,1,0)(1,0,1)[26] with a seasonal AR(1) component added to the model. Similarly as before, adding a seasonal AR(1) results in a higher AIC 5.646 and a number of residuals that are significantly correlated at lags 1,6 and 7 indicating that this model does not fit the data series. Since the interest is also in choosing a model that has high predictive power for 2016, as will be discussed in the following section, an ARIMA(1,1,0)(1,0,1)[16] was investigated further despite having a slightly higher AIC 5.696 than the other models that fit the data well. The residuals appear to be randomly scattered around zero and constant variance between 2005-2009 and a slightly more varying variance between 2010-2016 as displayed in Figure 9. Except a couple of lags, the residuals are not significantly correlated. Majority of the points in the Q-Q plot lie on the straight line, suggesting normality with a few points on the right side of the distribution having heavier tails.
Last model to investigate was an ARIMA(1,1,0)(1,0,0) with various numbers of periods (between 1:30) per season. From the 30 models fitted, ARIMA(1,1,0)(1,0,0)[26] has the lowest AIC 5.651. Residuals do not appear to be significantly correlated and the Q-Q plot suggests that they seem to be somewhat normally distributed with slightly heavier tails on both sides of the distribution. This model, however, has a larger AIC than ARIMA(1,1,0)(1,0,1)[26] and the residuals do not seem to follow normal distribution as closely as in ARIMA(1,1,0)(1,0,1)[26].
From all the models discusses the ARIMA(1,1,0)(1,0,1)[26] appears to be the best fit to the time series by investigating its AIC and residuals.

As mentioned earlier, there was an interest in choosing a model that is able to fit the test set accurately. In this report root mean squared error (RMSE) is used to evaluate the predictive power of the models on the test set. Despite ARIMA(1,1,0)(1,0,1)[26] being the best model to fit the test set, it didn't prove to be accurate in predicting the coffee prices in 2016, as displayed in Figure 11. The model (red line in Figure 11) predicts a fall in the price of coffee throughout 2016, but the price actually increased resulting in an RMSE of 73.56. The best model found to predict the prices for 2016 was an ARIMA(1,1,0)(1,0,1)[16] with a much lower RMSE of 58.85. This model correctly predicted the higher prices in the first half of 2016 and it also accurately predicted the slight reduction in the latter part of the year, as shown in Figure 12.

4. Non-parametric Spectral analysis

Spectral analysis in this paper was also conducted on the differenced time series as in the Frequency domain analysis.
There seem to be no apparent frequencies that would stand out in the raw periodogram displayed in Figure 13. A smoothed periodogram in Figure 14 with Kernel(4) takes into account eight neighboring observations and appeared to be the best choice to detect any potentially significant frequencies. By examining Figure 14, there seem to be two main frequencies at 18/135 and 52/135. However, the plot still had some low oscillating frequencies and therefore further smoothing could bring out the main frequencies clearer. Applying a Modified Kernel(3,4), displayed in Figure 15, resulted in a very smooth graph that clearly pointed out to three main frequencies at 2/135, 18/135 and 52/125. The peak at 18/135 (cycle every 7.5 months) is similar as identified in the ACF of the Frequency domain analysis. The lower ends of the confidence intervals for the three frequencies are 8.64, 9.78 and 5.37 respectively, as displayed in Figure 16. Deciding whether any of these frequencies are significant is problematic in this situation. At a given significance level a peak is claimed to be significant if the lower confidence limit for the spectral value is still greater than the baseline level. However, a baseline is a level of spectrum from which the peaks seem to emerge. In this case, as shown in Figure 15, there doesn't appear to be any obvious baseline for the spectrum. The first half of the frequency values seems to have a higher baseline than the second half of the frequencies. Therefore, we cannot prove that any of the peaks are significant.

4. Parametric Spectral analysis

The plotted AIC in Figure 17 of different fitted AR(p) models selected AR(1) to be the best fit to the time series with an AIC of 0.000. This is clearly displayed in Figure 17 with a maximum order of thirty, where the AIC increases with almost every additional AR component added to the model. The spectrum of the AR(1) model does not have any peaks as it is simply a decreasing function. This suggests that there is no periodicity in the differenced time series after applying the AR(1) model.

5. Next steps

As mentioned at the beginning of the paper, coffee prices are very sensitive to weather and therefore in the future analysis, I would like to incorporate weather as a multiple-series analysis. Secondly, there might potentially be long memory relationships in the series, and so analyzing a longer time span by getting access to a bigger data set might reveal some valuable relationships. Lastly, the price of Colombian Coffee Milds is very dependent on prices of coffees from other exporting countries, so including multiple-series and cross spectra analysis with other coffee prices might significantly improve the model.

6. Conclusion

In conclusion, using the differenced time series, there were many very good models that fit the train set well. The best model according to the AIC was a seasonal ARIMA(1,1,0)(0,0,1)[26] with residuals following the model assumptions quite well. However, at the same time, this model was one of the worst in predicting the Colombian Milds Coffee prices in 2016. It predicted the price to go down (Figure 11 red line), when in reality it went up.
Therefore, I decided to compromise and choose the seasonal ARIMA(1,1,0)(1,0,1)[16] to be the best model. This model has a slightly higher AIC then ARIMA(1,1,0)(0,0,1)[26], however, its predictive

power for 2016 was outstanding. It predicted both the increase (Figure 12 red line) in the first part of 2016 as well as the fall of the price in the latter part of the year.
Both parametric and non-parametric spectral analysis did not prove to be useful in explaining the differenced Colombian Milds Prices time-series.

Appendix

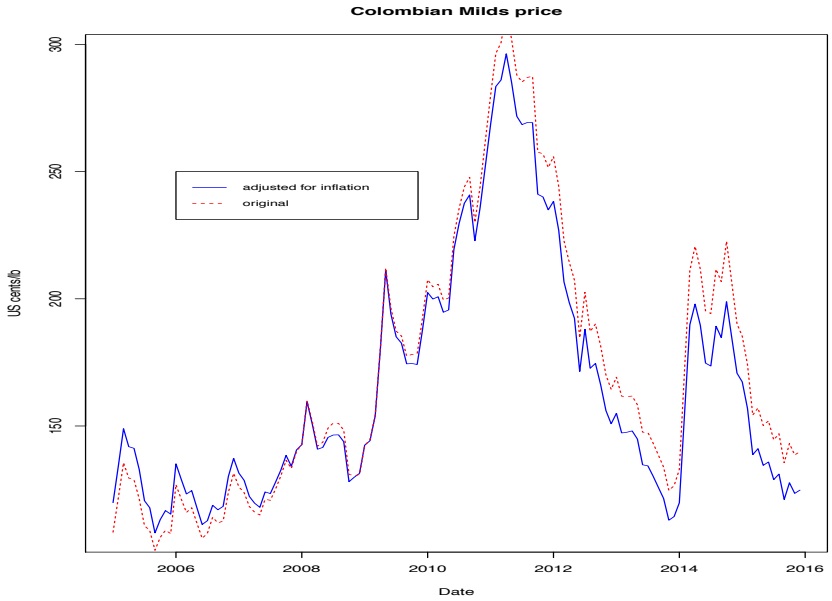

Figure 1

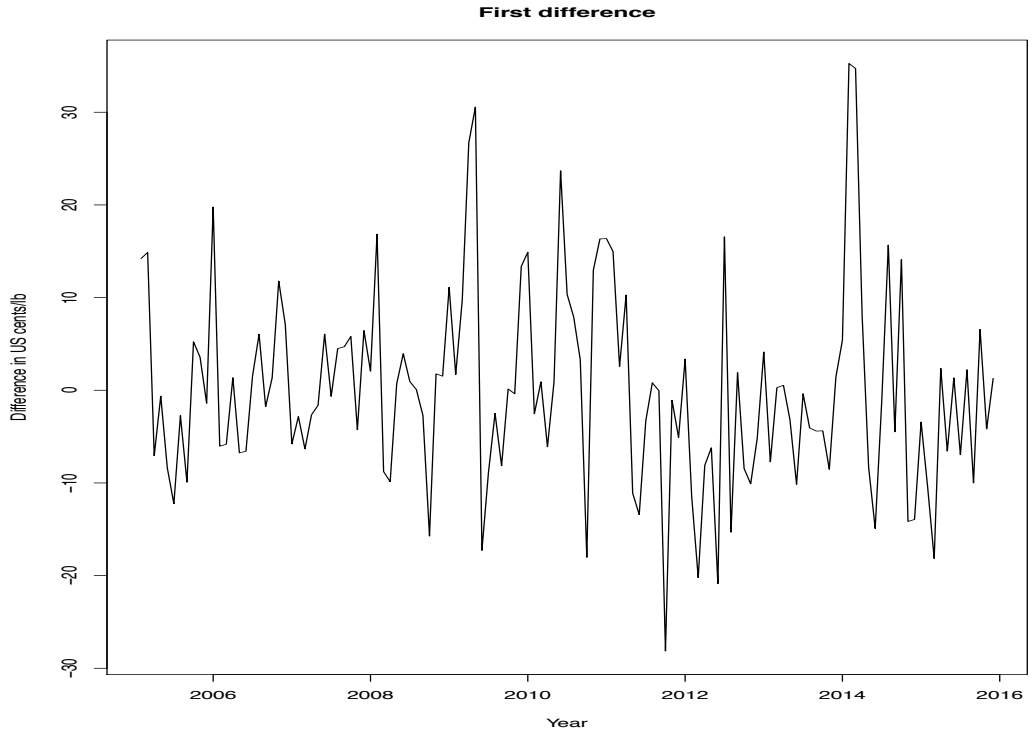
First difference
Difference in US cents/lb
30
20
10
0
-10
-20
-30
2006
2008
2010
2012
2014
2016
Year

Figure 2

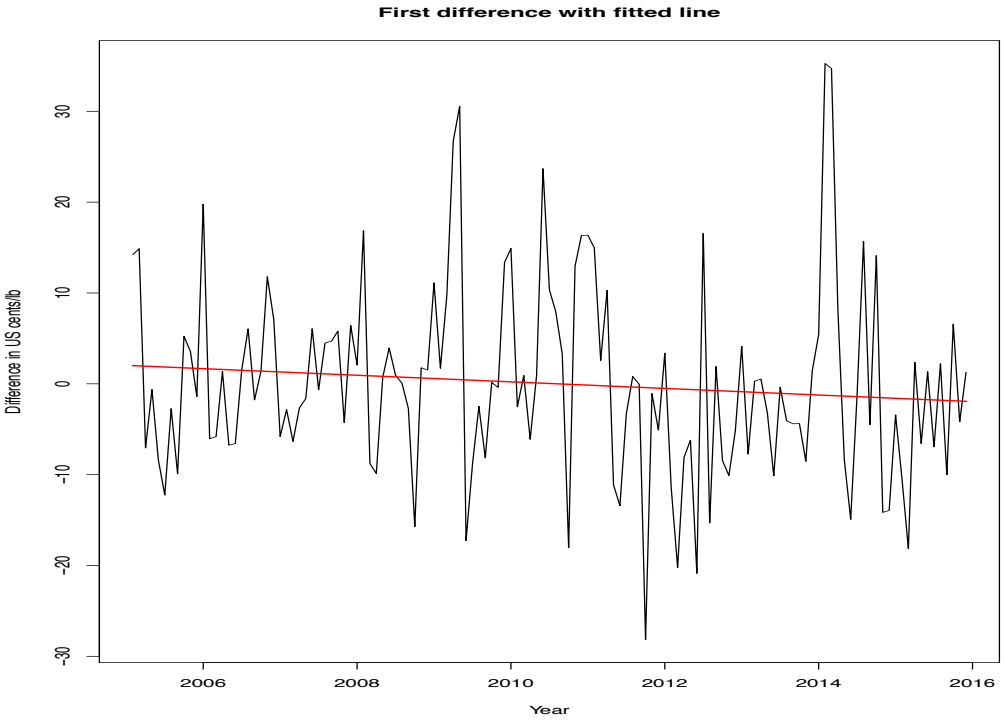

Figure 3

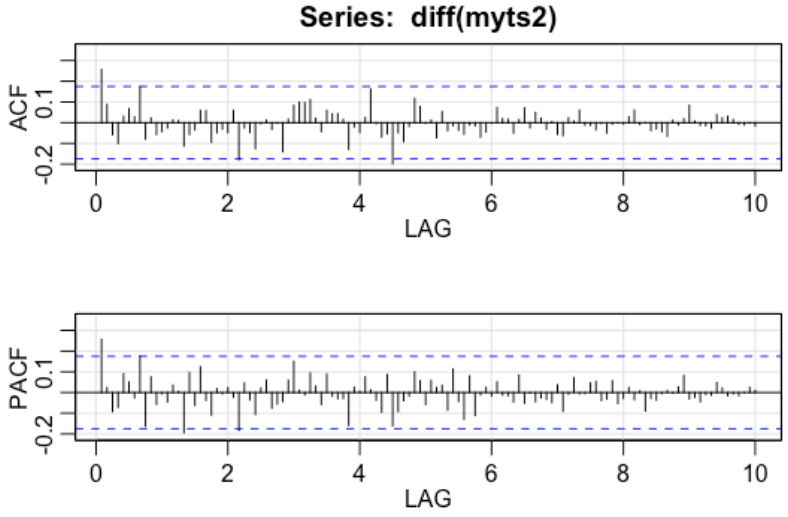

Figure 4

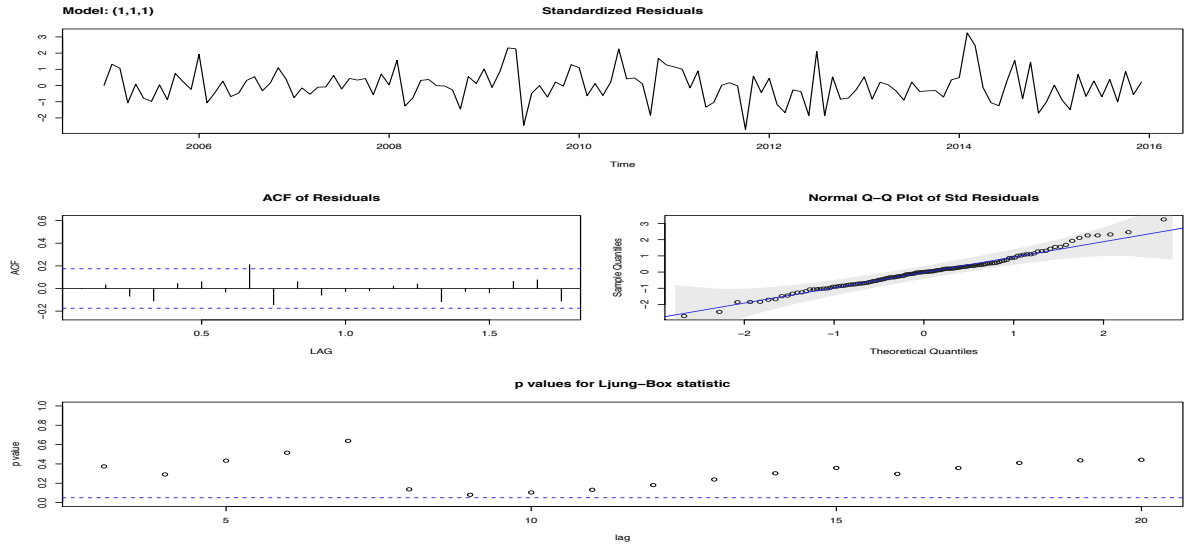

Figure 5

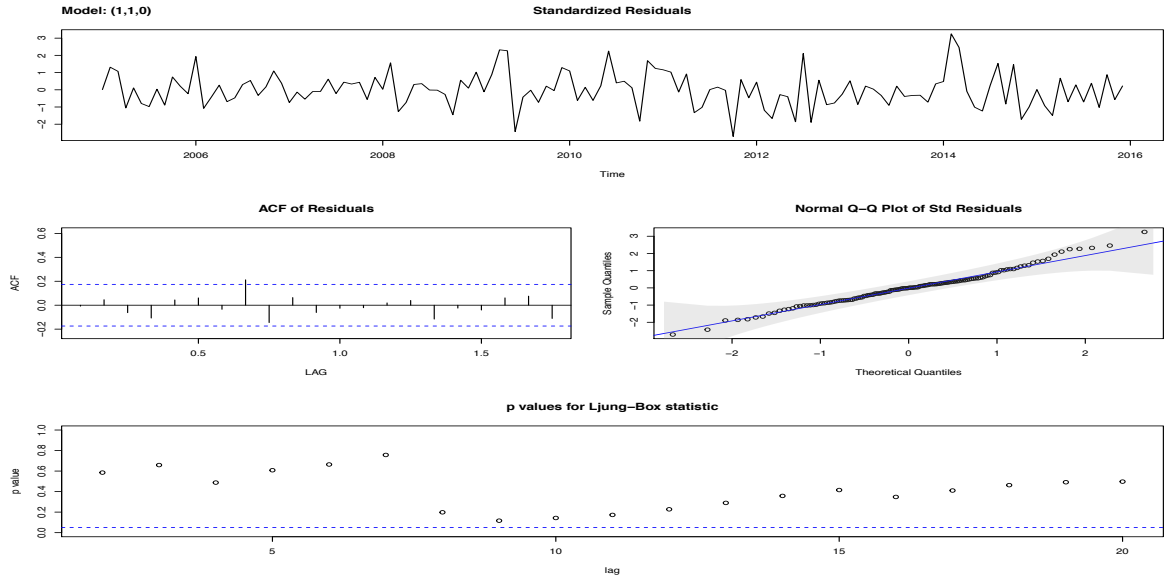

Figure 6

ARIMA	AIC	RMSE
(1,1,0,0,0,0,0)	5.713029242	66.54484966
(1,1,0,0,0,1,26)	**5.632495543**	**73.55522255**
(1,1,0,0,0,1,16)	5.70815213	62.05637137
(2,1,0,0,0,0,0)	5.727451428	66.62022915
(1,1,1,0,0,0,0)	5.727771281	66.62022915
(1,1,0,1,0,1,16)	**5.695709144**	**58.84798173**

Figure 7

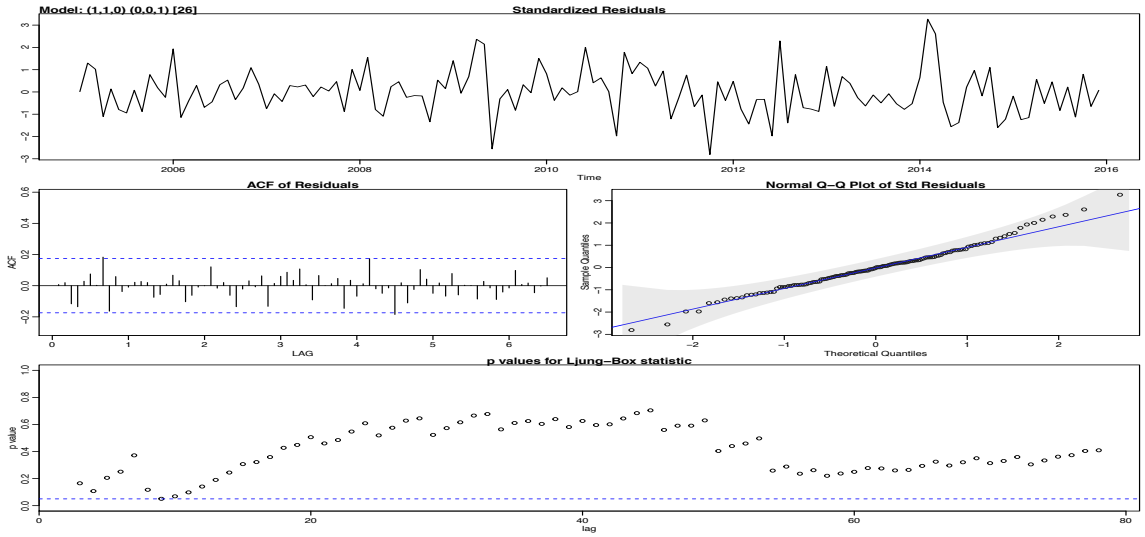

Figure 8

ARIMA(1,1,0,1,0,1,16)

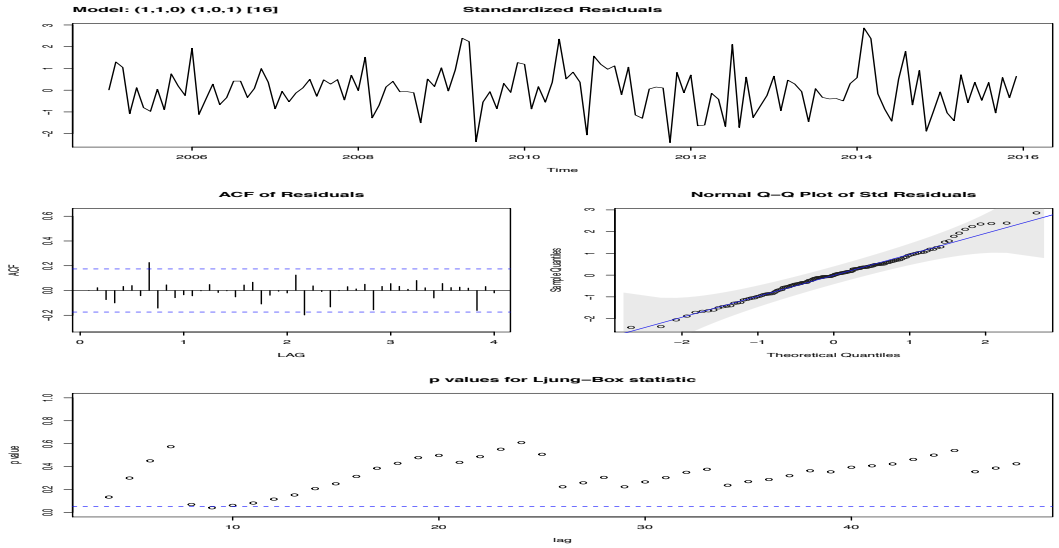

Figure 9

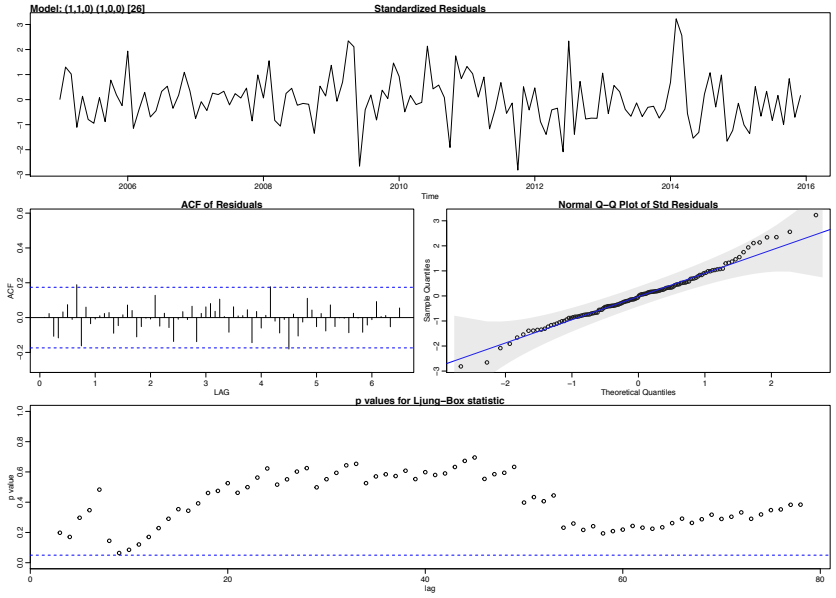

Figure 10

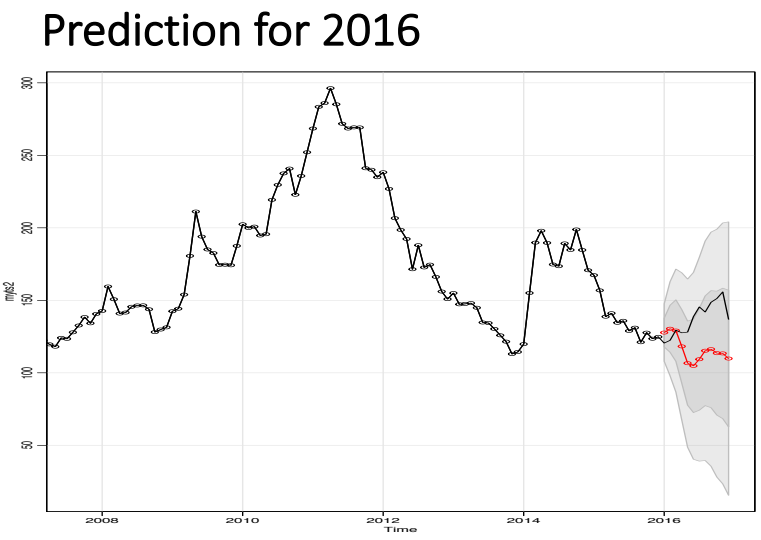

Figure 11

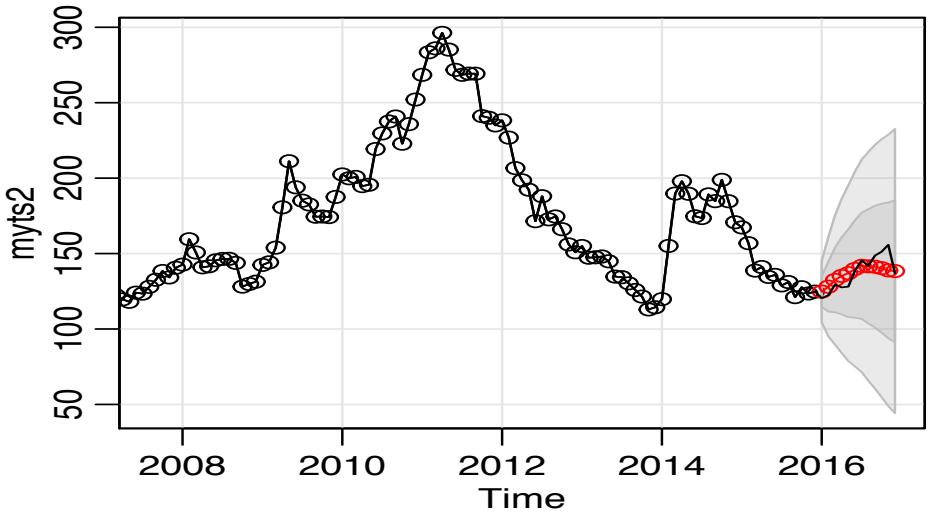

Figure 12

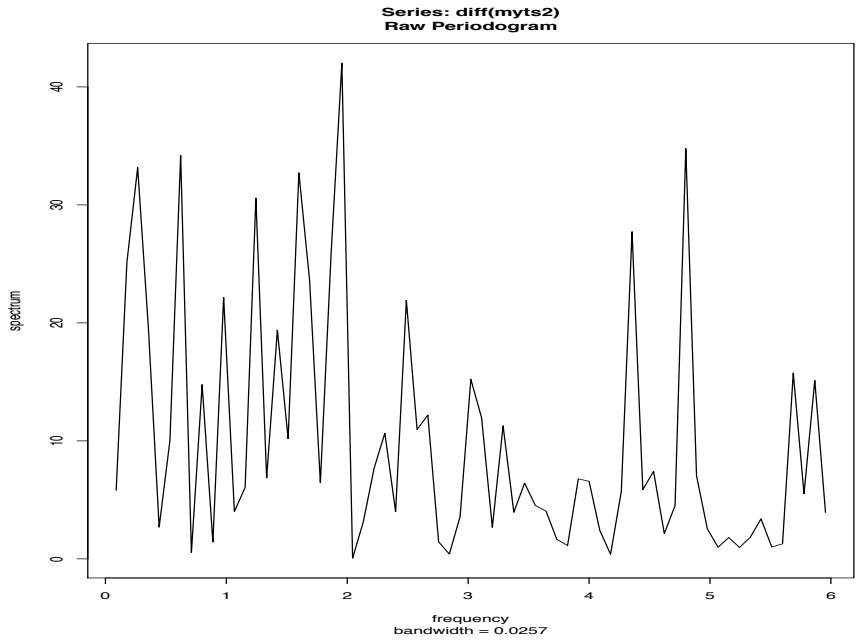

Figure 13

Kernell(4)

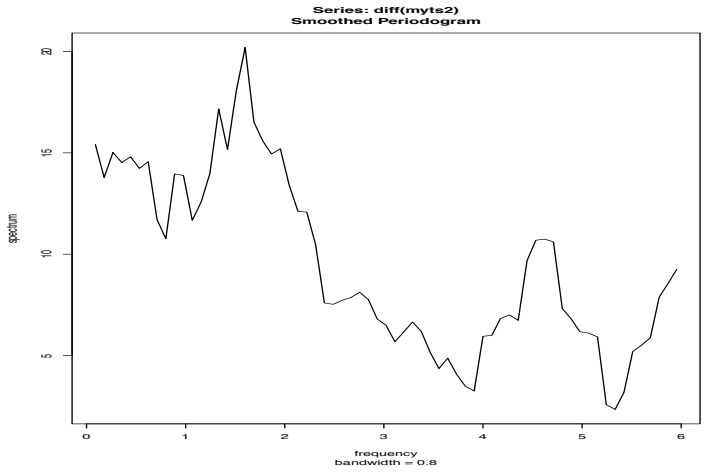

Figure 14

Modified Daniel(3,4)

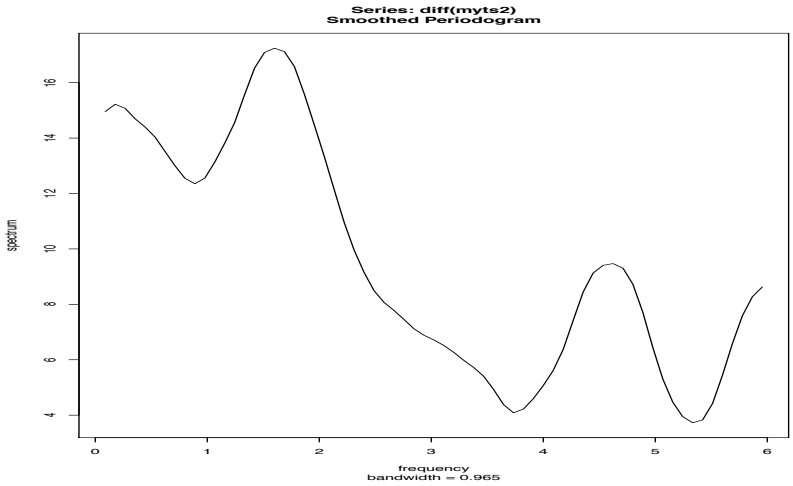

Figure 15

Frequency	Power	Lower	Upper
2/135	15.2186543	8.635536008	33.1401072
18/135	17.24408358	9.784827341	37.55067743
52/135	9.469695969	5.373398916	20.62118853

Figure 16

Parametric spectral analysis

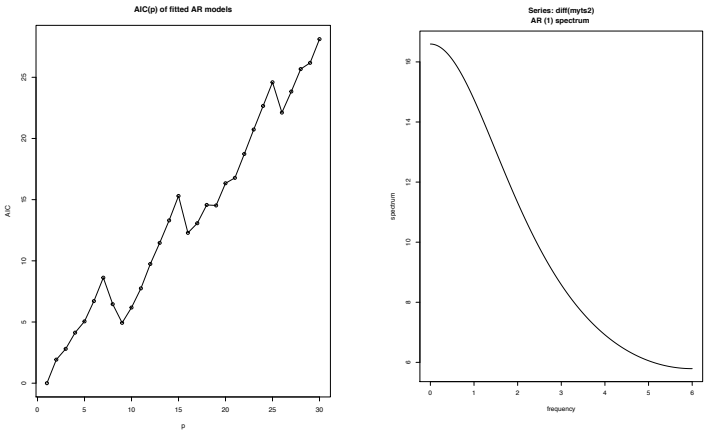

Figure 17