Amiel Chong
Spring 2011

Time Series – Student Project
Mild Arabica Coffee Bean Prices

Sunday – It's a day where at around 7pm, it hits – the beginning of the work week is tomorrow. We love our weekends and dread the weekdays, but however, our one saving grace that will not only keep us awake, but will alleviate the pain of Monday morning is none other than a delicious hot (or iced if its summertime) cup of, you guessed it, coffee. For my time series project I chose my favorite type of coffee bean; that being the mild Arabica bean. The information below will effectively use the 5 year monthly historical prices of the Arabica coffee bean to fit an autoregressive process and effectively predict 1 year's worth of coffee bean prices.

Data

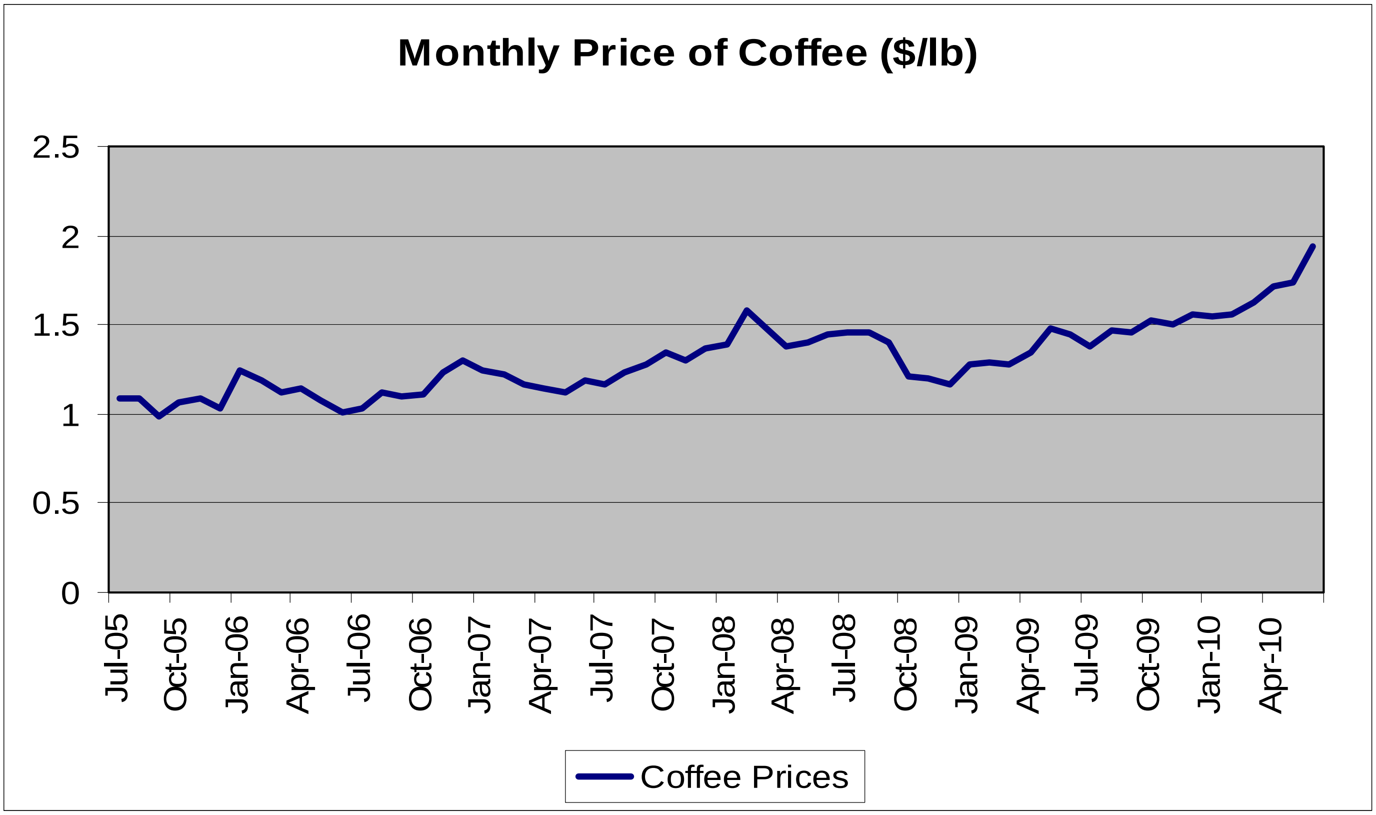
The data I used for my project comes from the following URL - http://www.indexmundi.com/commodities. A simple Google search led me to this website. Here I exported the data into an excel file – a great feature of the website. The data from this particular website was listed in cents/lb, which I converted to $/lb by simply dividing each price by 100. From here I found the monthly prices of the Arabica coffee bean from July 2005 to June 2011. I wanted to predict 1 year's worth of prices so my data set is simply the monthly prices from July 2005 to June 2010.

The data described above, along with the backup information and charts contained in the following report, can be found in the attached spreadsheet *chong - coffee statistics - final project.xls*. The full data set taken from the website can be found on the tab 'Original Data'.

Analysis of Original Arabica Coffee Bean Data

The initial step in my time series analysis was to graph the price of coffee over the five year time period that I had chosen, July 2005 to June 2010 shown in Figure 1, below. Also see tab 'Coffee Price Statistics (CPS)' in the attached Excel spreadsheet for the graph with data points. Coffee is delicious all year round, and Figure 1 below shows that there isn't much, if any, seasonality present in this time series. Because of this, I opted not to adjust for seasonality. The price of coffee beans slowly increases with the following anomalies – a tremendous increase between January and April 2008, with a slower but decreasing price between July and October 2008, a small spike between April and July 2009, with rising prices through June 2010. Using the Figure 1, we have a basis where using time series techniques can help model a process to determine the future price of coffee.

Figure 1



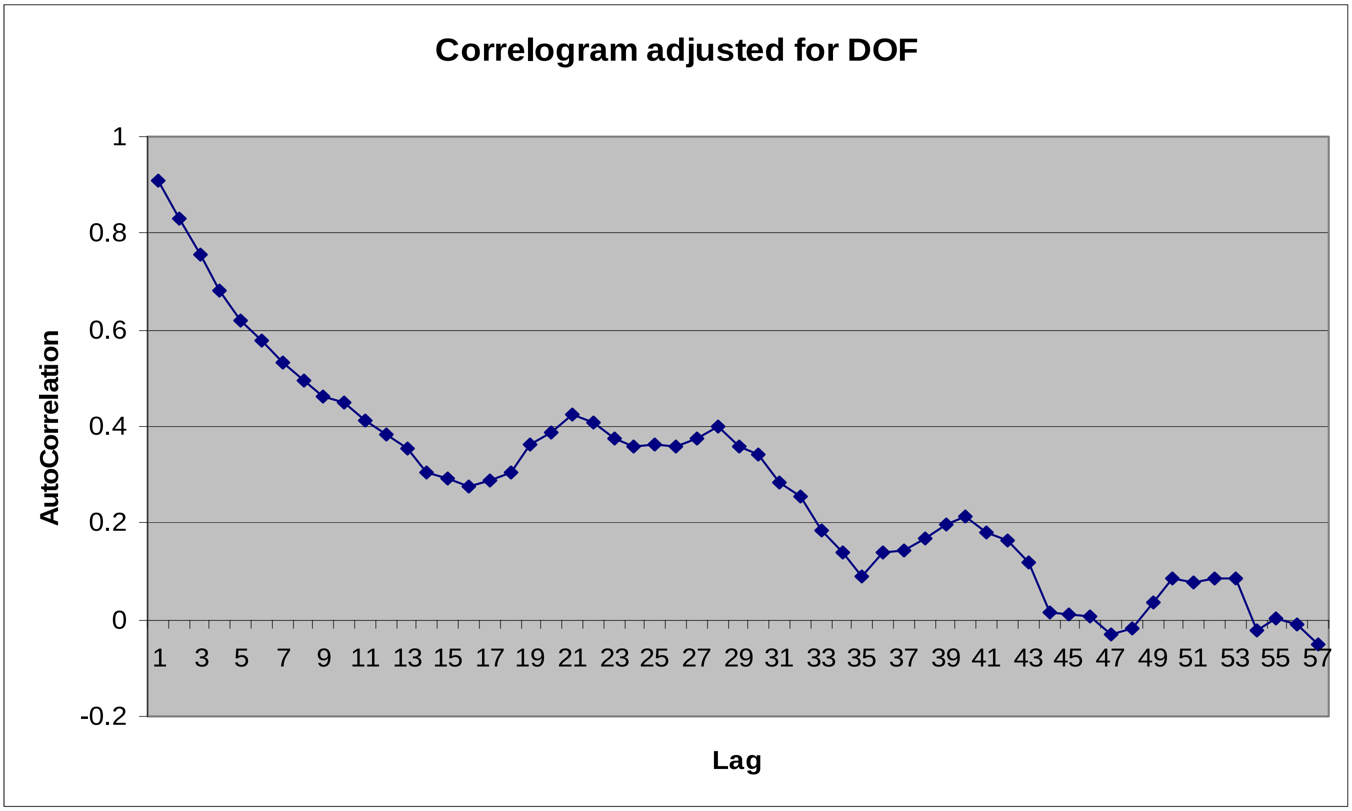
Autocorrelation of Original Data Series

Let's a graph a correlogram in order to further analyze the prices in Figure 1, above. This graph shows the sample autocorrelation function at each lag point. The formula used to calculate the sample autocorrelation function, found on page 46 of the text (3.6.2), is shown below.

$$r_k = \frac{\sum_{t=k+1}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \qquad \text{for } k = 1, 2, \ldots$$

The attached spreadsheet contains a tab called 'Coffee Price Statistics (CPS)' which shows the development of the points on this graph. Figure 2 below shows he graph of the sample autocorrelation against lag times.

Figure 2



The points on the graph are mainly positive but have a decreasing trend. At around lag 47 the autocorrelation becomes negative, then increases and remains relative constant from lags 49 through 53. After lag 54, the graph remains negative. If this series is stationary, we can assume the following for an AR(1) model:

$$|\phi| < 1$$

Using Excel's Data Analysis functions we get the following:

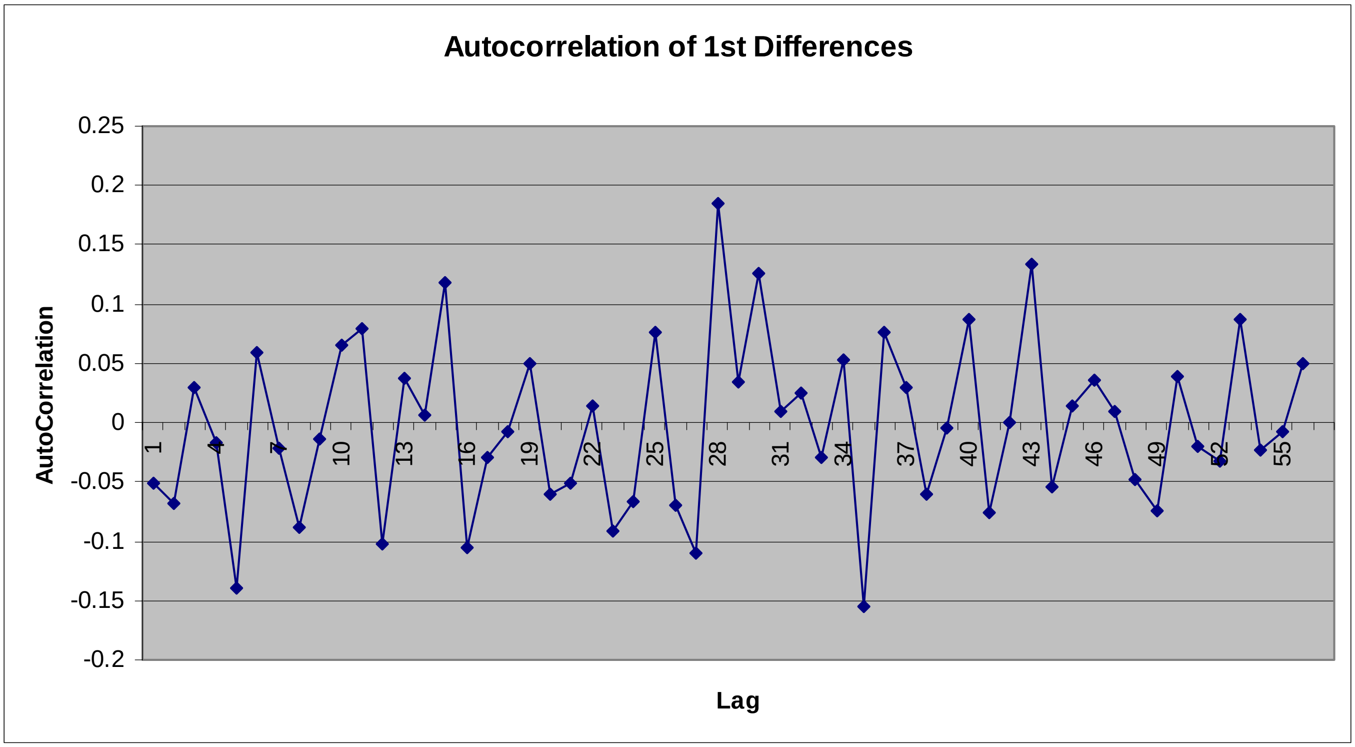
	Coefficients
Intercept	0.010703
Φ	1.002864

This information is shown on tab 'CPS AR(1)' of the attached spreadsheet. Since Φ > 1, although very, very slightly, we can say that this is a non- stationary process. We don't necessarily need to use this type of analysis to come to this conclusion. We can instead look at the graph visually and notice that Figure 2 does not fluctuate around a constant mean. We can try to find stationarity by taking the autocorrelation of first differences.

Autocorrelation of First Differences

The function used to calculate the fist difference was: $W_t = Y_t - Y_{t-1}$. Figure 3 below shows the correlogram for the sample autocorrelation function, r_k, versus the lag, k. The data and calculations used can be found in the attached spreadsheet on the tab '1st Diff for AR(X)'. Is the correlogram below stationary?

Figure 3

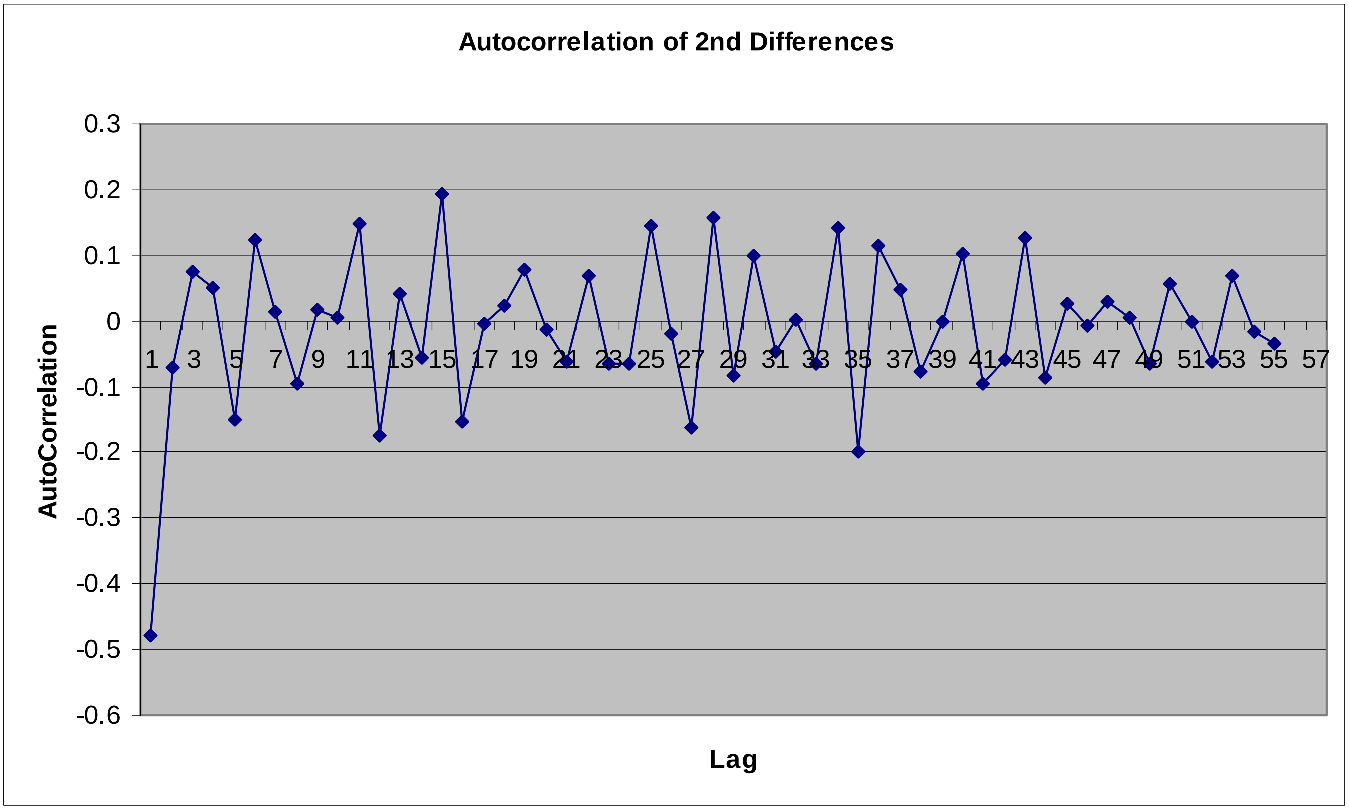


The graph of the sample autocorrelation of the first differences shows a trend towards zero, but the fluctuations do not seem to be decreasing. Figure 3 does not seem to be stationary. We can try the same analysis above using second differences for stationarity.

Autocorrelation of Second Difference

In order to calculate the second differences, I used the function: $Z_t = W_t - W_{t-1} = (Y_t - Y_{t-1}) – (Y_{t-1} - Y_{t-2})$.
I then used the autocorrelation function used in Figure 2 and Figure 3 to develop the sample autocorrelation graphed versus the monthly lag as shown in Figure 4. The data for Figure 4 can be found on tab '2nd Diff for AR(X)' in the attached Excel spreadsheet.

Figure 4



Since the graph above trends towards zero, and the fluctuations in the later lags become smaller – we can assume stationarity. We will use the second difference for our model analysis.

Model Analysis (Using 2nd Differences)

The next step in the process is to fit several models to the data and decide which model will be the most appropriate choice for predicting future monthly prices of Arabica coffee beans. We will look for the appropriate models using the following processes: AR(1), AR(2), and AR(3).

The formulas for these three models are as follows:

AR(1): $Y_t = \varphi Y_{t-1} + e_t$
AR(2): $Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + e_t$
AR(3): $Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varphi_3 Y_{t-3} + e_t$

I used the Regression Add-In for Excel to solve the above equations - the output is below:

AR(1)	*Coefficients*
Intercept	0.005477
Φ_1	-0.49386

AR(2)	*Coefficients*
Intercept	0.004256
Φ_1	-0.66353
Φ_2	-0.39086

AR(3)	*Coefficients*
Intercept	0.004582
Φ_1	-0.77208
Φ_2	-0.55802
Φ_3	-0.24889

The resulting equations are below

AR(1): $Y_t = -0.49386Y_{t-1} – 0.005477$
AR(2): $Y_t = -0.66353Y_{t-1} – 0.39086Y_{t-2} – 0.004256$
AR(3): $Y_t = -0.77208Y_{t-1} – 0.55802Y_{t-2} – 0.24889Y_{t-3} – 0.004582$

All of the data sets and the resulting regression add-in tool output can be found in the attached spreadsheet in tabs labeled ‘2nd Diff for AR(X)’ and ‘2nd Diff AR(X)’, for X = 1,2,3.

One thing to note is that the sum of the coefficients for each of the models above is less than 1, and $|\varphi_p|<1$ for p = 1, 2, 3 in each of the above models. These two inequalities are necessary for stationarity, however, this does not hold for my previous assumption (and is only stationary for the AR(1) and AR(2) process).

Figures 5, 6, and 7 below show the actual second differences vs an AR process with X parameters, for X = 1,2,3

Figure 5

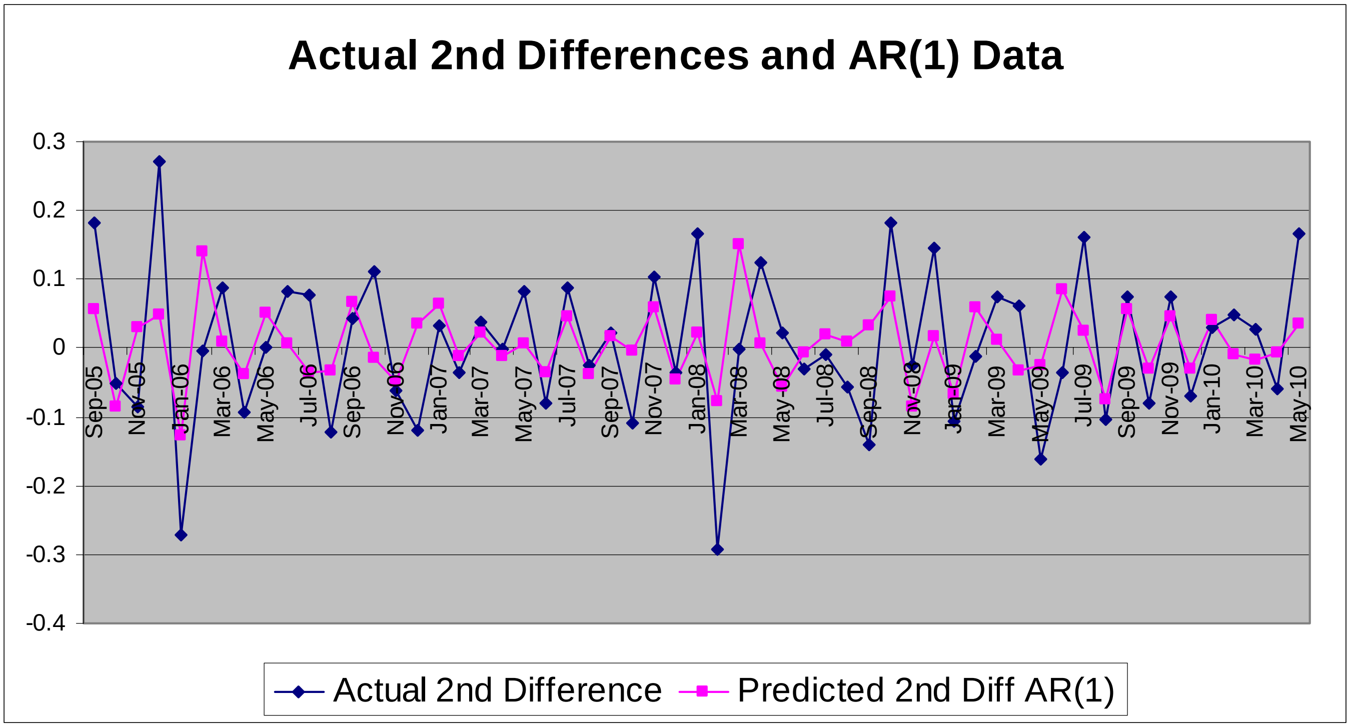


Figure 6

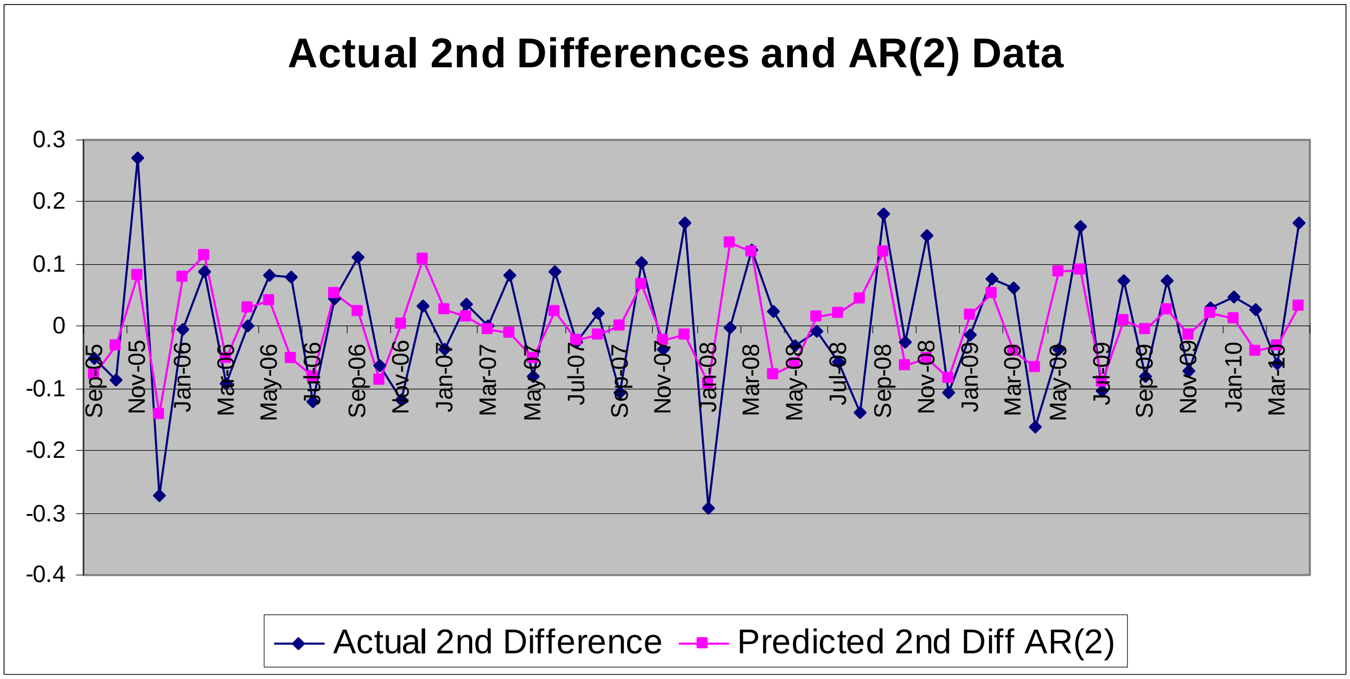
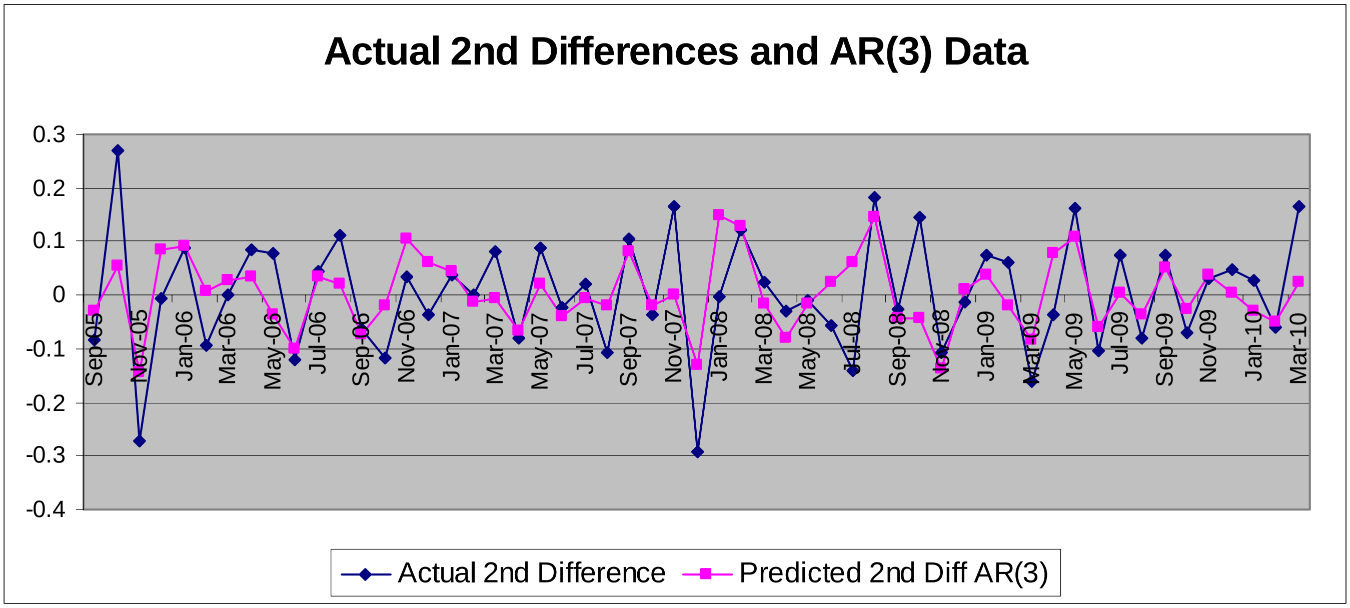


Figure 7



The data for Figure 5, 6, and 7 can also be found in the attached spreadsheet on the tab ‘2nd Diff for AR(X)’. Strictly looking at the graphs, I would say that the AR(2) and AR(3) graphs are a better fit for the data than the AR(1) graph. These two graphs were able to get closer to the spikes in March of 2004 and October of 2008. However, I will need to analyze the models more closely using various techniques to help me determine which model will be the best fit.

Model Statistics & Testing

Now that I have three potential models to choose from, I need to do some statistical testing to help with the final decision. The first statistic I will use is the Durbin-Watson Statistic. This statistic is used to test for serial correlation. I developed this statistic using the residuals from the regression tool output in excel. The calculations can be found in the attached spreadsheet on the tabs 'AR(1) Model', 'AR(2) Model' and 'AR(3) Model'. The result of this test for each model is shown below:

Model	Durbin-Watson Statistic
AR(1)	2.31534
AR(2)	2.27984
AR(3)	2.08967

The following is a summary of the Durbin-Watson (DW) statistic:

Table 1 Range of the Durbin-Watson Statistic

Value of DW	Result
$4 - d_l < DW < 4$	Reject null hypothesis; negative serial correlation present
$4 - du < DW < 4 - d_l$	Result indeterminate
$2 < DW < 4 - du$	Accept null hypothesis
$d_u < DW < 2$	Accept null hypothesis
$d_l < DW < du$	Result indeterminate
$0 < DW < d_l$	Reject null hypothesis; positive serial correlation present

The rule of thumb is that a DW Statistic whose values are close to 2 indicates no autocorrelation. In the Arabica coffee data, all three models result in a Durbin-Watson statistic of close to 2, however the AR(3) model seems the closest. Since a Durbin-Watson Statistic of 2 indicates no serial correlation, we can see from the results that these models have almost no serial correlation between them. It would seem that the AR(3) model is the best to use, but to be even more sure we use the Box-Pierce Q Statistic

The Box-Pierce Q statistic tests whether the time series used is a white noise process. The null hypothesis in this test is that the residuals are a white noise process. The test used to here was the Box-Pierce Q statistic. Using the calculations provided by the NEAS spreadsheet *TimeSeriesTechniques.xls*, I developed the Box-Pierce Q statistic for each model. The calculations can be found in the attached spreadsheet on tabs '2nd Diff AR(X)' for X = 1, 2, 3. In the table below, I have shown the Box-Pierce Q statistic, the degrees of freedom for this statistic and the corresponding χ^2 value at the 10% significance level.

Model	Box-Pierce Q Statistic	DOF	χ2 (10%)
AR(1)	34.65301	54	66.58420
AR(2)	27.96296	53	65.42241
AR(3)	18.82981	52	64.29540

The above results show that all three potential models have similar results. For all three of the models, the Box-Pierce Q Statistic is significantly lower than the χ^2 value. We do not reject the null hypothesis that the residuals are a white noise process.

Model Selection

The next step is to select the most appropriate model for this time series. Based on the results above, I would not choose the AR(1) model, since the Durbin-Watson statistic is the farthest from 2, and the Box-Pierce Q statistic is the highest of the three. My choice is now between the AR(2) and AR(3) models. In the table below, I show the adjusted R square values from the excel regression output for all three models.

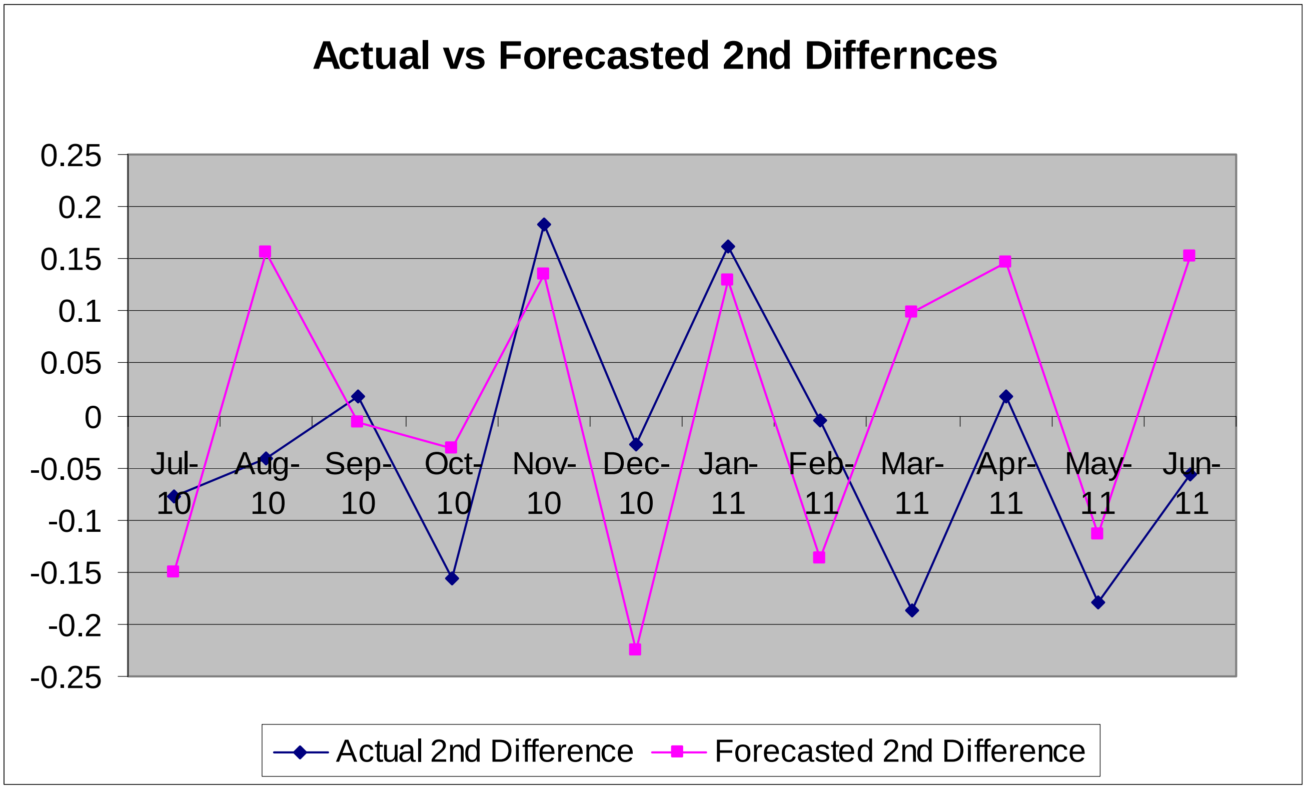
Model	Adjusted R Square
AR(1)	0.224437
AR(2)	0.320558
AR(3)	0.347827

Here, we use the use the Adjusted R Square to help us determine which process to use. The higher the Adjusted R Square, the better the fit of the model. Clearly, it seems that the AR(3) model will fit the data for 2nd differences of the Arabica coffee bean.

Forecast

Now that I have chosen the model, I want to use it to predict the monthly price of the Arabica coffee bean per pound and compare it to the actual results from July 2010 to June 2011. Recall from above that the formula for the AR(3) model is AR(3): $Y_t = -0.77208Y_{t-1} – 0.55802Y_{t-2} – 0.24889Y_{t-3} – 0.004582$. Using this formula I was forecasted the next 12 months of mozzarella prices. The forecasts in Figure 8 below, along with the graph comparing them to the actual second difference results, can be found in the attached spreadsheet on the tab ‘Forecasting 2nd Diff’. The graph is shown below. The graph of the forecasted second differences follows the increases and decreases in the 2nd difference very well, but it seems to fall short in periods of high peaks and be too low in periods of low peaks. However, this is to be expected based on the adjusted R square value being a value of .347827.

Figure 8



Conclusion

The best estimator examined in this project was the AR(3) model. However, since the trend of this analysis indicated that increasing parameters would bring the Durbin-Watson statistic closer to 2, lower the Box-Pierce Q statistic, and increase the adjusted R square value, a greater number of parameters for the AR(X) model would better predict the monthly prices of mild Arabica coffee beans.