John J. Clark
Fall 2014
Time Series Project

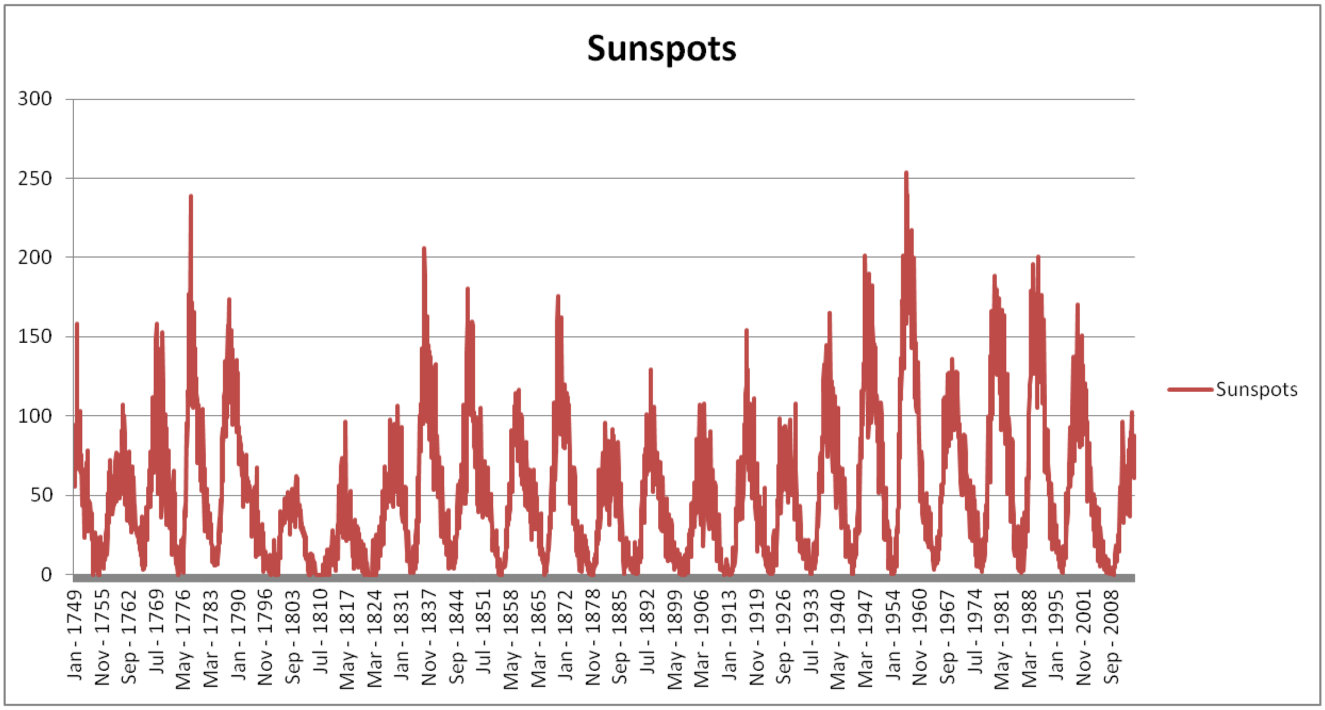
Monthly Sunspot Count

Introduction

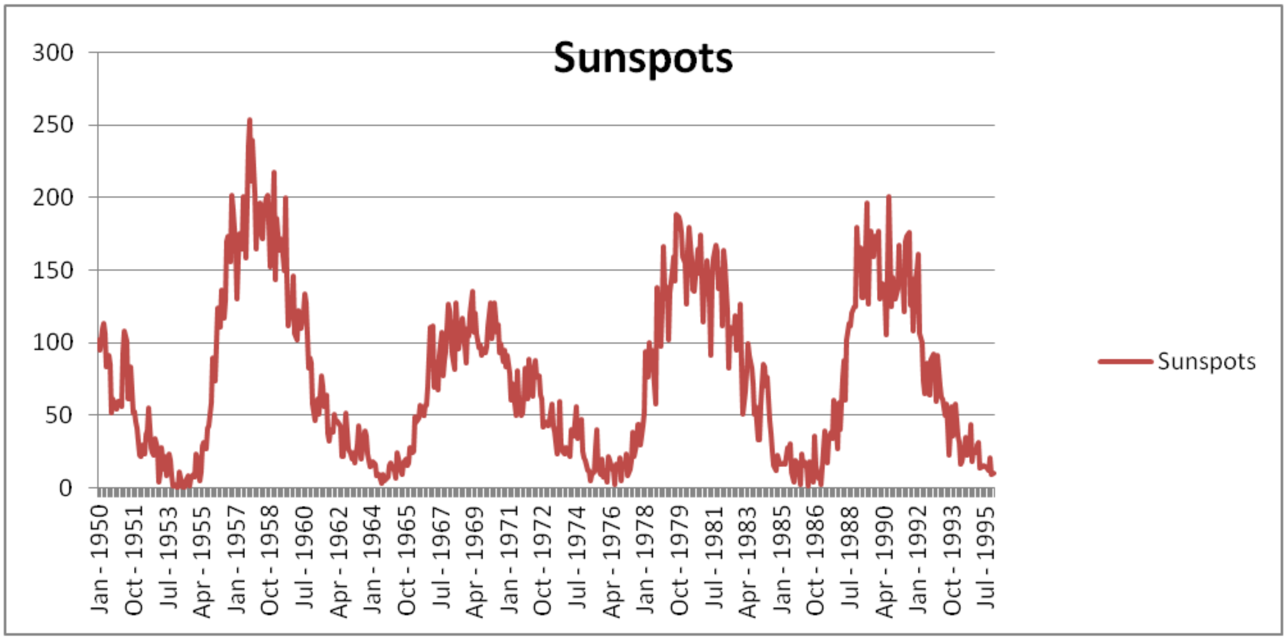
Sunspots are areas of the sun that appear darker in comparison to the rest of the sun's surface. Sunspots are closely tracked because they emit solar flares, which can be potentially harmful to the United State's power grid. Additionally, Solar cycles track very closely with temperature for much of this century, leading to the belief that temperature is more controlled by solar activity than CO$_{2.}$

Data

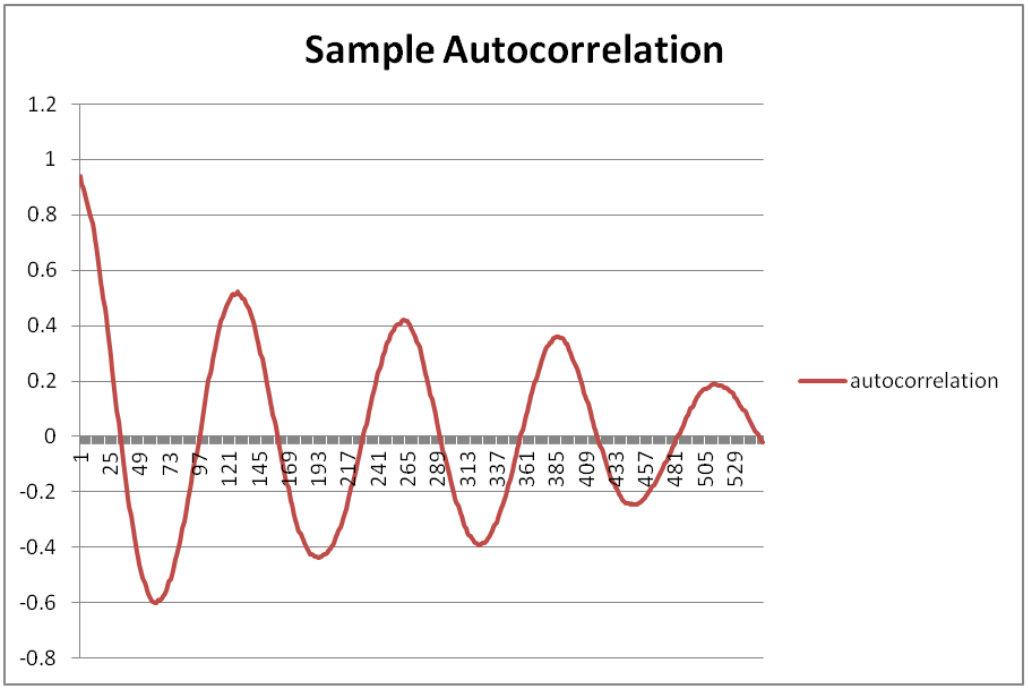
For this analysis, I will use data fromhttp://solarscience.msfc.nasa.gov/greenwch/spot_num.txt . This data gives monthly Suns back to the mid 18th century, shown below.



For the purpose of this exercise, I will take data beginning in 1950, below. The model will be completed on data through 1995, with the remaining points used as a "hold out" data set to test our model.

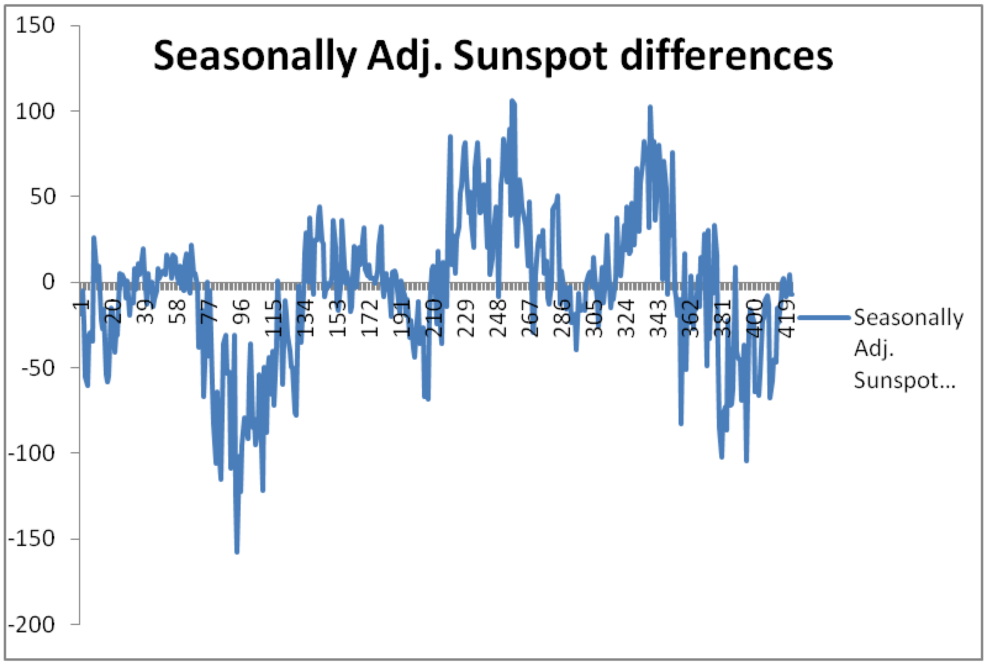


Looking at this time series, there is strong suggestion that the number of sunspots in a given month is not a stationary process. Also it is clear that there is seasonality in the data. We can confirm this by looking at the Sample Autocorrelation function, shown below:

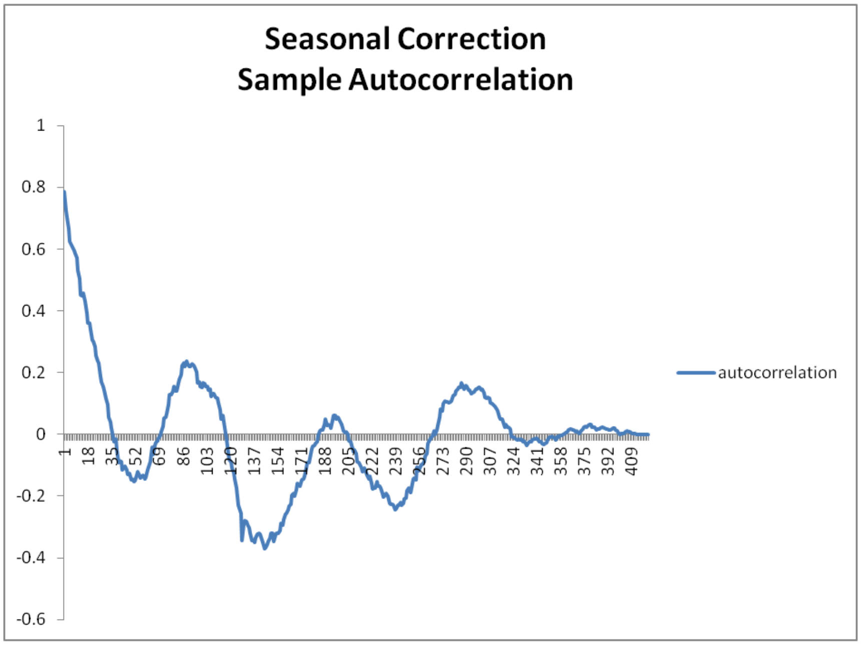


The sample autocorrelation function was developed using Excel. Per Cryer and Chan 6.1.1, the sample autocorrelation at lag k is given by the sample covariance at k divided by the sample variance. By looking at the peaks of the correlogram, we can see that there is approximately 130 month lag in the seasonality of the data.

After correcting for 130 month seasonality in the data, where S_t=Y_t-Y_{t-130} we get the below chart:

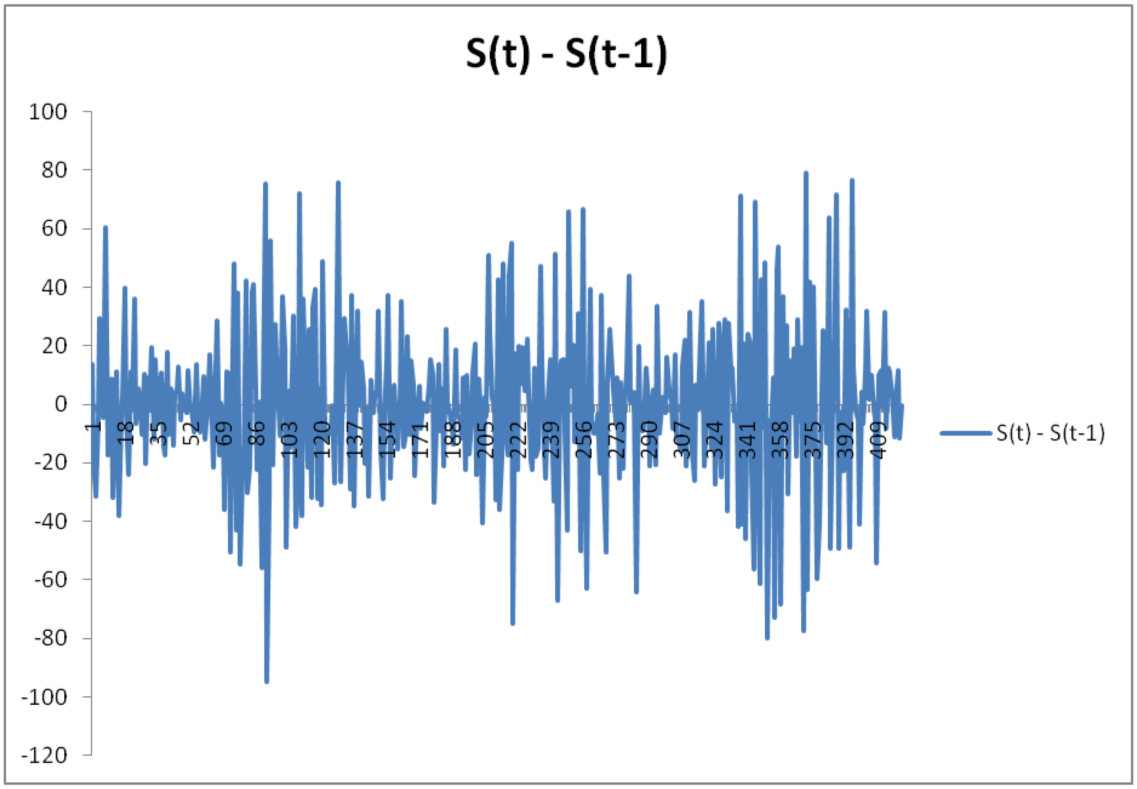


Given the long runs above and below 0, it still appears that this is a Non-Stationary process. To confirm, we once again look at the Correlogram. The large diversion from 0 confirms that further transformation is necessary.

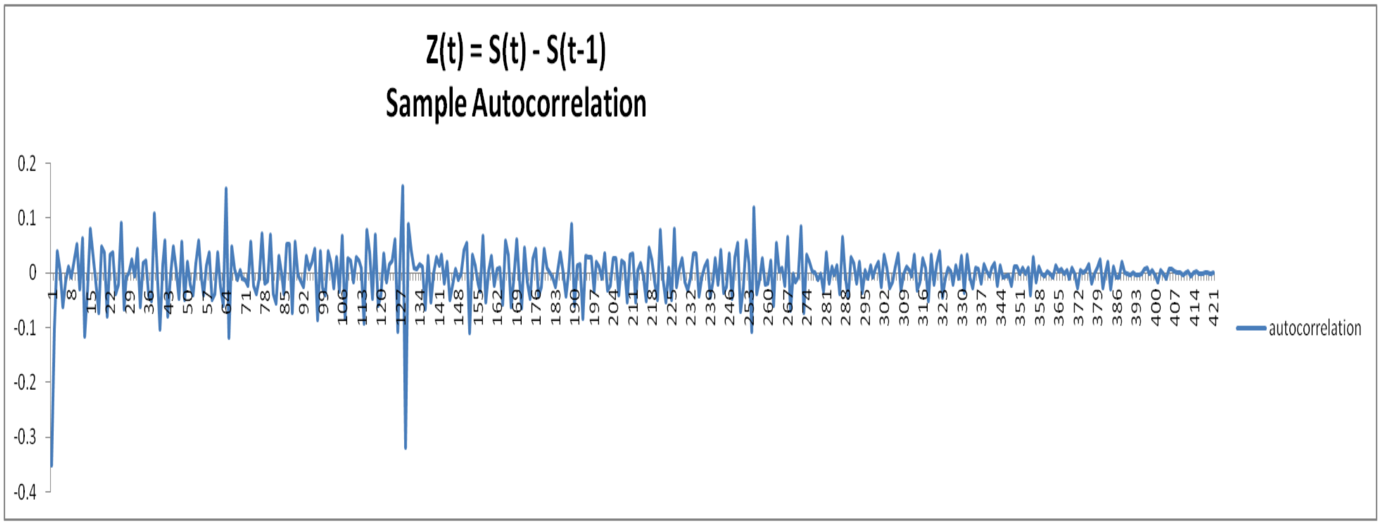


Transformation

We now have process $S_t = Y_t - Y_{t-130}$. Because this is still not a stationary process, I will take the first difference of $Z_t = S_t - S_{t-1}$, which does appear to be stationary.



To confirm that the series is stationary, we can once again check the Correlogram:



Now that this transformed process appears stationary, we can model this with an AR process.

Model Selection

An AR process takes the form of:

AR(1): $Z_t = a + B_1 Z_{t-1} + \varepsilon_t$

AR(2): $Z_t = a + B_1 Z_{t-1} + B_2 Z_{t-2} + \varepsilon_t$

AR(N): $Z_t = a + B_1 Z_{t-1} + B_2 Z_{t-2} \ldots + \ldots B_N Z_{t-N} + \varepsilon_t$

Using Excel's Regression tool, the equations for the AR processes are as follows:

An AR process takes the form of:

AR(1): $Z_t = .005 + -.351 Z_{t-1} + \varepsilon_t$

AR(2): $Z_t = .053 + -.256 Z_{t-1} + -.441 Z_{t-2} + \varepsilon_t$

AR(3): $Z_t = .145 + -.116 Z_{t-1} + -.306 Z_{t-2} + -.472 Z_{t-3} + \varepsilon_t$

The results can be summarized in a table:

Summary Table

	$\Sigma\phi_i$	Standard Error	R^2	Adj. R^2
AR(1)	-0.351	25.96	0.123	0.12
AR(2)	-0.697	25.15	0.181	0.177
AR(3)	-0.894	24.97	0.193	0.187

In examining this summary table, the R^2 and Adj. R^2 are very low. Because of this, I am likely not working with a stationary process and I will re-examine the transformation step.

Transformation Take II

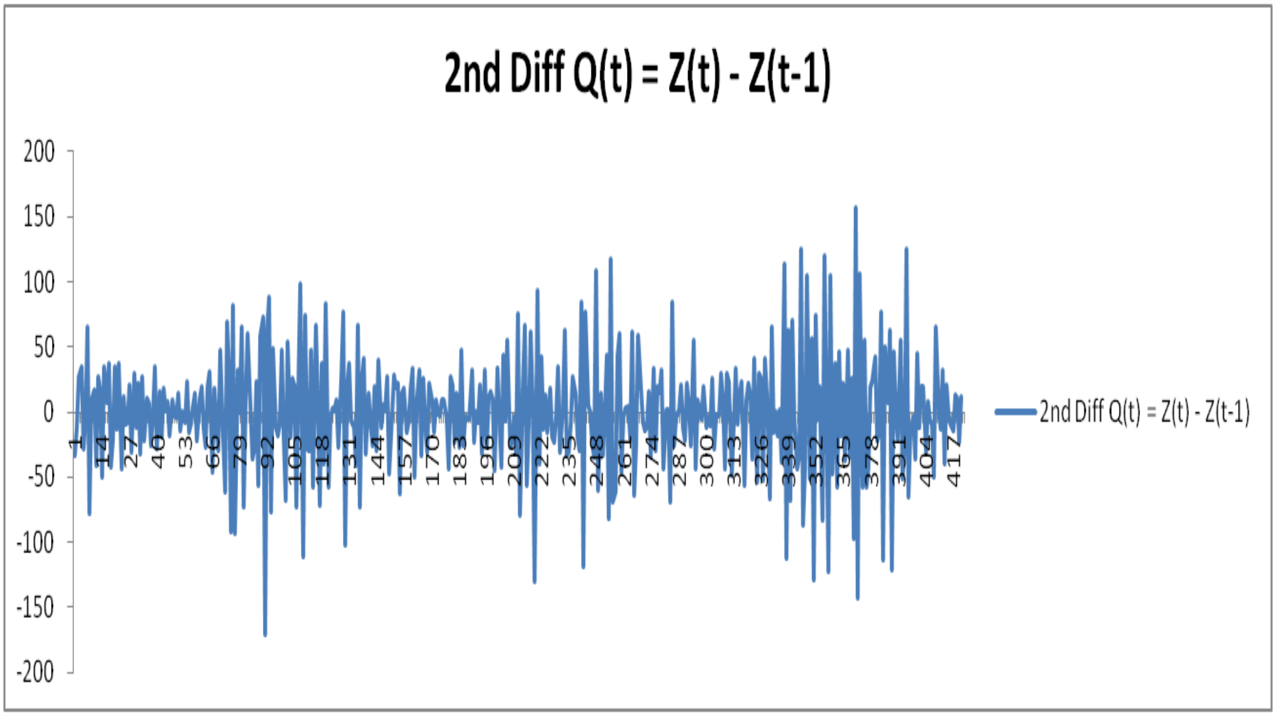
Rather than using First differences, I will now use the second difference of the Seasonally adjusted process. Thus far, we have the following transformations:

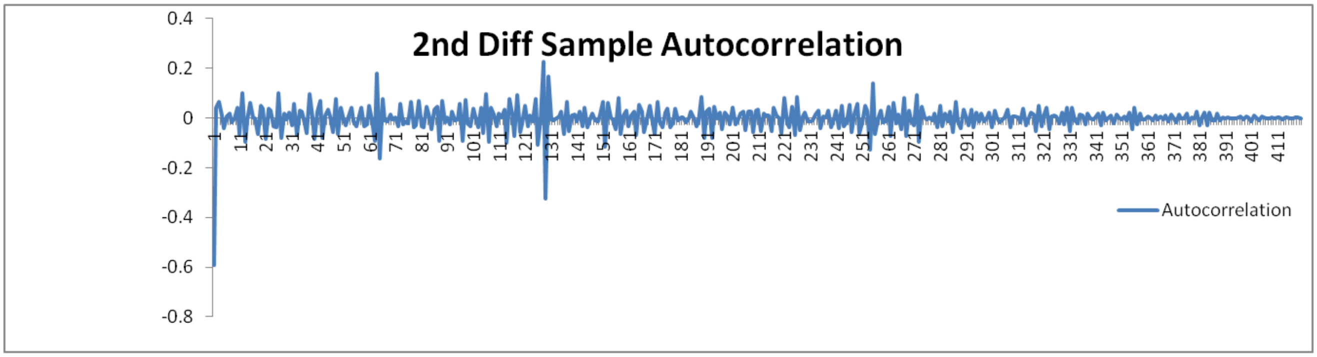
$S_t = Y_t - Y_{t-13}$ $\rightarrow$Seasonal Adjustment

$Z_t = S_t - S_{t-1}$ $\rightarrow$First Difference

$Q_t = Z_t - Z_{t-1}$ $\rightarrow$Second Difference

The charts for the data points Q(t) and Sample Autocorrelation function are as follows:





From the above charts, we see no improvement over the 1st difference transformation that was attempted earlier. Nonetheless, we will again try to model the time series using various AR(N) processes, as described above. Due to Parsimony, An AR process with N > 3 is likely more complex than necessary. After fitting using the Excel Regression tool, We get the following 3 equations:

AR(1): $Q_t = .009 + -.592Q_{t-1} + \varepsilon_t$

AR(2): $Q_t = .087 + -.478Q_{t-1} + -.877Q_{t-2} + \varepsilon_t$

AR(3): $Q_t = .108 + -.367Q_{t-1} + -.801Q_{t-2} + -1.053Q_{t-3} + \varepsilon_t$

The results can be summarized in a table:

Summary Table

	$\Sigma\phi_t$	Standard Error	R^2	Adj. R^2
AR(1)	-0.593	36.75	.351	.350
AR(2)	-1.355	32.30	.501	.498
AR(3)	-2.22	30.11	.568	.565

Durbin-Watson Test

A Durbin-Watson statistic of 2 indicates no serial correlation. A Durbin-Watson <2 Indicates a remaining positive correlation between the residuals. A Durbin-Watson >2 indicated there remains a negative correlation between the residuals. The following are the results for the three regressions

	DWS
AR(1)	2.567
AR(2)	2.352
AR(3)	2.181

With all Durbin-Watson statistics >2, we know that the residuals for each model have a strong negative correlation with the previous residual. Ideally we would have a DWS = 2. It is possible that we have over-differenced our model and should have stayed with the 1st difference. Nonetheless, we will continue working with the 2nd difference model.

Box-Pierce Q Statistic

The Box-Pierce statistic is used to test the following using a χ^2 distribution:

H_0 = Residuals are from a white noise process

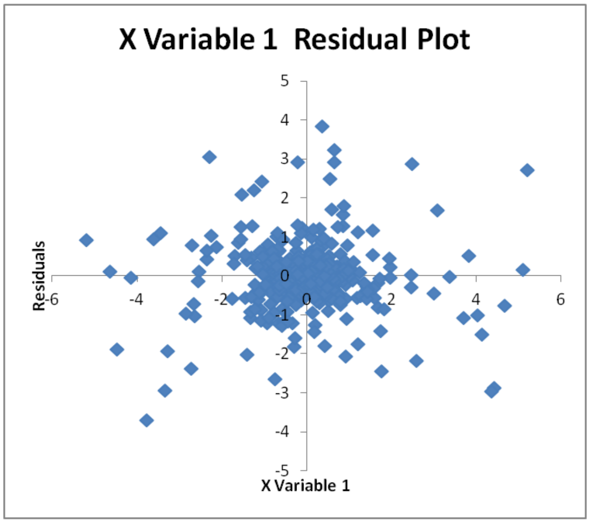
H_1 =Residuals are ***not*** from a white noise process

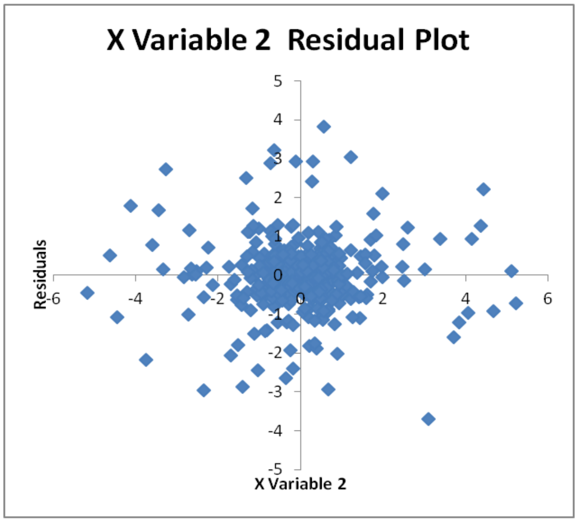
Because the AR(1) model exceeds the critical value at the 10% significance for 414 degrees of freedom, which is ~451, we reject H_0 ; that the residuals are white noise. Both the AR(2) and AR(3) models may have residuals that are white noise.

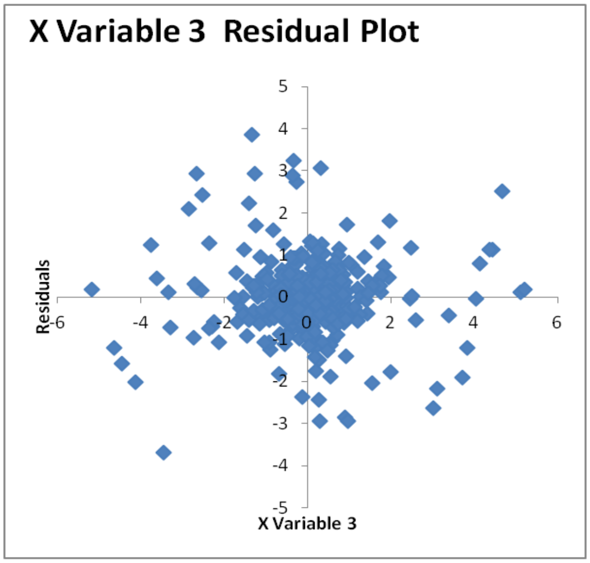
The Following chart shows the respective B-P Q statistics.

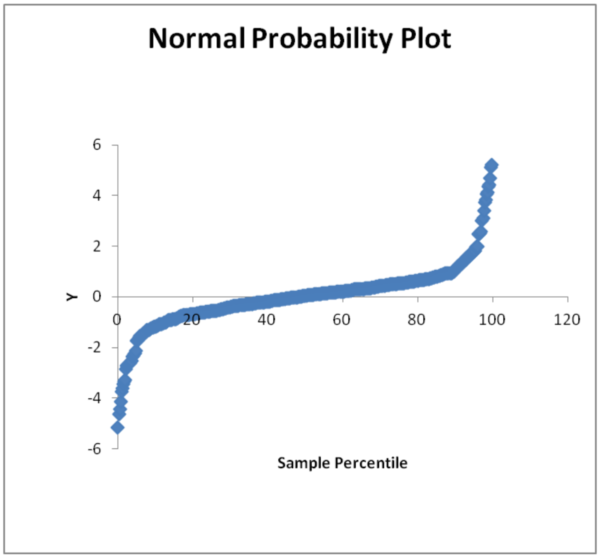
	BPQS
AR(1)	523.3
AR(2)	346.4
AR(3)	328.3

The Durbin Watson test suggests that we may not have a white noise process. Looking at the Box Pierce Q statistic, we are not able to reject a white noise process as a possibility at 10% significance for our AR(2) and AR(3) process. Furthermore, $\Sigma\phi_\iota < -1$ which indicate we may not have a stationary process. Nonetheless, Of the 3 models of 2nd differences above, I will select the AR(3) model because it performs marginally better in our Durbin-Watson tests and B-P Q test, but performs much better in the overall regression analysis. the following are the residual graphs and Normal Plot for our AR(3) process. The residuals scattered around the origin for the residuals indicate a fairly random process. The Normal plot indicates that our distribution performs much worse in the tails.


X Variable 1 Residual Plot


X Variable 2 Residual Plot


X Variable 3 Residual Plot


Normal Probability Plot

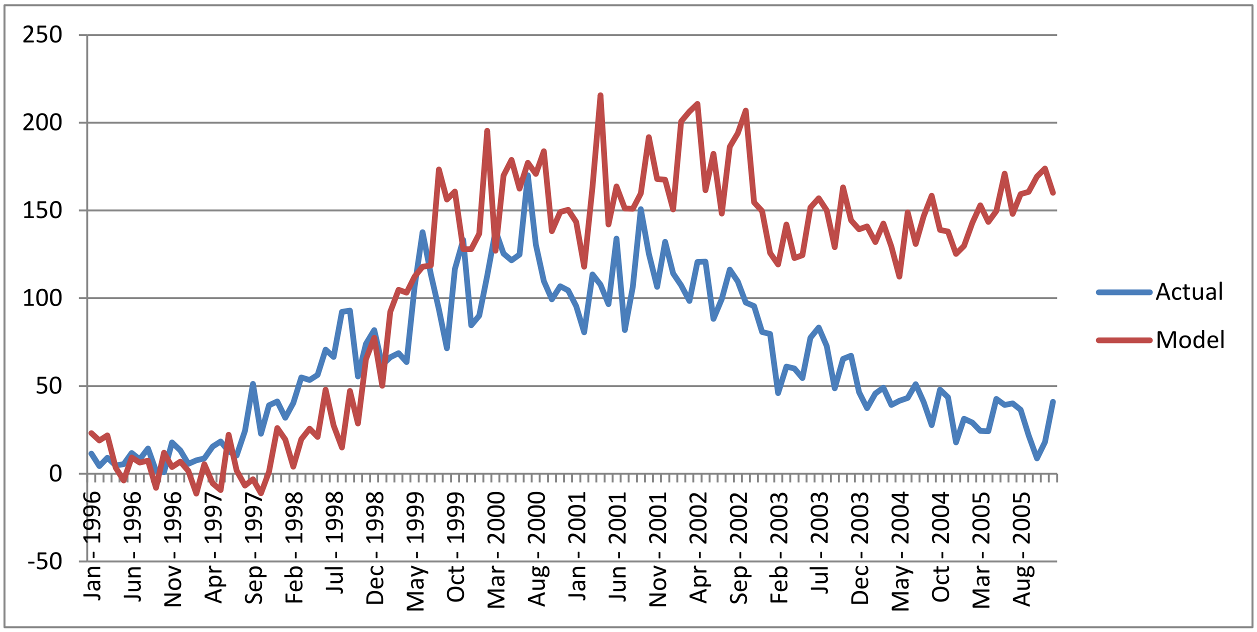
Forecasting

In order to test a model, it is necessary to test forecast a “hold out” data set. We compare our forecast to actual results to determine if we have chosen an appropriate model.

Our selected model is

AR(3): $Q_t = .108 + -.367Q_{t-1} + -.801Q_{t-2} + -1.053Q_{t-3} + \varepsilon_t$

The below graph shows the predicted values vs. the actual values of this time series. It is clear from the graph that our model is non-stationary. It is possible that further seasonality existed or that there was a MA component that was missed in my analysis. Further out, the model begins diverging exponentially from the actual value.


Jan - 1996
Jun - 1996
Nov - 1996
Apr - 1997
Sep - 1997
Feb - 1998
Jul - 1998
Dec - 1998
May - 1999
Oct - 1999
Mar - 2000
Aug - 2000
Jan - 2001
Jun - 2001
Nov - 2001
Apr - 2002
Sep - 2002
Feb - 2003
Jul - 2003
Dec - 2003
May - 2004
Oct - 2004
Mar - 2005
Aug - 2005
Actual
Model