Peter Faber
Time Series Student Project

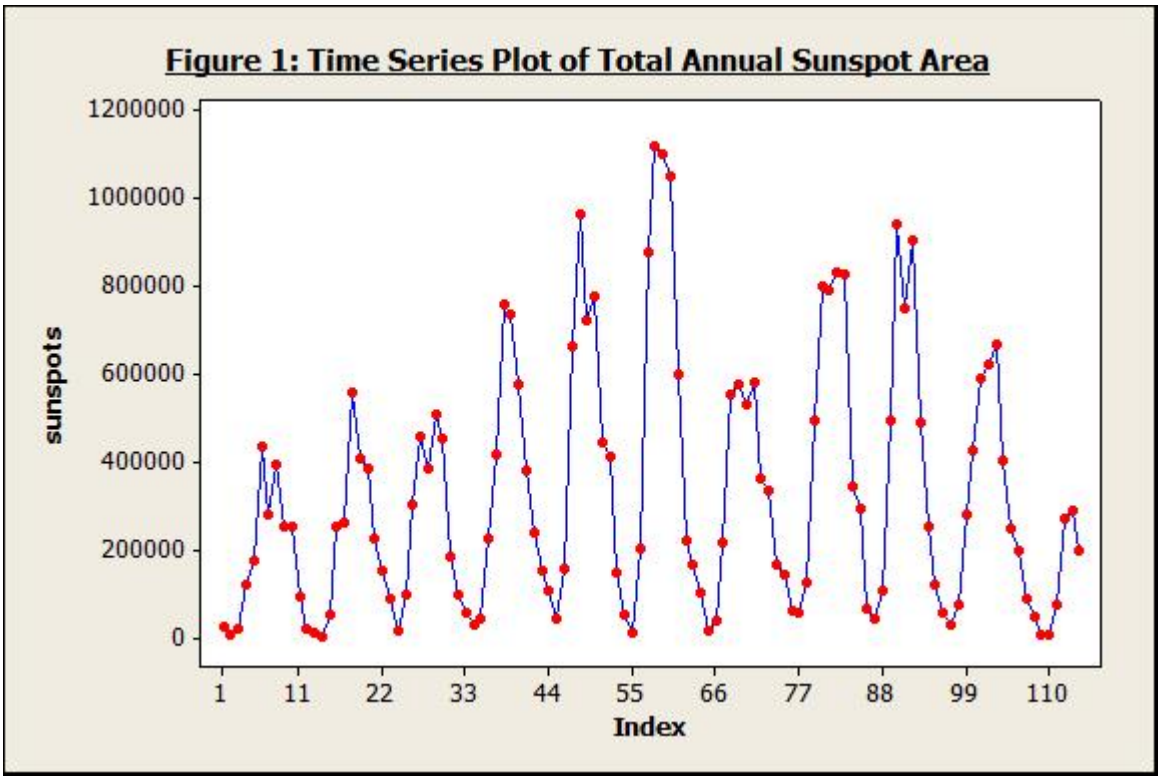
Goal

The goal of this analysis is to match total annual sunspot activity (measured by area of sunspots in millionths of a hemisphere) to the correct ARIMA model to predict future sunspot activity. Daily sunspot rates from 1900 to 2012 will be utilized to construct a model to predict 2000-2013 activity. Actual 2000-2013 sunspot activity will be used to validate the model.

Data

Data was obtained from the Royal Greenwich Observatory (RGO) and is maintained through the National Aeronautics and Space Administration's Solar Physics at the Marshal Space Flight Center's website: http://solarscience.msfc.nasa.gov/greenwch.shtml. The data set used can be found by visiting the following site directly: http://solarscience.msfc.nasa.gov/greenwch/daily_area.txt. While funding for this database has been revoked, the database is still maintained. Data as of October 1, 2013 will be considered[1]. Throughout this analysis, sunspot activity will be measured in millionths of a hemisphere.

Figure 1 displays daily sunspot activity from 1900-2012. There is a clear seasonal cycle during this period.

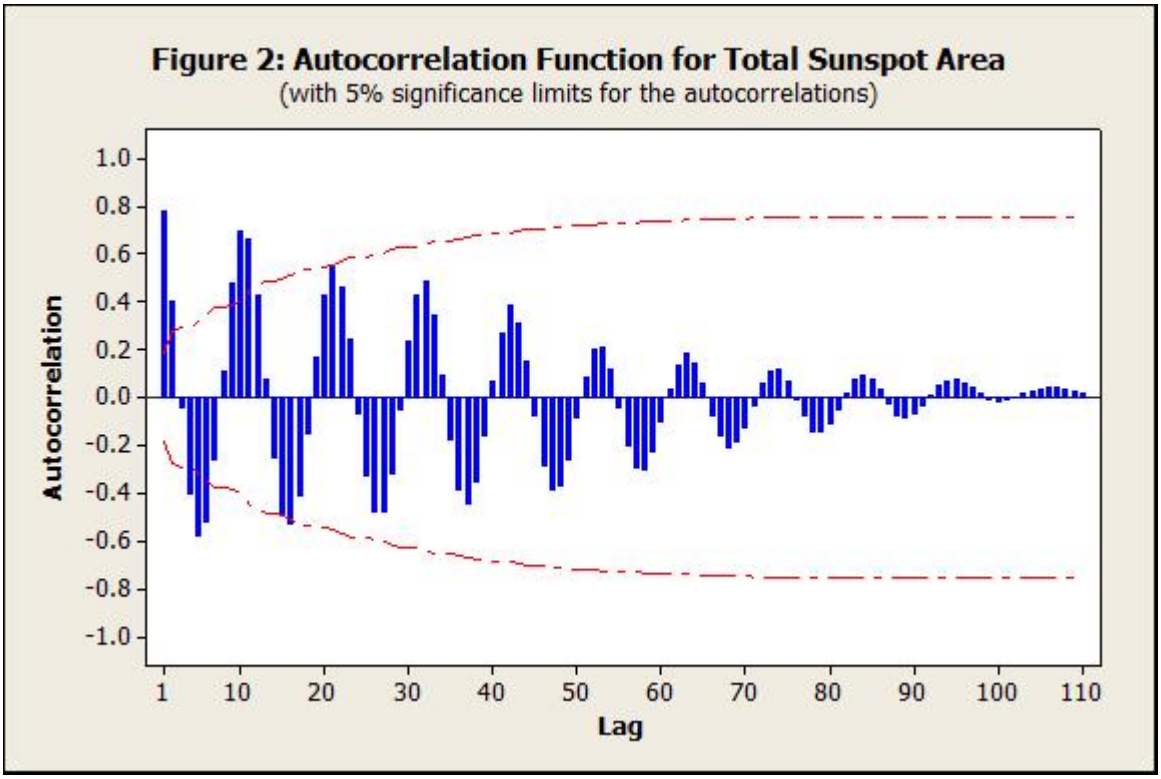


[1] The data used in my analysis is contained in the attached excel spreadsheet titled "Raw Sunspot Data.xlsx". Please note that this spreadsheet contains annual data even though the source material contained daily data. That daily data was very large and impractical to include due to file size considerations. The annual values were calculated by taking the average of the daily values.

Peter Faber
Time Series Student Project

Stationarity Test

In order to apply an ARIMA model, the data in question must be stationary. Sunspots are areas on the sun that experience such strong magnetic activity that the temperature of the gas in that area drops and appears dark to the observers. Even though the rational is not yet entirely understood, sunspots are known to by cyclical. The general shape of the time series suggests that the data in question has cyclical highs and lows and is therefore not stationary. To quantify the seasonality of the data, we consider the autocorrelation plot of the time series plot of daily sunspot activity in Figure 2.

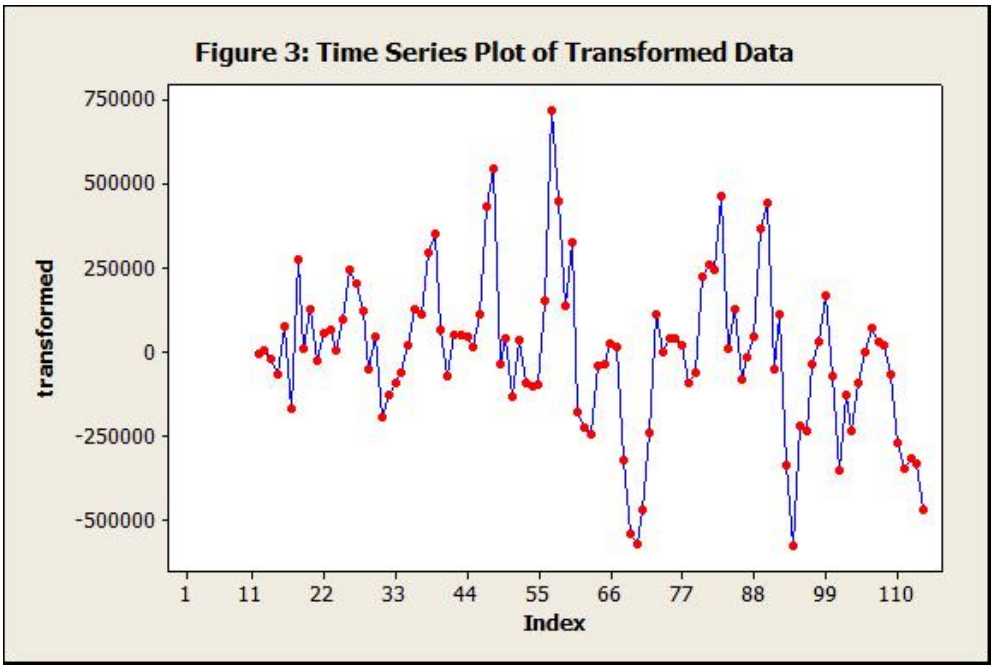


Since the autocorrelation function does not decline quickly and remain near zero, we see that the time series is not stationary. Figure 2 suggests that peaks occur around Lag 1, 11, 22, 33, 44 and again about every 11 years. This implies a period of seasonality of about 11 years.

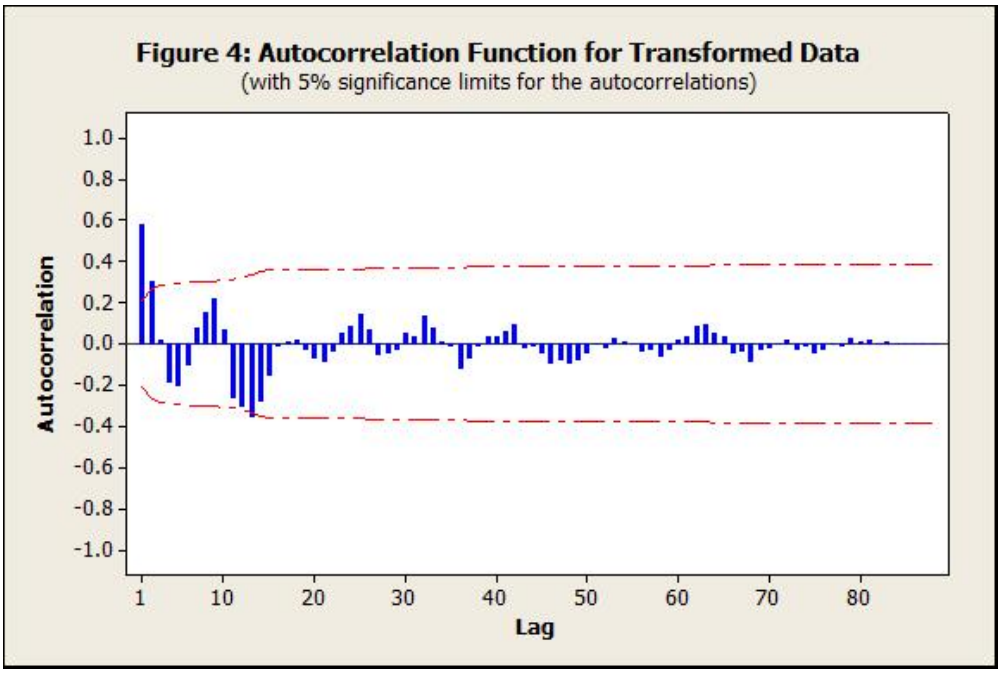
Peter Faber
Time Series Student Project

Seasonal Adjustment

In order to predict future sunspot activity, we must adjust our model and remove the seasonality. By transforming the time series by taking the first order differences between the periods, we see Figure 3 below.



We consider the autocorrelation of the transformed data in Figure 4.



The transformed time series shown in Figures 3 and 4 is clearly stationary. The autocorrelation quickly goes to 0 and remains there as the time series goes on.

Model Estimation

Since the transformed model is stationary, we can now develop a model with may contain autoregressive and/or moving average components.

Since the autocorrelation function increases and decreases in a sinusoidal manner as lag increases, we know that the time series cannot be described via a first order autoregressive model. The model would decrease in one direction if it was first order autoregressive. The model must be at least second order or contain moving average components.

Using Minitab (see attached file), we consider increasing order autoregressive and moving average models until an acceptable model is found.

AR(2)

```
Final Estimates of Parameters

Type       Coef  SE Coef      T      P
AR   1   0.6047   0.1077   5.62  0.000
AR   2  -0.0398   0.1080  -0.37  0.713

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   29.8   39.0   56.8   65.0
DF             10     22     34     46
P-Value     0.001  0.014  0.008  0.034
```

The p-value for the second order term suggests that this term is not statistically significant.[2] This model should be rejected and we include a moving average term to see if first order white noise provides more accurate coefficients for the ARMA model.

ARMA(2,1)

```
Final Estimates of Parameters

Type       Coef  SE Coef      T      P
AR   1   0.0016   2.8475   0.00  1.000
AR   2   0.3522   1.6235   0.22  0.829
MA   1  -0.5669   2.8777  -0.20  0.844

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   32.5   42.0   59.4   67.3
DF              9     21     33     45
P-Value     0.000  0.004  0.003  0.017
```

[2] Note: All statistical tests are performed at a 5% significance level.

The p-values for all three coefficients suggest that none of the terms are statistically significant. The model should not be considered. Since the autoregressive graph of the transformed time series does not cross zero until lag 3, it is likely that the ARMA function be of at least order 3.

AR(3)

```
Final Estimates of Parameters

Type       Coef  SE Coef      T      P
AR   1   0.5954   0.1059   5.62  0.000
AR   2   0.0913   0.1245   0.73  0.466
AR   3  -0.2141   0.1063  -2.01  0.047

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   25.6   31.6   49.7   60.3
DF              9     21     33     45
P-Value     0.002  0.064  0.031  0.063
```

The p-values for the coefficients of the first and third order autoregressive term suggest that they are statistically significant. However, since the p-value for the second order term is greater than 0.05, we next consider an ARMA model which includes three autoregressive terms as well as a moving average term.

ARMA(3,1)

```
Final Estimates of Parameters

Type       Coef  SE Coef        T      P
AR   1  -0.3725   0.1089    -3.42  0.001
AR   2   0.5470   0.0985     5.56  0.000
AR   3  -0.0678   0.1107    -0.61  0.542
MA   1  -0.9716   0.0088  -110.07  0.000

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   30.1   38.8   58.0   66.8
DF              8     20     32     44
P-Value     0.000  0.007  0.003  0.015
```

We see that all but one of the moving average and first and second autoregressive terms are statistically significant. We consider a second order moving average term.

ARMA(3,2)

```
Final Estimates of Parameters

Type       Coef  SE Coef       T      P
AR   1   1.7955   0.1295   13.86  0.000
AR   2  -1.5374   0.1843   -8.34  0.000
AR   3   0.3915   0.1203    3.25  0.002
MA   1   1.2614   0.0607   20.78  0.000
MA   2  -0.9451   0.0437  -21.62  0.000

Modified Box-Pierce (Ljung-Box) Chi-Square statistic
```

Lag	12	24	36	48
Chi-Square	17.5	24.8	40.0	48.4
DF	7	19	31	43
P-Value	0.015	0.166	0.130	0.263

We see that all five coefficients have p-values below 0.05 and are therefore statistically significant. Since all coefficients are statistically significant, this is a good model to consider for further validation tests.

We want to confirm that the residuals present in this model are random and not a function of the model.

First, since the p-values under the modified Box-Pierce Chi-Square test under 24, 36 and 48 lags is larger than 0.05, the residuals are similar to the residuals from a random selection of data and are therefore random.

Next, we consider whether or not the residuals are normally distributed by plotting them next to the normal distribution.

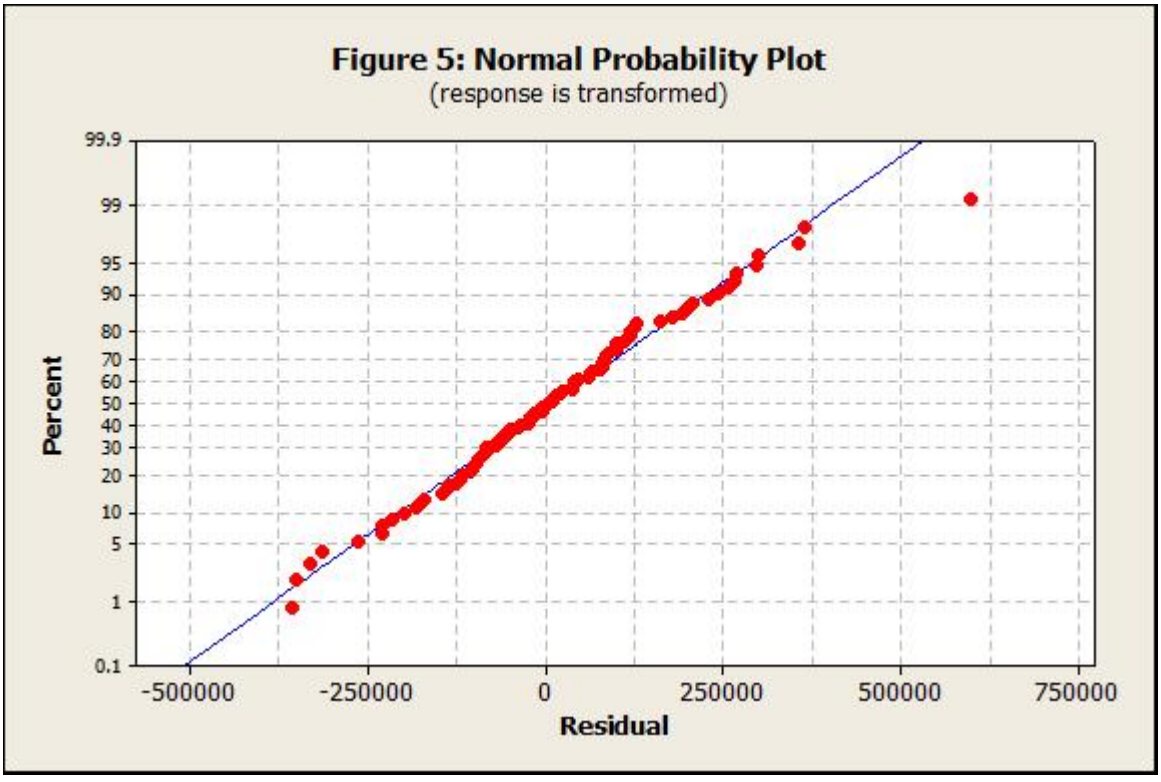


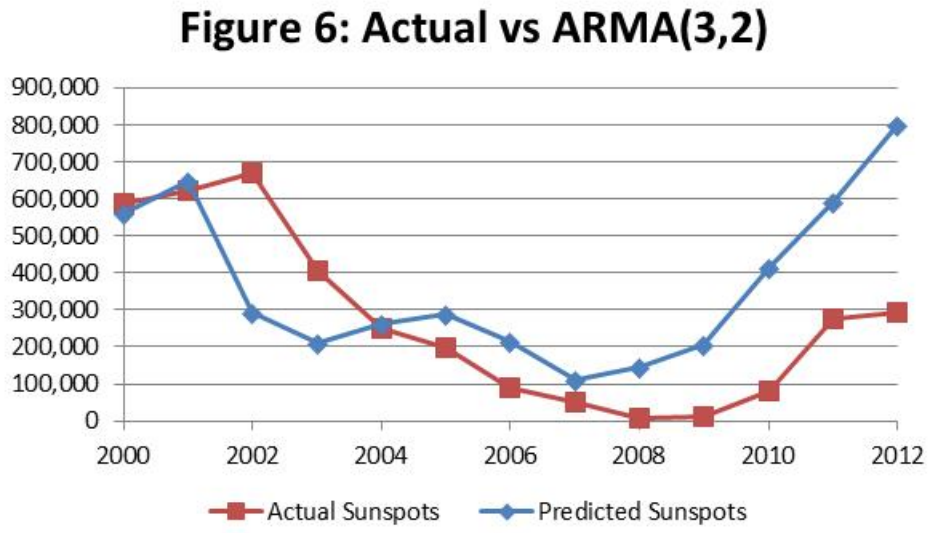
Figure 5 shows that the residuals very closely follow the normal distribution.

Since the model has low p-values for its coefficients, passes the Modified Box-Pierce Chi Square Statistic Test, and has residuals which are very close to the normal distribution, we accept the model. Its coefficients are statistically significant; its residuals are random and normally distributed.

Peter Faber
Time Series Student Project

Model Evaluation and Conclusion

We now predict total annual sunspot activity for 2000-2012 using the model.



The source data for Figure 6 was obtained from Minitab and the graph was created using the attached Excel file.

The graph suggests that our model more accurately predicts the general trend of sunspot activity rather than the actual number of sunspots which will occur in a given year. Sunspot activity is predicted to drop similarly to actual events but specific year measurements are off by several hundred millionths of a hemisphere.

Sunspot activity, while complicated, is based on magnetic fields present in the Sun. While much of the thermodynamics describing the Sun's physics are known, there are any other aspects which are not. The time series developed here provides one method of prediciting sunspot activity but it may not be the best. Perhaps a more accurate prediction tool would come from using our understanding of the Sun's thermodynamics to model solar particles using the always increasing computer power present today.