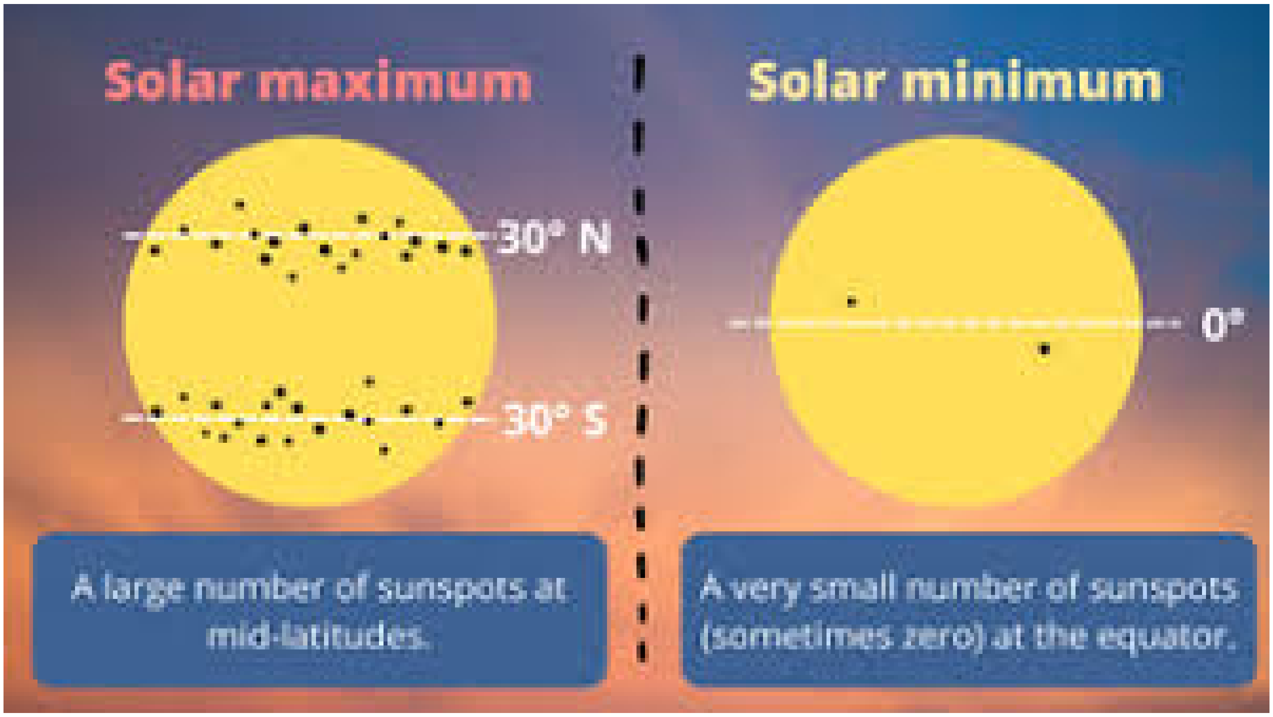
Time Series Analysis of the Sunspot Cycle using the ARIMA and SARIMA Models

Insights and predictions of the count and density of sunspots. This will act as the data input to feed into Acme Telecommunications' proprietary model attempting to more accurately predict solar flares, which pose a significant financial risk to their fleet of satellites.


Solar maximum
30° N
30° S
A large number of sunspots at mid-latitudes.
Solar minimum
0°
A very small number of sunspots (sometimes zero) at the equator.

Max Sussman

04.12.2024

Acme Telecommunications
Satellite Safety Division

INTRODUCTION

Sunspots are temporary phenomena on the Sun's photosphere that appear as spots darker than the surrounding areas. They are regions of reduced surface temperature caused by concentrations of magnetic field flux that inhibit convection. Sunspots usually appear in pairs of opposite magnetic polarity. Their number varies according to the approximately 11-year solar cycle. These spots are visible from the sun's surface or "photosphere" and are known to appear as spots which are darker than surrounding areas of the sun. The layers of plasma from within and on the surface of the sun churning against one another at different rates (differential rotation) generates the sun's very powerful magnetic field. This is the driving cause of not only sunspots, but plasma loops and solar flares as well. When a strong magnetic field disrupts the flow of plasma, sunspots begin to appear, which is typically the signal that a coronal mass ejection is soon approaching.

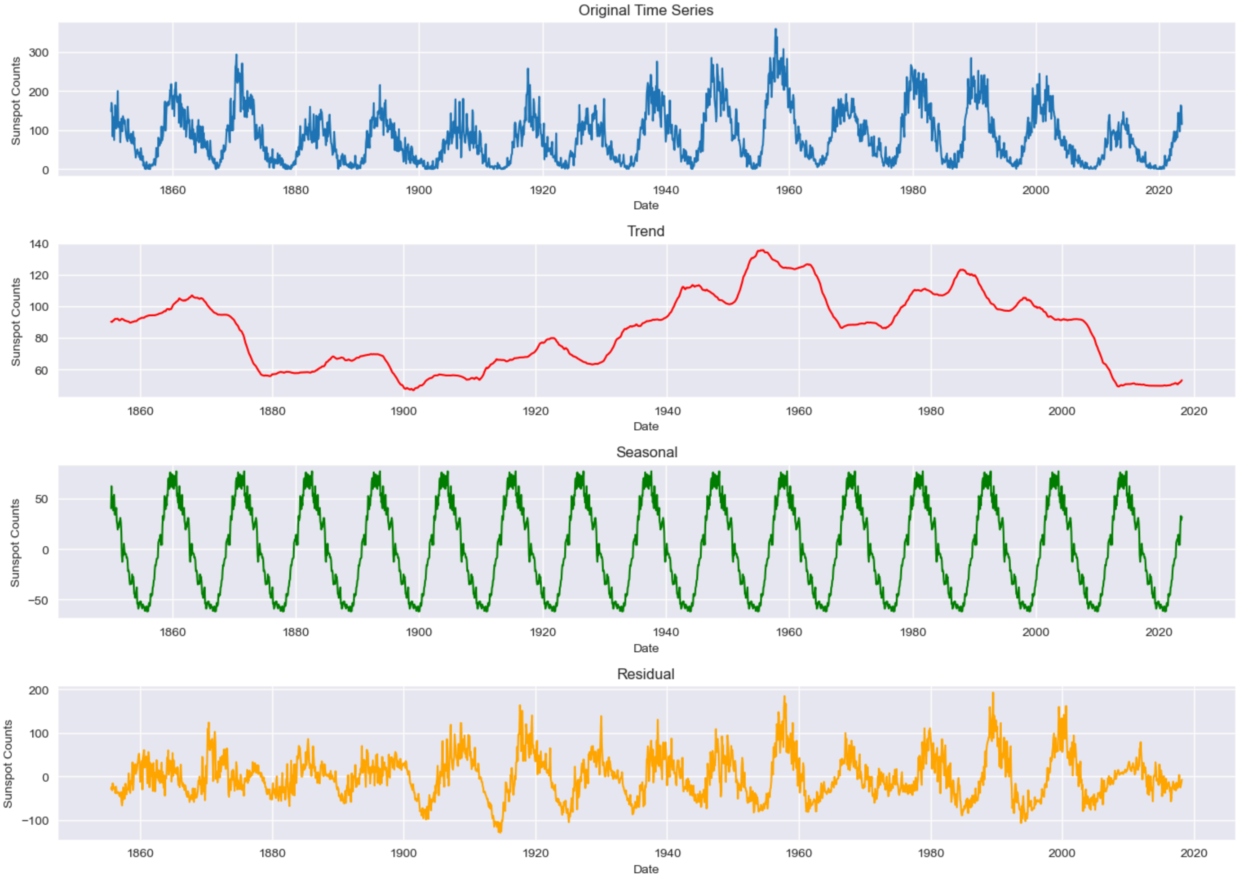
I am using the Daily Sunspot Dataset from the World Data Center SILSO, Royal Observatory of Belgium, Brussels. The Daily total sunspot number is derived by the formula: R = Ns + 10 * Ng, with Ns the number of spots and Ng the number of groups counted over the entire solar disk. The original data contained some values between 1818 and 1850 but I removed those years because there were too many missing values. The standard error of the daily Sunspot count can be computed by: sigma/sqrt(N) where sigma is the listed standard deviation and N the number of observations for the day. Before 1981, the number of observations is set to 1, as the Sunspot Number was then essentially the raw Wolf number from the Zürich Observatory. The following are the column descriptions of the dataset:

- Column 1-3: Gregorian calendar date
  - Year
  - Month
  - Day
- Column 4: Date in fraction of year
- Column 5: Daily total sunspot number. A value of -1 indicates that no number is available for that day (missing value): there should not be any missing values as I removed all years before 1850 where there was missing values.
- Column 6: Daily standard deviation of the input sunspot numbers from individual stations.

- Column 7: Number of observations used to compute the daily value.
- Column 8: Definitive/provisional indicator. A blank (NaN) indicates that the value is definitive. A '*' symbol indicates that the value is still provisional and is subject to a possible revision (Usually the last 3 to 6 months).

EDA

After cleaning up the dataset and graphing the original time series I performed a decomposition function to separate the trend, seasonal and residual components and graph them respectively. I then added a time column and replaced any NaN by linear interpolation. I then converted the counts column to a series.



Data Preprocessing

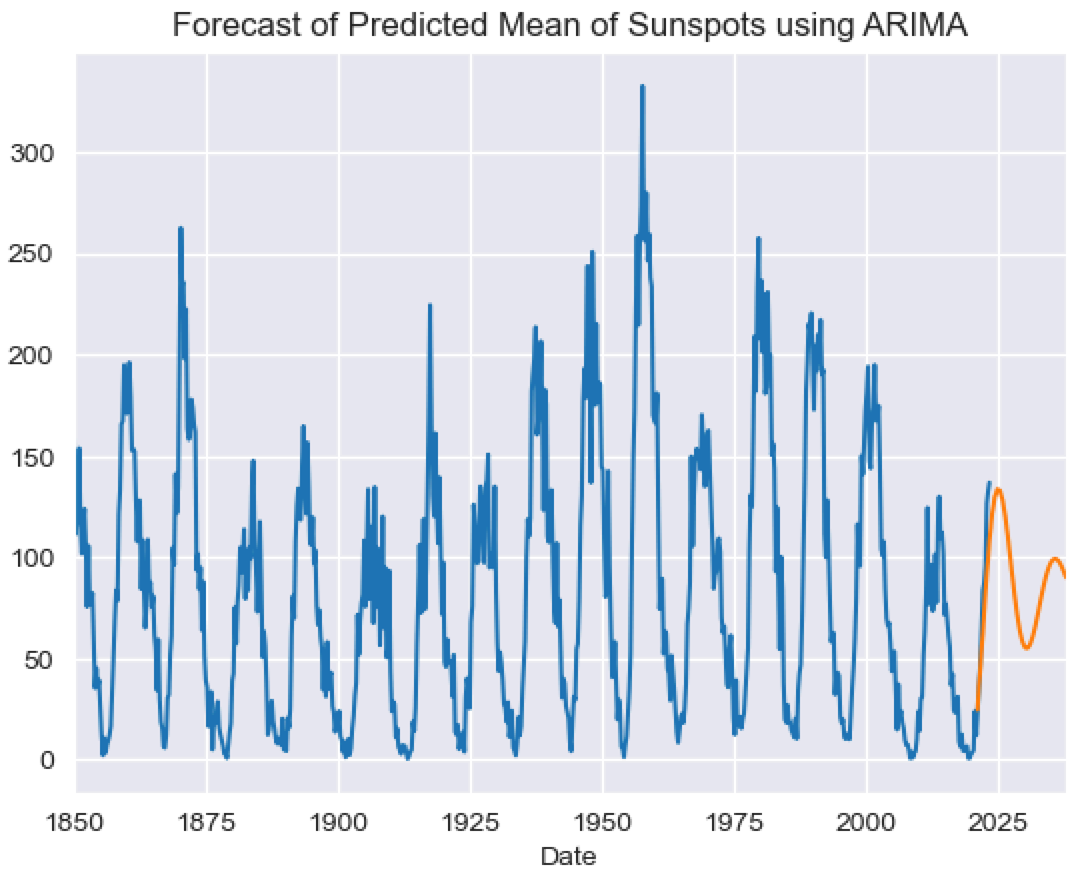
From the graphs of the decomposed time series I selected the values for (p,d,q) and (P, D, Q, S). From the ACF plot, the first high lag is around 43, so let's take S=43. D=1, the series has a stable seasonal pattern over time. PACF plot : p = first lag where the value is above the significance level. p=3. ACF plot : q = first lag where the value is above the significance level. q=10. ACF plot : P=1, ACF is positive at lag 43 AND P+Q≤2. ACF plot : Q=0 ACF is negative at lag 43 AND P+Q≤2. Then performed a Dickey-Fuller Test to check for stationarity. Here the null hypothesis is that the TS is non-stationary. The test results consist of a Test Statistic and some Critical Values for difference confidence levels. If the 'Test Statistic' is less than the 'Critical Value', we can reject the null hypothesis and say that the series is stationary.

Modeling

I tried out the ARIMA and SARIMA models to best make future predictions for the sunspots. The following is the results from the ARIMA model:

```
Dep. Variable:            counts   No. Observations:            695
Model:              ARIMA(3, 0, 10)   Log Likelihood          -3099.604
Covariance Type:               opg
==============================================================
              coef      std err     z        P>|z|     [0.025     0.975]
--------------------------------------------------------------
const      83.3876     5.999    13.901     0.000     71.631     95.144
ar.L1       1.0039     0.066    15.209     0.000      0.874      1.133
ar.L2       0.8298     0.124     6.689     0.000      0.587      1.073
ar.L3      -0.8748     0.064   -13.569     0.000     -1.001     -0.748
ma.L1      -0.3146     0.070    -4.492     0.000     -0.452     -0.177
ma.L2      -0.9428     0.089   -10.589     0.000     -1.117     -0.768
ma.L3       0.4160     0.043     9.569     0.000      0.331      0.501
ma.L4       0.1558     0.052     2.972     0.003      0.053      0.259
ma.L5      -0.0245     0.046    -0.534     0.593     -0.114      0.065
ma.L6      -0.0533     0.052    -1.016     0.309     -0.156      0.049
ma.L7      -0.0473     0.050    -0.951     0.342     -0.145      0.050
ma.L8       0.1086     0.055     1.972     0.049      0.001      0.217
ma.L9       0.0417     0.039     1.055     0.292     -0.036      0.119
ma.L10     -0.0965     0.041    -2.359     0.018     -0.177     -0.016
sigma2    435.7232    20.471    21.285     0.000    395.601    475.846
==================================================================
Ljung-Box (L1) (Q):              0.02   Jarque-Bera (JB):         53.85
Prob(Q):                         0.89   Prob(JB):                  0.00
Heteroskedasticity (H):          0.97    Skew:                     0.34
Prob(H) (two-sided):             0.83   Kurtosis:                   4.18
===================================================================
```

Using this to forecast we end up with the following:



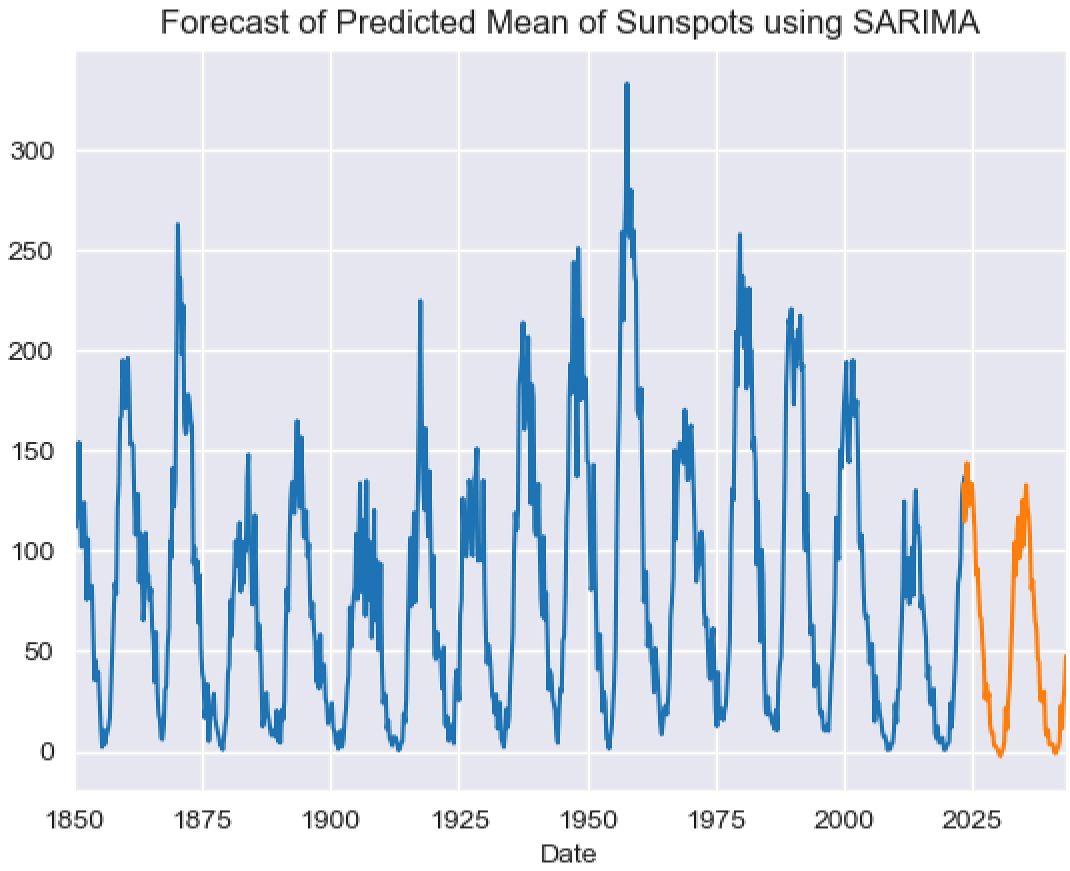
With the SARIMA model we end up with the following:

```
Dep. Variable:                   counts   No. Observations:               695
```

```
Dep. Variable:                   counts   No. Observations:               695
```

```
Model:           SARIMAX(3, 0, 10)x(1, 1, [], 43)   Log Likelihood            -3031.596

Covariance Type:                         opg

==============================================================================

              coef    std err          z      P>|z|      [0.025      0.975]

------------------------------------------------------------------------------

ar.L1         0.9031      0.121      7.472      0.000       0.666       1.140

ar.L2         0.7916      0.169      4.675      0.000       0.460       1.124

ar.L3        -0.7522      0.104     -7.206      0.000      -0.957      -0.548

ma.L1        -0.2946      0.124     -2.381      0.017      -0.537      -0.052

ma.L2        -0.9234      0.135     -6.832      0.000      -1.188      -0.658

ma.L3         0.4243      0.054      7.855      0.000       0.318       0.530

ma.L4         0.1840      0.061      3.039      0.002       0.065       0.303

ma.L5        -0.0305      0.061     -0.498      0.618      -0.150       0.089

ma.L6        -0.0927      0.053     -1.737      0.082      -0.197       0.012

ma.L7        -0.0536      0.052     -1.023      0.306      -0.156       0.049

ma.L8         0.1807      0.051      3.540      0.000       0.081       0.281

ma.L9         0.0499      0.050      0.994      0.320      -0.049       0.148

ma.L10       -0.1232      0.050     -2.456      0.014      -0.222      -0.025

ar.S.L43     -0.4548      0.031    -14.626      0.000      -0.516      -0.394

sigma2      626.2016     31.424     19.928      0.000     564.612     687.791

==============================================================================
====

Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):               17.14

Prob(Q):                              0.90   Prob(JB):                        0.00

Heteroskedasticity (H):               1.18   Skew:                 -0.02          Kurtosis:                      3.79

Prob(H) (two-sided):                  0.22   Kurtosis:             3.79b       (H) (two-sided):               0.22
==============================================================================
```

Using this to forecast our sunspots we end up with the following:



Conclusion:

As we can see from the results of our models against our test data as well as from the graphs the SARIMA model performs better than the ARIMA model. This information will be passed on to my client, Acme Telecommunications for use to better predict the occurrences of solar flares.