**Time Series Analysis Project**

**Predicting Sunspots Number**

**2202 STAT 4160 1/5160 1/MSA 5160 50 CS**

**Project Report**

**Presented By:**

**Venkata Sai Suneeth Ravilla**

**Submitted By:**

**Venkata Sai Suneeth Ravilla**

**Academic Supervisor:**

**Dr. John Chen**

**Spring 2020**

**Bowling Green State University, Ohio**

**Executive Overview:**

Time series is a sequence of data points, measured typically at successive points in time at uniform time intervals. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical, finance, weather forecasting, earthquake prediction, electroencephalography, control engineering and communications engineering etc. The sunspots dataset in this study consist of a monthly count of sunspots from 1749 to 1983 and there are 2,820 observations. The source of the dataset is credited to Andrews & Herzberg (1985)

**Visualizing time series:**

Visualizing is one of the key aspects to identify patterns in the data. Firstly, we start with plotting time series then from the plot, we try to identify patterns related to the series like trend and seasonality. In this report of sunspots, data have number of sunspots for different months are fluctuating across the years with no constant mean and variance and seasonality is also not observed even though sun spots has correlation with the solar magnetic cycle where every time sun finishes an approximate 11 year cycle the magnetic pole swap places.

**Stationary time series:**

Stationary time series define one whose statistical properties such as mean, variance, autocorrelation and other factors which are constant over time. A stationary series is relatively easy to predict we simply predict that its statistical properties will be the same in the future as they have been in the past. For example, in this report the sunspots data is fluctuating over time and this will always underestimate the mean and variance in future periods and if the mean and variance of a series are not well-defined, then neither are its correlations with other variables. So, using the augmented dickey fuller test non-stationary check is performed and it shows data is stationary. The seasonality effect and the trend if dictated is removed from time series through ordinary differencing of order one.

**Fitting model:**

Once we have our optimal model parameters, we can fit an ARIMA model to learn the pattern of the series. In this report the sunspotsdata is used, and the model fitted is ARIMA (2,0,1) (1,0,1) [11]

**Predictions:**

Time series methods are better suited for short-term forecasts. We analyze the dataset using timeseries technique and fit an appropriate timeseries model to predict the number of sunspots for next 10 years using R language. The future study can be implementing the other solar phenomena’s and predicting trend of them.

**Data Collection:**

Data is chosen from machinelearningmastery a site that has blog “7 Time Series Datasets for Machine Learning” published by Jason Brownlee. Sunspots dataset consists of monthly count of the number of sunspots observed for 210 years i.e. from 1749 to 1983. Total dataset has 2,820 observations. These sunspots are noted based on number of dark spot visible on the surface of the sun. This dark spot called the sunspot last from a few days to months before eventually decaying.

**Why this dataset:**

The number of sunspots is related to temperature, sun’s activities, the effect of other planets, etc. However, it has a strong relationship with time. In this way, the analysis for the number of sunspots and time is a good topic for time series analysis. Also, we can infer the activities of the sun, the temperature, and a lot of solar system events from the analysis result.

**Why sunspot variable:**

The commonest study about the solar activity is the prediction of sunspot number along with the solar cycle and various radiations. The sunspot can vary from 16 km to 160k km in diameter. The larger ones can be seen from the earth with our naked eye. Determining the sunspot cycle period is important to compare the period estimate with disruptions to radio and satellite communications and with weather cycles. There are strong indications that the cooling and warming of the earth might be due to the changes in the number of observed sunspots.

**Why interest to anyone:**

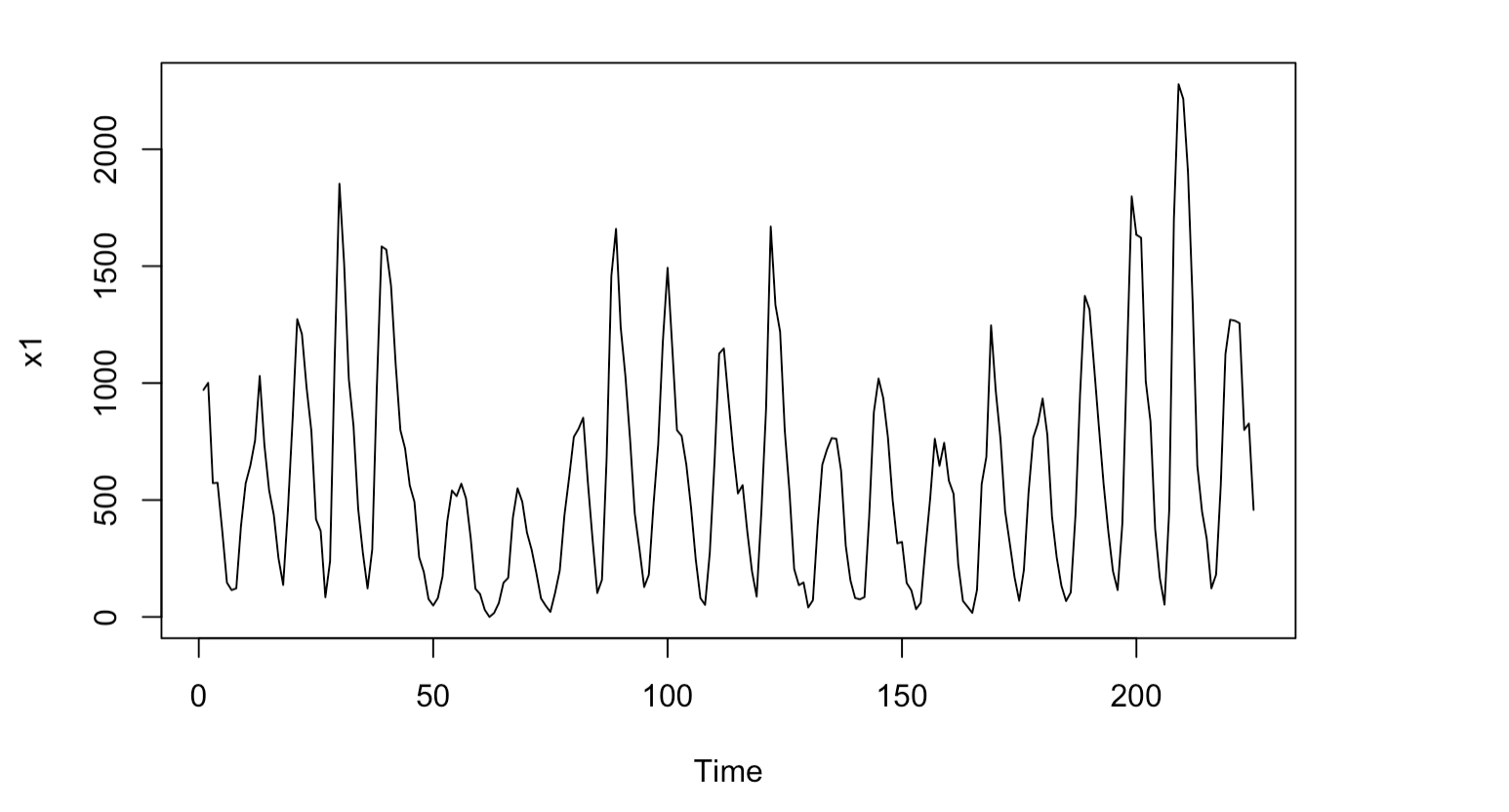
Understanding the entire solar phenomenon is very complex, yet it remains an important aspect of solar physics, as solar activity is closely associated with biosphere, space weather and the field of space technology. Prediction of sunspot number are also very important in planning space related activities particularly for low earth orbiting spacecraft.

**Analysis:**

Performed analysis on complete dataset first starting with the analysis on monthly sunspots data then followed with yearly data by aggregating the sum of the monthly data. But in presentation we analyzed both monthly and yearly data and, in this report, we only discuss about yearly data as it is performed by myself and one of my team members.

**Time-Series Data Modeling**

The dataset chosen has yearly details along with sunspots, plotting the data by converting the dataset to time series,



From the above time series plot we can say that, the number of sunspots over the time are fluctuating across the years without any constant mean and variance. By looking at the pattern we cannot draw much information from it as the data is huge and there is no specific trend following in it.

After observing the above characteristics from the plot, proceeding to further analysis to test whether data is stationary or not. By looking at the plot it seems to be stationary but checking with Augmented dickey fuller test we can confirm whether data is stationary or not.

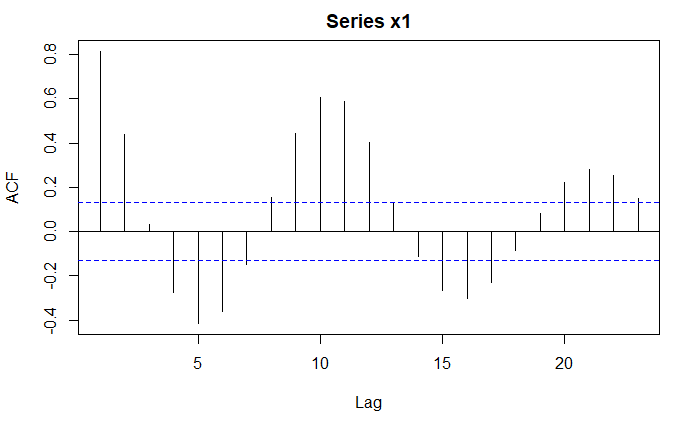
After performing the Dickey-Fuller test we observed the p value is below the significance level (5%).

* Reject the null hypothesis that the data is stationary
* Data is stationary at this point.

Now we confirmed that the data is stationary and can proceed for the further analysis and predictions.

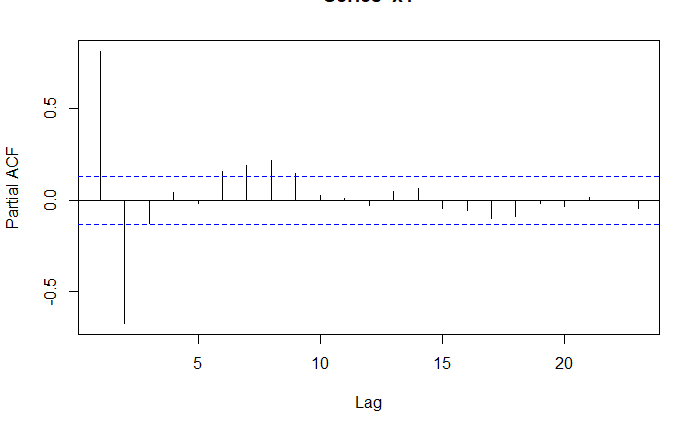
Plotting the ACF and PACF plots for understanding which pattern the data follows whether AR or MA to choose the best model by fitting multiple values.

ACF is an (complete) auto-correlation function which gives us values of autocorrelation of any series with its lagged values. In simple terms, it describes how well the present value of the series is related with its past values. A time series can have components like trend, seasonality, cyclic and residual. ACF considers all these components while finding correlations hence it is a ‘complete auto-correlation plot’.



From the above ACF plot of sunspots, we can see the seasonality in the data for every 11 lags as mentioned before that these spots are related to magnetic cycle where every time sun finishes an approximate 11 year cycle the magnetic pole swap places. At this point we cannot draw conclusion about the pattern of the data, but the pattern indicates it is following MA.

PACF is a partial auto-correlation function. Basically, instead of finding correlations of present with lags like ACF, it finds correlation of the residuals (which remains after removing the effects which are already explained by the earlier lag(s)) with the next lag value hence ‘partial’ and not ‘complete’ as we remove already found variations before we find the next correlation. So, if there is any hidden information in the residual which can be modeled by the next lag, we might get a good correlation and we will keep that next lag as a feature while modeling.



From the above PACF plot of sunspots, the two peak points indicates AR (2) component expected. But with this information we cannot determine which pattern the data is following. So, performing auto arima gives us better results.

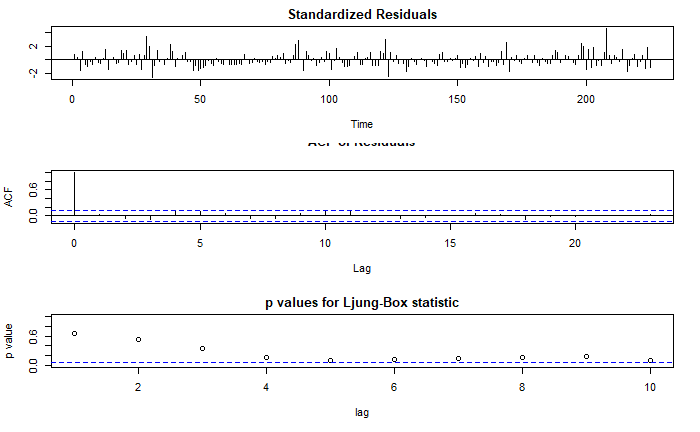
**Model Selection:**

Performing model selection using ARIMA models and tried different possibilities. Below is the observed results given in table with sigma square, likelihood and AIC

|  |  |  |  |
| --- | --- | --- | --- |
| **Order** | **Sigma^2** | **Likelihood** | **AIC** |
| **ARIMA (3,0,1)** | **39805** | **-1509.5** | **3031.01** |
| **ARIMA(1,0,1)(1,0,0)[11]** | **46428** | **-1529.16** | **3066.32** |
| **ARIMA(2,0,1)(1,0,0)[11]** | **38921** | **-1509.5** | **3029.01** |
| **ARIMA(3,0,1)(1,0,0)[11]** | **38917** | **-1509.49** | **3030.98** |
| **ARIMA(2,0,1)(1,0,1)[11]** | **35880** | **-1500.65** | **3013.29** |

Performing autoarima gave result with no seasonality but with ACF plot we have seen seasonality pattern following in the data. So we tried different possibilities with seasonality effect. The model with the lowest AIC and lowest amount of error is a multiplicative seasonal model: ARIMA (2,0,1) (1,0,1)[11]. Not only the AIC we can also see it has the highest likelihood and the lowest sigma square. By considering all these observations we conclude the last observations the best model to continue the analysis.

**Diagnostics of the model:**

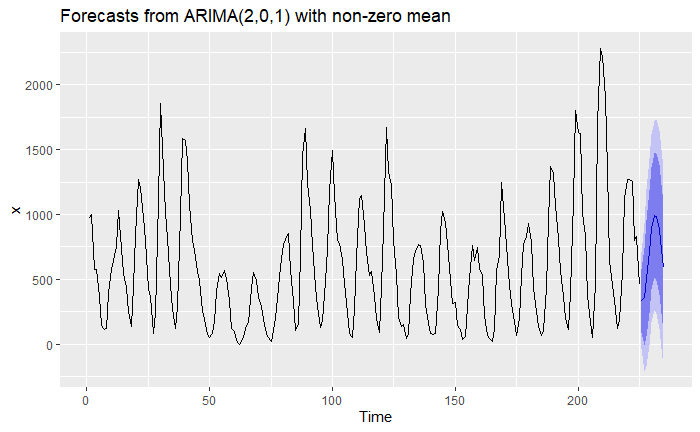


* From Ljung-Box statistic we can see almost all the p values are significant but with a slight white noise.
* Conducted Box-Pierce test to assess the model fit, which yielded a p-value of 0.1 where we failed to reject null hypothesis, that the model is good fit for the data.

**Summary:**

Forecasting:

After we choose the right model for the data, we used the model to predict the sunspots for the next 10 years. Before the analysis of the data we have dropped last 10 records for comparing the predicted values and actual values. Forecasting for the next 10 years is below.



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 1974 | 1975 | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
| Actual | 412.9 | 185.5 | 150.6 | 329.8 | 1111.9 | 1863.3 | 1855.8 | 1684.5 | 1395.5 | 799.6 |
| Upper | 523.92 | 652.84 | 857.5 | 1788.45 | 1259.9 | 1362.5 | 1364.9 | 1272.1 | 1123.2 | 979.4 |
| Pred | 334.58 | 361.10 | 508.37 | 710.47 | 891.07 | 990.33 | 982.95 | 882.57 | 732.46 | 587.47 |
| Lower | 145.08 | 69.36 | 159.1 | 342.35 | 522.09 | 618.08 | 600.93 | 493 | 341.7 | 195.45 |
| S E | 189.42 | 291.74 | 349.2 | 368.05 | 368.91 | 372.25 | 382.02 | 389.57 | 390.76 | 392.02 |

The predicted values and actual values are nearly far to each other, but they lie in the 95% prediction interval. The predicted numerical values are as follows along with interval

We can see the prediction is almost accurate for the year 1974 as 334.58 with 95% confidence interval of upper limit 523.92 and lower limit 145.08

**Example in real-life:**

With the results obtained above, we can analyze the accurate sunspots with the prediction interval which helps in indicating that the cooling and warming of the earth. For example, when scientists are planning to deploy any new technology into space which is indirectly linked with the earths atmosphere, with help of this results they can delay or advance their launch of the product accordingly for better success rate.

**Conclusions:**

* Analyzed the aggregated data of sunspots from year 1749 to 1983 we built time series model to predict sunspots for 10 years.
* This data has almost 130 decades which might reduce the ability of applying as todays global warming might also get added to the factors which are considered.
* Data is stationary which is confirmed through dickey fuller hypothesis test.
* From ACF and PACF plots, we identified the data follows AR and MA components along with seasonal effect.
* After fitting different models, we came up with candidate model ARIMA(2,0,1)(1,0,1)[11] with low AIC and sigma square value.
* The data is forecasted for next 10 years along with the standard error where the predicted values lie within the 95% confidence interval.