Predicting Sunspot Number by SARIMA Model

School of Science, RMIT University, Melbourne.
May 2017
Niranjan Adhikari

Abstract

The dark spots visible on the surface of the sun due to the variation of the temperature caused by the concentration of the solar magnetic flux is called the sunspot. The dataset of a monthly sunspot from 1998 to 2016 prepared by Australian Bureau of Metrology (ABM) is analyzed using time series approach. Some seasonal effect and autocorrelation are noticed in the time series. The SARIMA(0,1,2)x(0,1,1)_4 model best explained this time series with all the parameter significant. The predicted trend of the series shows lesser variation in the series resembling the variation experienced between 2007 and 2011. Diagnostic check on the residuals is not as convincing which is an obvious limitation of the study.

1. Introduction

The dark spot visible on the surface of the sun is a temporary incongruity of the temperature in the sun's mass, caused by the concentration of solar magnetic field flux. This dark spot called the sunspot last from a few days to few months before eventually decaying. A sunspot can vary from 16 kilometers to 160,000 kilometers in diameter. The larger ones can be seen from the earth with our naked eyes. Study of sunspot time series has found many uses and gained the attention of many researchers for a long time. Determining the sunspot cycle period was and 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 (Filipe E. Olvera, 2005).

The dataset in this study consist of a monthly count of a sunspot from 1998 to 2016 prepared by ***Australian Bureau of Meteorology Space Weather Services (ABM)***. As quoted by **ABM** "Although the prediction of solar cycles is difficult, it has practical applications. For example, in planning High Frequency (HF) communication links it is important to estimate what frequencies will be supported by the ionosphere well into the future. Predictions of sunspot number are also very important in planning space-related activities particularly for low Earth orbiting spacecraft".

Understanding the entire solar phenomenon is very complex, yet it remains an important aspect of solar physics, as solar activity is closely associated with the biosphere, space weather and the field of space technology. The commonest study about the solar activity is the prediction of sunspot number along with the solar cycle and various radiations. We analyze the dataset using time series technique and fit an appropriate time series model to predict the number of sunspot for next 5 years.

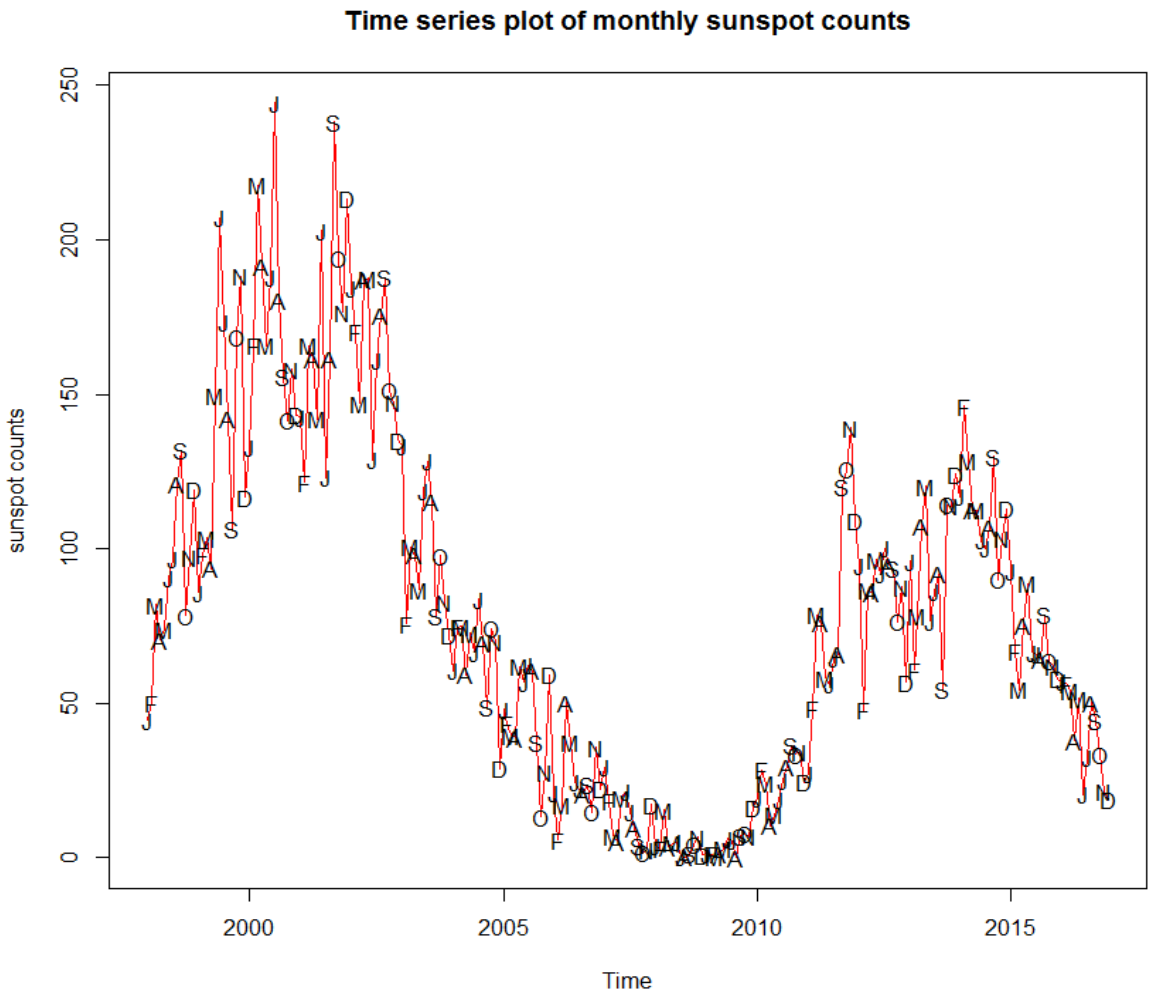
2. Methodology

With the help of visual plots presence of any trend and seasonality in the data set is examined. Using the Augmented-Dickey Fuller test non-stationarity check is performed. The seasonality effect and the trend if dictated is removed from time series through differencing. Finally, appropriate models are shortlisted from which the best one is selected to make the forecast. The selection is based on the residual analysis.

2.1. Time Series plot and data description

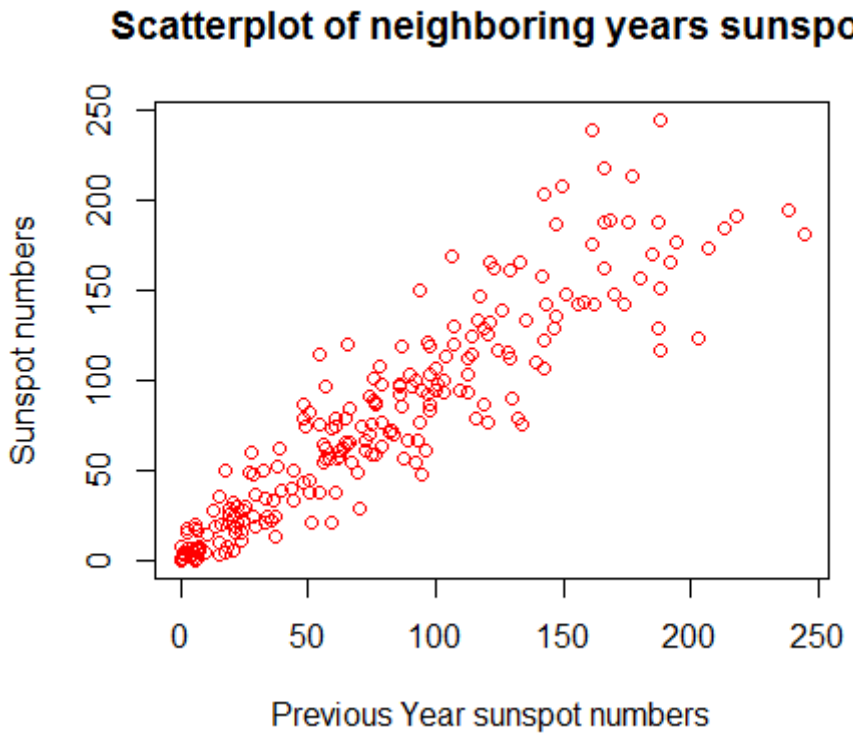
The dataset exhibits an obvious trend in monthly sunspot numbers from 1998 to 2016. It can be noted from the plot(fig.1.) below that the highest is recorded during 2002 and there is gradual fall hitting the lowest during 2008. The sunspot counts during this period ranged from 0 to 244.

Fig.1.Time Series Plot



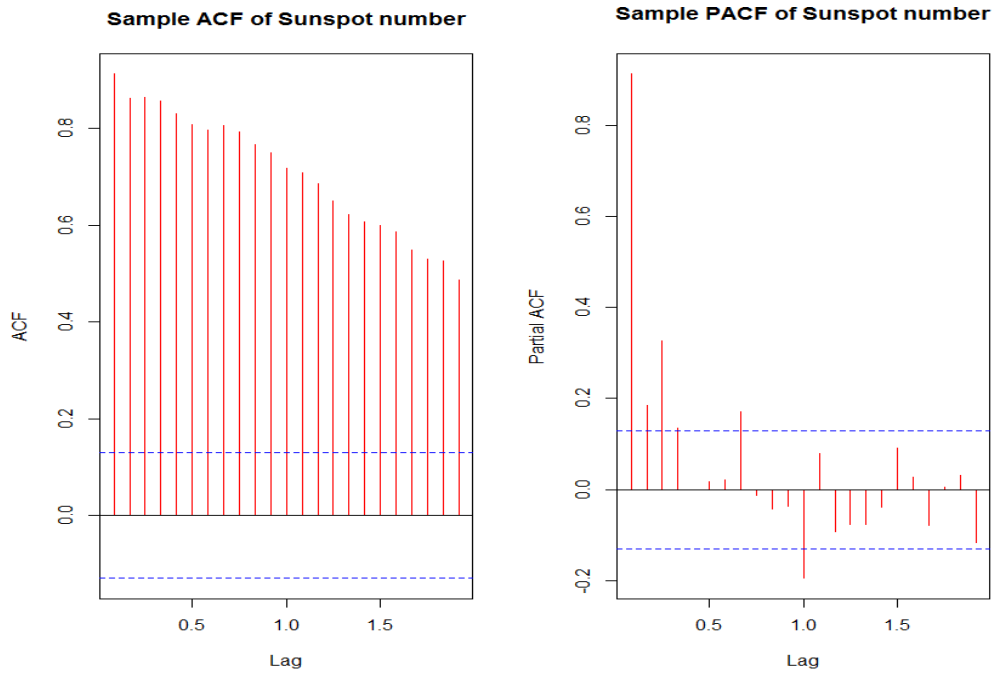
It is also clear from the plot that there exists change of variance along the time series owing to the huge range of sunspot number. There is a strong positive correlation of 0.91 among the numbers of sunspot between the consecutive years as it can be seen in the below scatter plot(fig.2). This means that the sunspot number of present year can have an effect on the sunspot in the following year.

Fig.2.Scatter plot



The following two plots are auto correlogram and partial auto correlogram for the original time series respectively. The slow decaying autocorrelation function (ACF) and the significant initial lags in the partial autocorrelation function (PACF) shows that there exists trend in the time series. Further, from the ACF it is seen that there is a slight increase in every fourth lag along the gradual fall in the process. This confirms the presence of seasonality with period four.

Fig.3.ACF and PACF plots



2.2.Testing Non-Stationarity

As evident from the plot the times series does not follow stationarity as the mean doesn't seem to be constant and the variance largely varies along the time. The Augmented Dickey-Fuller(ADF) unit-root test is used to test the null hypothesis that the process is difference nonstationary which means the process is nonstationary but becomes stationary after differencing. The alternative hypothesis is that the process is stationary. The following output is executed in R studio. It proves that the time series is non-stationary as it fails the reject the null hypothesis.

```
      Augmented Dickey-Fuller Test
data:  data.ts
Dickey-Fuller = -1.6897, Lag order = 11, p-value = 0.7063
alternative hypothesis: stationary
```

The non-stationarity is removed through seasonal differencing and through ordinary differencing of order one.

2.3. Fitting Models

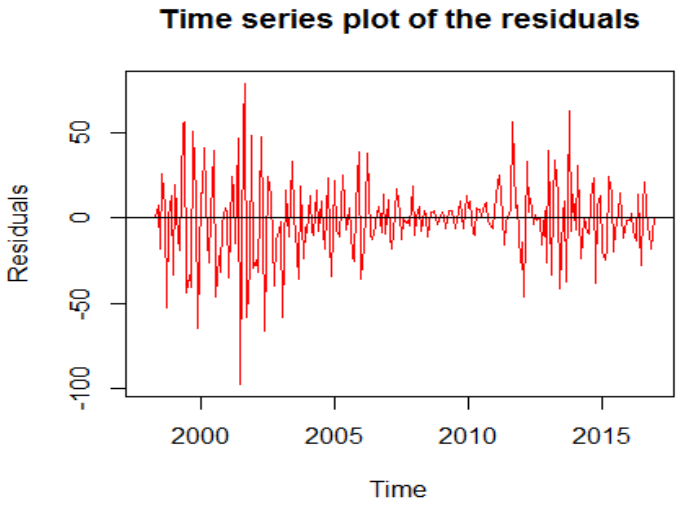
In the first place, it is important to get rid of the seasonal effect in the time series. Following a trial and error method, ARMA process with a first order seasonal difference at D =1 and period 4, is examined and inspected if the seasonal effect is removed. This helped in removing the trend, however, the ACF and PACF plots still showed significant seasonal lag.

Further, ARIMA model with MA (1) process with D =1 helped resolve first seasonal lag while higher order lags are significant. Next, first-order ordinary differencing on the times series is

performed. Taking the ARIMA model with d= 1 and inspecting the ACF and PACF it confirmed that both the trend and seasonality is removed from the time series.

The plot below(fig.4) shows the residual plots of the ARIMA process of seasonal difference with MA (1) process and D=1 and ordinary difference of first order. There is no trend in the residuals but the variance is still not constant. The ACF and PACF also showed that not many lower order lags are significant and there are no seasonal lags, so we can go ahead and shortlist possible models from the above ARIMA process with an ordinary difference and seasonal difference of order one and period of four.

Fig.4.Residual Plot



3. Model selection

It is clear from the above analyses that the possible models we can derive are seasonal ARIMA models with a first-order difference of the general form

ARIMA $(p, d, q) \times (P, D, Q).S$, where S is the period of the seasonal lag, the p & P are the order of AR process and q & Q MA processes, and the d & D represents the order of difference for ordinary and seasonal differencing respectively.

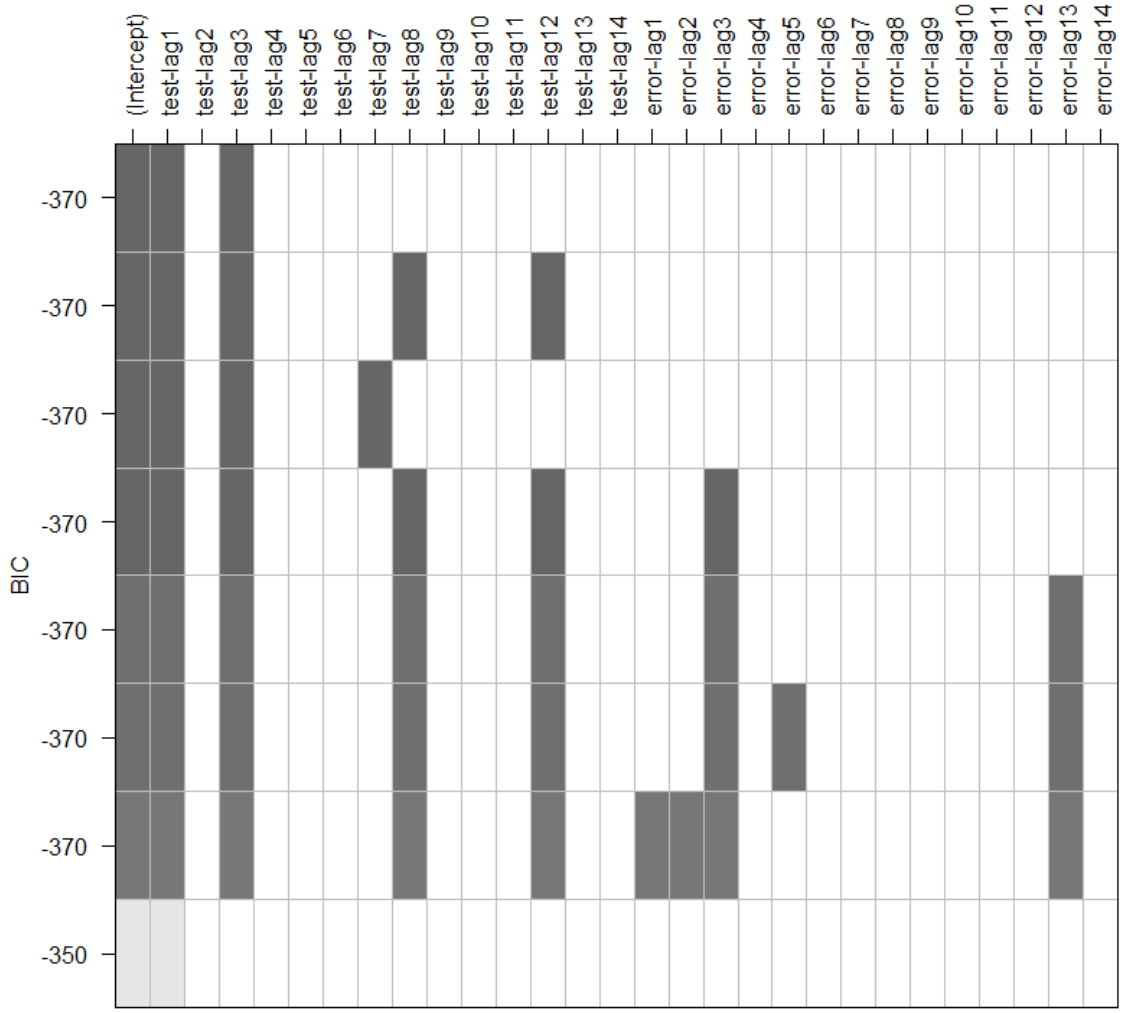
The models are derived from the extended partial correlation (EACF) and BIC plots.

The following is the matrix of the EACF of the residual of seasonal ARIMA (0,1,0)x(0,1,1).4

AR/MA

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	o	o	o	o	o	o	o	o	o	x	o	o
1	x	x	o	x	o	o	o	o	x	o	o	o	o	o
2	x	o	o	o	o	o	o	o	o	o	o	o	x	o
3	o	o	o	o	x	o	o	o	o	o	o	o	o	o
4	x	o	o	o	x	o	o	o	o	o	o	o	o	o
5	x	x	o	o	x	o	o	o	o	o	x	o	o	o
6	x	o	o	x	x	o	o	o	o	o	x	o	o	o
7	o	x	o	x	x	o	x	o	o	o	x	o	o	o

Fig.5.BIC plot for the residuals



The above plot is the BIC plot for the residual of the seasonal ARIMA discussed above. From the EACF and BIC plot, the following candidate models can be derived:

From the EACF

i. SARIMA(0,1,2)x(0,1,1)_4
ii. SARIMA(0,1,3)x(0,1,1)_4
iii. SARIMA(1,1,2)x(0,1,1)_4
iv. SARIMA(2,1,1)x(0,1,1)_4

From BIC table

i. SARIMA(1,1,0)x(0,1,1)_4
ii. SARIMA(3,1,0)x(0,1,1)_4

3.1. Test of Significance of Parameter

Each of the above candidate models is evaluated using maximum likelihood method to see if all the processes are significant in the model. The table below represents each of the time series

components in the seasonal ARIMA model. Some of the components are significant and others are not. The best model as per this analysis would be the one with all the components significant and with lesser the number of parameter. From this table, the seasonal ARIMA models are narrowed to first and the sixth as all the components are significant. Further evaluation can be conducted using the AIC and BIC criterion of the residuals.

Table.1. The significance of time series process

Sl. No.	Models	AR(1)	AR(2)	AR(3)	MA(1)	MA(2)	MA(3)	SMA(1)
1	SARIMA(0,1,2)x(0,1,1)_4				**	**		**
2	SARIMA(0,1,3)x(0,1,1)_4				**	**	*N.S*	**
3	SARIMA(1,1,2)x(0,1,1)_4	*N.S*			*N.S*	**		**
4	SARIMA(2,1,1)x(0,1,1)_4	*N.S*	**		**(SS)*			**
5	SARIMA(1,1,0)x(0,1,1)_4	*N.S*	**		**(SS)*			**
6	SARIMA(3,1,0)x(0,1,1)_4	**	**	*(SS)				**

*Note: ** means significant, SS means slightly significant and N.S is not significant. The blanks mean an absence of that process.*

The table (table.2) shows the AIC and BIC for candidate models sorted in increasing order of the AIC and BIC values. As per these criterion it to choose between *SARIMA(3,1,0)X(0,1,1)_4* and *SARIMA(0,1,2)X(0,1,1)_4* as they have the least AIC and BIC values. However, as it a general practice to avoid the models with a higher number of parameter, the better model would be *SARIMA(0,1,2)X(0,1,1)_4* which does not involve an MA process.

Table.2. AIC and BIC

SL. No.	Models	AIC	Models	BIC
1	SARIMA(3,1,0)x(0,1,1)_4	2023.08	SARIMA(0,1,2)x(0,1,1)_4	2039.055
2	SARIMA(2,1,1)x(0,1,1)_4	2023.218	SARIMA(3,1,0)x(0,1,1)_4	2040.116
3	SARIMA(0,1,3)x(0,1,1)_4	2025.323	SARIMA(2,1,1)x(0,1,1)_4	2040.254
4	SARIMA(0,1,2)x(0,1,1)_4	2025.426	SARIMA(0,1,3)x(0,1,1)_4	2042.359
5	SARIMA(1,1,2)x(0,1,1)_4	2026.239	SARIMA(1,1,2)x(0,1,1)_4	2043.275
6	SARIMA(1,1,0)x(0,1,1)_4	2053.833	SARIMA(1,1,0)x(0,1,1)_4	2064.055

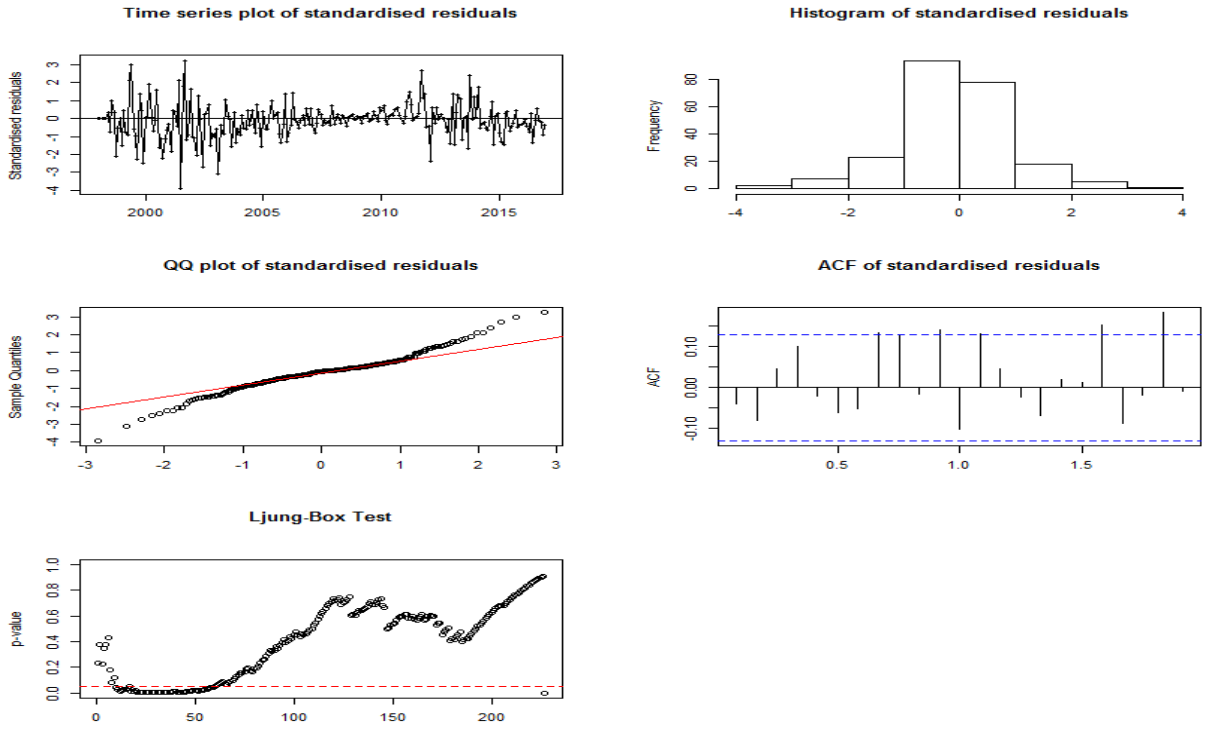
This process has the following components and the coefficients:

```
z test of coefficients:
      Estimate Std. Error  z value  Pr(>|z|)
ma1  -0.317472   0.067339  -4.7145 2.423e-06 ***
ma2  -0.249580   0.063282  -3.9439 8.015e-05 ***
sma1 -0.982177   0.074568 -13.1716 < 2.2e-16 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*3.2.*Residual Analysis

The histogram and the QQ-normal plot shows that the residuals are somewhat normally distributed. The ACF proves that the residual exhibits a white noise process and the time series plot of the residuals don't show much change in the variance. However, the Shapiro-Wilk test does not support the normality in the residuals.

Fig.6.Residual Plots for SARIMA(0,1,2)X(0,1,1)_4



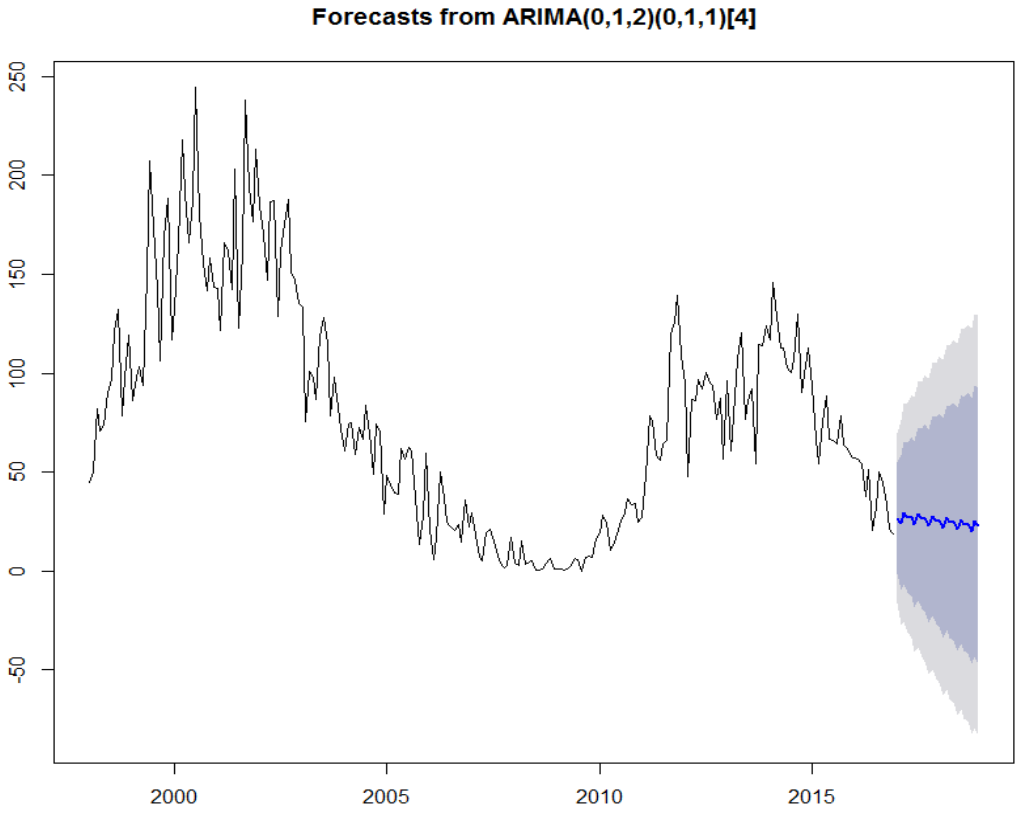

```
        Shapiro-Wilk normality test

data:  res.model
W = 0.96267, p-value = 1.106e-05
```

4. Predictions

The forecast of the sunspot number using the SARIMA(0,1,2)x(0,1,1).4 with the 95 percent confidence interval is plotted(fig.7). The 80 percent confidence interval has a wider range of shade grey. The predicted values for the next 24 months are however varying between 20 and 30. The sunspot numbers don't seem to demonstrate a high variation from 2016 to 2018 as compared to the past. The similar pattern was also experienced between 2007 and 2011.

Fig.7.Forecast



5. Limitation of Model

Some limitations of SARIMA(0,1,2)x(0,1,1)_4 model are

i. The change in variances of the residuals persists in the model which suggests that the entire series could not convert into a series with constant variance. This might affect performance accuracy of the model.

ii. Residuals of the model are not normally distributed which are not reliable to demonstrates a good accuracy from the model. The deviation points on the tail of fitted QQ norm plot suggests that the high spark on the series is not covered by the model.

iii. The higher order lags on the ACF plot of the residuals of the model are significant that may suggest some part of series might not be covered by the model. The expected ACF plot should be white noise.

6. Conclusion

The SARIMA(0,1,2)x(0,1,1)_4 is a suitable model for this data set as there were an obvious trend and seasonal effect. The adequacy of the model shown by the residual analysis are not promising but we can get quite a good prediction as shown in the plots. As per this model, we can conclude that the sunspot numbers seem to be steady compared to historical records.

Acknowledgment

We would like to express our sincere thanks of gratitude to our project mentor and guide Dr. Haydar Demirhan for giving us the opportunity to do this wonderful project. It is a great honor to get continuous support and guidance throughout the project from Dr. Demrihan. The meetings we had with Dr. Demirhan helped us visualize the problem more practically and gained numerous other ideas on handling time series problem. This made us more equipped and kept us encouraged for the task. Our, acknowledgment would be incomplete without the mention of the weekly laboratory hand-on task sessions delivered by Dr. Demrihan which contributed hugely to our learning and compiling of the entire project.

References

Anon., n.d. *Sunspot.* [Online]
Available at: https://en.wikipedia.org/wiki/Sunspot [Accessed 25 April 2017].

Filipe E. Olvera, J., 2005. *A Spectral Analysis of the Sunspot Time Series Using the Periodogram,* s.l.: Portland State University Maseeh College of Engineering and Computer Science.

Appendix

```
> rm(list = ls())
> library(TSA)
> library(forecast)
> library(lmtest)
> library(fGarch)
> library(readr)

# This  function sort the AIC and BIC accoring to their score
> sort.score <- function(x, score = c("bic", "aic")){
  if (score == "aic"){
    x[with(x, order(AIC)),]
  } else if (score == "bic") {
    x[with(x, order(BIC)),]
  } else {
    warning('score = "x" only accepts valid arguments
("aic","bic")')
  }
}

# This function produce the out put for residual analysis
> residual.analysis <- function(model, std = TRUE){
  library(TSA)
  library(FitAR)
  if (std == TRUE){
    res.model = rstandard(model)
  }else{
    res.model = residuals(model)
  }
  par(mfrow=c(3,2))
  plot(res.model,type='o',ylab='Standardised residuals',
main="Time series plot of standardised residuals")
  abline(h=0)
  hist(res.model,main="Histogram of standardised residuals")
  qqnorm(res.model,main="QQ plot of standardised residuals")
  qqline(res.model, col = 2)
  acf(res.model,main="ACF of standardised residuals")
  print(shapiro.test(res.model))
  k=0
  LBQPlot(res.model, lag.max = length(model$residuals)-1 ,
StartLag = k + 1, k = 0, SquaredQ = FALSE)
}

# This code read the data set
```

```
> data <- read_csv("C:/Users/ndnad/Desktop/Time Series
Analysis/Project/sunspotnumbers.csv",
                                 col_types = cols(Apr =
col_number(),
                                                                  Aug
= col_number(), Dec = col_number(),
                                                                  Feb
= col_number(), Jan = col_number(),
                                                                  Jul
= col_number(), Jun = col_number(),
                                                                  Mar
= col_number(), May = col_number(),
                                                                  Nov
= col_number(), Oct = col_number(),
                                                                 Sep
= col_number()))

# This code change the data frame to time seris
> data.ts = as.vector(t(data[,-1]))
> data.ts = ts(data.ts,start=c(1998,1), end=c(2016,12),
frequency=12)
> class(data.ts)

# plot  of time series plot
> plot(data.ts,ylab='sun spots no',xlab='Year',type='o',
      main = "Time series plot of monthy sun spot number ")

# seasonility check (can you help me here)
> plot(data.ts, type = "l", ylab='sun spot',main = "Time
series plot.")
> points(y=data.ts,x=time(data.ts),
pch=as.vector(season(data.ts)))

# scatter plot
> plot(y=data.ts,x=zlag(data.ts),ylab='sun spot',
xlab='Previous Year sun spot' , main = "Scatter plot of
neighboring sun spots")

# this is not working
> y=data.ts
> x = zlag(data.ts)  # Generate first lag of the Spawners
series
> index = 2:length(x)
> cor(y[index],x[index])
```

```
> par(mfrow=c(1,2))
> acf(data.ts, main="The sample ACF of sun spot number
series")
> pacf(data.ts, main="The sample PACF of sun spot number
series")

# The Dickey-Fuller Unit-Root test (ADF test)
> ar(diff(data.ts)) # To find the value of lag in the
following adfTest() function
#adfTest(data.ts, lags = 11, type = "ct", title =
NULL,description = NULL)
> adf.test(data.ts, k = 11)
# The p value of 0.7063 > 0.5, tells us we can not reject the
null hypothesis of data is non stationary.
# with this unit roots test we are conforming that the series
is *seasonal nonstationary*. We need to look for
 # Seasonality and existence of trend are apparent from the
ACF and PACF plots

#--------------- seasonal arima model-------------------------
# First fit a plain model with only the first seasonal
difference with order D = 1
# and see if we can get rid of the seasonal trend effect
# by inspecting the autocorrelation structure of the
residuals.
> m1.sunspot =
arima(data.ts,order=c(0,0,0),seasonal=list(order=c(0,1,0),
period=4))
> res.m1 = residuals(m1.sunspot);
> plot(res.m1,xlab='Time',ylab='Residuals',main="Time series
plot of the residuals")
> par(mfrow=c(1,2))
> acf(res.m1, lag.max = 36, main = "The sample ACF of the
residuals")
> pacf(res.m1, lag.max = 36, main = "The sample PACF of the
residuals")
#  see if we get rid of seasonal component.
# adding MA(1) in sasoanl part
> m2.sunspot =
arima(data.ts,order=c(0,0,0),seasonal=list(order=c(0,1,1),
period=4))
> res.m2 = residuals(m2.sunspot);
> plot(res.m2,xlab='Time',ylab='Residuals',main="Time series
plot of the residuals")
```

```
> par(mfrow=c(1,2))
> acf(res.m2, lag.max = 36, main = "The sample ACF of the
residuals")
> pacf(res.m2, lag.max = 36, main = "The sample PACF of the
residuals")

# we will apply differentiating on the ordinary seris and see
if we can see the trend more clearly.
> m4.sunspot =
arima(data.ts,order=c(0,1,0),seasonal=list(order=c(0,1,1),
period=4))
> res.m4 = residuals(m4.sunspot);
> plot(res.m4,xlab='Time',ylab='Residuals',main="Time series
plot of the residuals")
> par(mfrow=c(1,2))
> acf(res.m4, lag.max = 36, main = "The sample ACF of the
residuals")
> pacf(res.m4, lag.max = 36, main = "The sample PACF of the
residuals")

# We are going to see the possible sets of models using each
and BIC of residuals of the
# above models

> eacf(res.m4)
> res = armasubsets(y=data.ts,
nar=14,nma=14,y.name='test',ar.method='ols')
> plot(res)

# From the EACF, we will include AR(1) order as well.
# SARIMA(0,1,2)x(0,1,1)_4
# SARIMA(0,1,3)x(0,1,1)_4
# SARIMA(1,1,2)x(0,1,1)_4 and
# SARIMA(2,1,1)x(0,1,1)_4 will be fitted

# from BIC table we wil include AR(3)
# SARIMA(1,1,0)x(0,1,1)_4
# SARIMA(3,1,0)x(0,1,1)_4

> m4_012.sunspot = arima(data.ts
,order=c(0,1,2),seasonal=list(order=c(0,1,1), period=4),method
= "ML")
> res_012 = residuals(m4_012.sunspot);
```

```
> plot(res_012,xlab='Time',ylab='Residuals',main="Time series
plot of the residuals")
> par(mfrow=c(1,2))
> acf(res_012, lag.max = 36, main = "The sample ACF of the
residuals")
> pacf(res_012, lag.max = 36, main = "The sample PACF of the
residuals")
# still have some significant lags in ACF and PACF

> m4_013.sunspot = arima(data.ts
,order=c(0,1,3),seasonal=list(order=c(0,1,1), period=4),method
= "ML")
> res_013 = residuals(m4_013.sunspot);
> plot(res_013,xlab='Time',ylab='Residuals',main="Time series
plot of the residuals")
> par(mfrow=c(1,2))
> acf(res_013, lag.max = 36, main = "The sample ACF of the
residuals")
> pacf(res_013, lag.max = 36, main = "The sample PACF of the
residuals")
# still have some significant lags in ACF and PACF

> m4_112.sunspot = arima(data.ts
,order=c(1,1,2),seasonal=list(order=c(0,1,1), period=4),method
= "ML")
> res_112 = residuals(m4_112.sunspot);
> plot(res_112,xlab='Time',ylab='Residuals',main="Time series
plot of the residuals")
> par(mfrow=c(1,2))
> acf(res_112, lag.max = 36, main = "The sample ACF of the
residuals")
> pacf(res_112, lag.max = 36, main = "The sample PACF of the
residuals")

> m4_211.sunspot = arima(data.ts
,order=c(2,1,1),seasonal=list(order=c(0,1,1), period=4),method
= "ML")
> res_211 = residuals(m4_211.sunspot);
> plot(res_211,xlab='Time',ylab='Residuals',main="Time series
plot of the residuals")
> par(mfrow=c(1,2))
> acf(res_211, lag.max = 36, main = "The sample ACF of the
residuals")
```

```
> pacf(res_211, lag.max = 36, main = "The sample PACF of the
residuals")

# from BIC table ma(3)
> m4_110.sunspot = arima(data.ts
,order=c(1,1,0),seasonal=list(order=c(0,1,1), period=4),method
= "ML")
> res_110 = residuals(m4_110.sunspot);
> plot(res_110,xlab='Time',ylab='Residuals',main="Time series
plot of the residuals")
> par(mfrow=c(1,2))
> acf(res_110, lag.max = 36, main = "The sample ACF of the
residuals")
> pacf(res_110, lag.max = 36, main = "The sample PACF of the
residuals")

> m4_310.sunspot = arima(data.ts
,order=c(3,1,0),seasonal=list(order=c(0,1,1), period=4),method
= "ML")
> res_310 = residuals(m4_310.sunspot);
> plot(res_310,xlab='Time',ylab='Residuals',main="Time series
plot of the residuals")
> par(mfrow=c(1,2))
> acf(res_310, lag.max = 36, main = "The sample ACF of the
residuals")
> pacf(res_310, lag.max = 36, main = "The sample PACF of the
residuals")

> coeftest(m4_012.sunspot) # all significant
> coeftest(m4_013.sunspot) # MA(3) is not significant
> coeftest(m4_112.sunspot) # MA(1) and AR(1) is not
significant
> coeftest(m4_110.sunspot) # all significant
> coeftest(m4_211.sunspot) # AR(1) is not significant
> coeftest(m4_310.sunspot) # all are signifcant

> sc.AIC=AIC(m4_012.sunspot, m4_013.sunspot,
m4_112.sunspot,m4_110.sunspot, m4_211.sunspot, m4_310.sunspot)
> sc.BIC=BIC(m4_012.sunspot, m4_013.sunspot,
m4_112.sunspot,m4_110.sunspot, m4_211.sunspot, m4_310.sunspot)

> sort.score(sc.AIC, score = "aic")
> sort.score(sc.BIC, score = "bic")
```

```
# using coef test, AIC and BIC m4_012.sunpost or
[sarima(0,1,2)X(0,,1)_4] is best model

> residual.analysis(model = m4_012.sunspot) # good but still
have some lags signicant
> residual.analysis(model = m4_013.sunspot) # still good but
proble with lag 12
> residual.analysis(model = m4_112.sunspot) # still good but
problem with lag 12
> residual.analysis(model = m4_110.sunspot) # good
> residual.analysis(model = m4_211.sunspot) # still good but
slight significan line in ACF
> residual.analysis(model = m4_310.sunspot) # still good but
slight significan line in ACF

> m4_012.sunspot = Arima(data.ts, order=c(0,1,2),
seasonal=list(order=c(0,1,1), period=4), method = "ML")
> preds1 = forecast(m4_012.sunspot, h = 24)
> plot(preds1)

> m4_310.sunspot = Arima(data.ts, order=c(3,1,0),
seasonal=list(order=c(0,1,1), period=4), method = "ML")
> preds2 = forecast(m4_310.sunspot, h = 24)
> plot(preds2)

# choose either of above, I think (m4_012.sunspot) is a better
model because this model is quite small than the
(m4_310.sunspot)
```

~ Thank you ~